Aus dem
CharitéCentrum 2 für Grundlagenmedizin
Institut für Molekularbiologie und Bioinformatik
Direktor: Professor Dr. Burghardt Wittig

# Habilitationsschrift

## Discovering transcriptional networks for cardiac development, function and disease with an systems biology approach

Zur Erlangung der Lehrbefähigung
Für das Fach Molekularbiologie und Bioinformatik

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

**Frau Dr. med. Silke Rickert-Sperling**
geboren am 31.3.1971 in Chemnitz

| | |
|---|---|
| Eingereicht: | Oktober 2008 |
| Dekanin: | Professorin Dr. med. A. Grüters-Kieslich |
| 1. Gutachter: | Professor Dr. med. Dr. rer. nat. T. Braun |
| 2. Gutachter: | Professor Dr. techn. Z. Trajanoski |

# Inhaltsverzeichnis

**1 Introduction** ........ **3**

**2 Results and Discussion** ........ **6**

**2.1 Regulatory transcription networks controlling cardiac muscle development and function** ........ **6**

2.1.1 Cardiovascular genetics database – integrating clinical and molecular data ........ 6

2.1.2 Genome-wide expression profiling of congenital heart disease ........ 17

2.1.3 Analysis of cardiac transcription factors and epigenetic marks at a global scale ........ 30

2.1.4 Prediction of cardiac transcription networks ........ 46

2.1.5 Genomic organization of the transcriptome ........ 64

**2.2 DNA-binding and epigenetic transcription factors relevant for cardiac development and disease** ........ **82**

2.2.1 Mutations of the transcription factor CITED2 lead to congenital heart defects ........ 82

2.2.2 Characterization of T-box transcription factor TBX20 in human hearts ........ 93

2.2.3 Regulation of cardiac and skeletal muscle development by DPF3 – a novel epigenetic transcription factor ........ 109

**2.3 Advanced biotechniques to analyze patient material** ........ **130**

2.3.1 A novel technique for quantitative PCR analysis – cr-real-time PCR ........ 130

2.3.2 Refined genotyping for sparse sample material – LDR-TaqMan ........ 137

**3 Summary** ........ **150**

**4 References** ........ **151**

**6 Acknowledgment** ........ **161**

For my family

# 1 Introduction

The heart is the first functional organ during embryogenesis and the one most susceptible to disease. A rapidly growing number of factors have been shown to be involved in regulating the pattern and timing of expression of genes responsible for the cardiac lineage determination, heart chamber formation, valvulogenesis and conduction-system development (Clark et al., 2006). Spatiotemporal and quantitative regulation of cardiac transcription factors (TFs) must occur in a precise manner to ensure fine regulation of downstream targets. The complexity of these molecular cascades during development may explain the sensitivity of the heart to perturbations before birth and into old age. Congenital heart diseases (CHD) are the most common birth defects in humans. They arise during development of the embryo and affect 1 in every 100 live births and an even higher number in miscarriages (Hoffman, 1995a; Hoffman, 1995b).

Although major insight of the cardiac developmental process has been gained by the study of animal models like mice, chicken and zebrafish, little is known about the genetic basis in human. The overwhelming majority of congenital heart malformations do not segregate in Mendelian ratios, although they show familial aggregation, which has long suggested a role of genetic factors in their development. Approaching 30% of major cardiac malformations are associated with additional developmental abnormalities and result from a recognized chromosomal anomaly or occur as part of a syndrome. To gain insight into the formation of cardiac anomalies molecular genetic studies of human patient populations have been carried out. Linkage analyses and candidate-gene approaches have led to the identification of several gene mutations causing CHD (*e.g. GATA4, NKX2-5* and *ZIC3, CITED2*) (Schott et al., 1998; Garg et al., 2003; Ware et al., 2004; Sperling et al., 2005). However, most heart malformations display variable expressivity and penetrance pointing to a multifactorial and multigenic basis. In humans and mice similar mutations can cause a variety of phenotypes from one family, individual or inbred strain, respectively to another. Heterozygous mutations in the homeobox transcription factor NKX2-5 in human can lead to diverse abnormalities as atrial septal defects (ASDs), ventricular septal defects (VSDs), Ebstein's anomaly of the tricuspid valve, AV block, or Tetralogy of Fallot (TOF), either alone or in combinations (Benson et al., 1999). Haploinsufficiency is thought to be at the root of the malformations. A similar situation exists for the t-box factor TBX5, in which heterozygous mutations cause a variety of CHD in the context of Holt-Oram syndrome (Basson et al., 1997). The linkage to haploinsufficiency is supported by the occurrence of the same syndrome in mouse with one deleted copy of *Tbx5* (Bruneau et al., 2001). The symptom severity of cardiac defects also depends on the type of mutation. Some missense mutations result in a non-functional protein, whereas others may lead to altered properties of unknown nature (Cross et al., 2000). Certain mutations abolish binding of Tbx5 to its DNA-binding sites (Ghosh et al., 2001), whereas others influence collaboration with other proteins (Hiroi et al., 2001). For example, Nkx2-5 physically interacts with Tbx5 and Gata4 to synergistically activate downstream target genes (Small and Krieg, 2003; Takeuchi et al., 2003). Disruption of the stoichiometry of the TF interaction by decreased amount of either protein may lead to similar effects on transcriptional targets. Intriguingly, mutations in human α-myosin heavy chain (MYH6), a direct target of NKX2-5, TBX5 and GATA4, also cause ASDs (Ching et al., 2005). Additionally, the disease manifestation of decreased TF dosage may vary due to stochastic events of unknown nature or parameters comprising environmental influences and genetic modifiers. For example, it has been shown that a decreased level of Tbx20 affects heart development via a breakdown of transcription factor networks (Takeuchi et al., 2005). This suggests that the regulatory context of TFs plays an important role and its function must be viewed in the context of transcriptional networks including the interplay between different TFs.

The ability of TFs to bind to DNA is highly influenced by the accessibility of their binding sites. In eukaryotic cells, DNA is packaged into chromatin by association with histone proteins. A high compaction of chromatin renders the DNA inaccessible to TF binding, silencing the genes in these regions.

Genomic DNA is packaged into nucleosomes, the basic unit of chromatin structure formed by DNA wrapped around a histone octamer. Chromatin remodelling and covalent histone modifications facilitate DNA access for DNA-binding transcription factors (Simone, 2006; Bernstein et al., 2007; Sperling, 2007). Specific patterns of histone tail modifications attract or repel regulatory proteins of the chromatin remodelling complex. Histone modifications, such as acetylations and methylations, can influence one another and thus not just the level of modification but also the pattern may dictate biological outcome (Fischer et al., 2008b).

Chromatin remodelling complexes are recruited to their target nucleosomes via two mechanisms, the guidance by DNA-binding transcription factors and the binding to acetylated histone tails. Mammalian chromatin remodelling complexes (SWI/SNF-like complexes, BAF complexes) are characterized by central core subunits BRG1 and BRM and 10 further subunit elements, e.g. SMARCD3 (BAF60c) representing a muscle specific component. BRG1 and BRM contain an ATPase domain and a bromodomain that recognizes acetylated lysine in histone tails and other proteins (Sif, 2004; Simone, 2006). Thus, BRG1 acts as a ubiquitously expressed targeting molecule to anchor chromatin remodelling complexes on promoters with particular histone modification marks (Hassan et al., 2002; Hassan et al., 2007). SMARCD3 is a promiscuous partner for several DNA-binding transcription factors, including nuclear receptors PPARγ, RXRα, RAR and muscle regulatory factors like MEF2, MyoD, Nkx2.5, Tbx5 and Gata4 (Debril et al., 2004; Lickert et al., 2004; Palacios and Puri, 2006; Simone, 2006; Flajollet et al., 2007; Li et al., 2007). Tissue specific transcription can be initiated by ligand-dependent activation of signalling cascades, e.g. phosphorylation of SMARCD3 and MEF2 through p38 MAP-kinase leads to translocation of MEF2 to the nucleus, potentially enhances their interaction and finally the BAF complex is targeted to muscle specific loci (Simone et al., 2004; Rauch and Loughna, 2005).

A number of transcription factors have been implicated to interact with histone modifying enzymes. For example the histone acetyl transferase (HAT) p300 not only acetylates lysine residues on histone 3 but also on Gata4, thereby enhancing the DNA-binding and activating potential of this factor (Kouzarides, 2007). Furthermore, the Srf-cofactor Myocardin (Myocd) has been reported to recruit p300 to Srf binding sites whereby histone 3 acetylation is induced and gene expression enhanced (Cao et al., 2005). The effect of histone acetylases is counteracted by histone deacetylases (HDACs). If embryonic stem cells are treated with inhibitors of HDACs the level of acetylated Gata4 increases and the cells differentiate into cardiomyocytes (Kawamura et al., 2005), demonstrating that the acetylation of Gata4 as well as histone 3 is a critical step in the formation of cardiomyocytes. The deletion of HDACs in mice leads to early lethality and a spectrum of cardiac abnormalities (Zhang et al., 2002; Montgomery et al., 2007). The recruitment of chromatin remodelling complexes is highly affected by histone acetylation, which might explain the sever phenotype. Recently the muscle expressed epigenetic transcription factor DPF3 was discovered, which links these modifications to the BAF complex and displays an essential role for skeletal and cardiac muscle development (Lange et al., 2008).

In summary, to understand the networks directing gene expression not only the interplay between different TFs and co-regulatory elements but also epigenetic factors such as histone modifications has to be considered. In the following a panel of studies is described that aimed to discover and analyze key nodes of the transcription network underlying normal and diseased cardiac development and function. Furthermore, a cardiac transcription network is predicted based on obtained molecular data and clinical phenotypes.

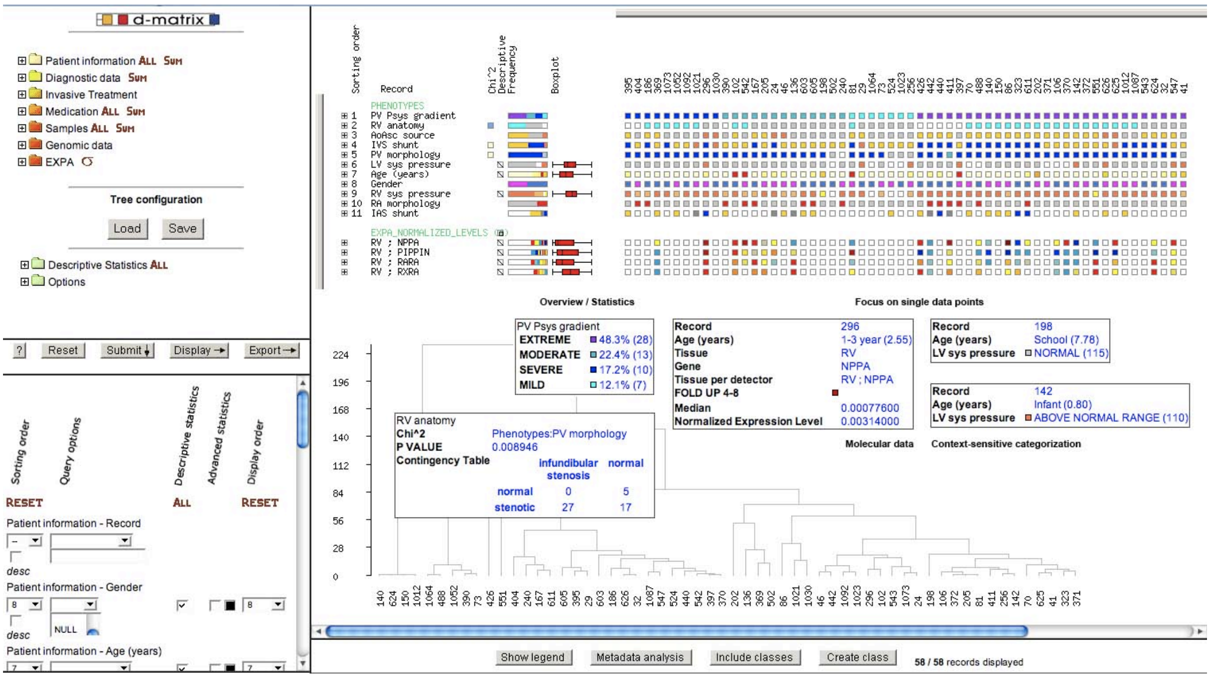
# 2 Results and Discussion

## 2.1 Regulatory transcription networks controlling cardiac muscle development and function

### 2.1.1 Cardiovascular genetics database – integrating clinical and molecular data

> Seelow D, Galli R, Mebus S, Sperling HP, Lehrach H, **Sperling S**. d-matrix - database exploration, visualisation and analysis. *BMC Bioinformatics* 2004;**5**:168.

Both the generation and the analysis of genome, transcriptome and proteome data are becoming increasingly widespread, and these data must be integrated to generate a molecular phenotype. Moreover, the correlation of molecular with phenotype data requires both with comparable profoundness, which lead to the development of the CardioVascular Genetic database (CVGdb). CVGdb stores the detailed clinical phenotype of patients with congenital heart diseases as well as molecular data such as gene expression analysis results (Kaynak et al., 2003) and genotypes. Since the majority of CHD can be traced back to abnormalities in specific developmental milestones, the detailed phenotypical description of analyzed heart defects remains to be the first step for their association with genetic and epigenetic data as well as environmental influences. A dedicated phenotyping scheme for CHD was developed, which is based on cardiac segments raising at certain developmental milestones. For storage of these phenotype information and their association with obtained molecular data a relational database (Oracle 8i) was set-up. Until now, samples and phenotype information of around 500 patients were collected after informal consents. As the phenotyping scheme consists of 150 different attributes and 400 values for each dataset, a major challenge was the development of a graphical front-end, allowing a dedicated user-specified display of the stored information, satisfying the association of phenotype with molecular data and handling simple analyzing procedures. Based on these needs "d-matrix" was developed, a data mining software with a display of three dimensions in form of colour-coded boxes or bars arranged as a matrix (**Figure 1**).



PV Psys gradient

| | |
|---|---|
| EXTREME | 48.3% (28) |
| MODERATE | 22.4% (13) |
| SEVERE | 17.2% (10) |
| MILD | 12.1% (7) |

RV anatomy

| | |
|---|---|
| Chi^2 | Phenotypes:PV morphology |
| P VALUE | 0.008946 |
| Contingency Table | |

| | infundibular stenosis | normal |
|---|---|---|
| normal | 0 | 5 |
| stenotic | 27 | 17 |

| | |
|---|---|
| Record | 296 |
| Age (years) | 1-3 year (2.55) |
| Tissue | RV |
| Gene | NPPA |
| Tissue per detector | RV ; NPPA |
| FOLD UP 4-8 | ■ |
| Median | 0.00077600 |
| Normalized Expression Level | 0.00314000 |

| | |
|---|---|
| Record | 198 |
| Age (years) | School (7.78) |
| LV sys pressure | □ NORMAL (115) |

| | |
|---|---|
| Record | 142 |
| Age (years) | Infant (0.80) |
| LV sys pressure | ■ ABOVE NORMAL RANGE (110) |

**Figure 1. Data selection and graphical output of d-matrix**. Shown are phenotype features of the Tetralogy of Fallot and their association with gene expression levels of relevant genes.

Querying and analyzing stored data to uncover the valuable information hidden in the databases are difficult tasks. With some exceptions, these are approached by a two-step procedure, in which a database specific front-end serves the query and extraction of data, which are subsequently imported in stand-alone analysis tools for visualization, mining and statistics (Walker et al., 1999; Falkman, 2001; Wegman, 2003). Moreover, the visualization and mining tools frequently focus on presenting overall views of data sets for a specific task and seldom permit single-case addressability or have drill-down capability. In today's systems, the perceptual abilities of human users are only used to a limited extend. However, it is essential to make users part of the overall process through computer support of their intelligence, creativity and perceptual abilities. Hence, a major research challenge was to find human-oriented forms of representing information and enabling rapid interaction between humans and computers in the query, visualization and analysis process.

One visual representation, which motivated the graphical display of the d-matrix, was the data matrices handled in microarray studies, in which rows in the matrices typically represent genes and columns individual samples (Eisen et al., 1998). Rather than showing a numerical 'spreadsheet', it is convenient to display microarray data in such matrices, which indicate varying expression levels in a grid of varying colours.

Taken together, d-matrix is a generic front-end solution capable of extracting, exploring, visualizing and analyzing complex data. It can be interfaced with the most common relational database management systems without any intervention on the schema or pre-processing phase, and represents the front-end of CVGdb (**Figure 1**). As the name suggests, the visual model proposed has the form of a matrix. Its elements are boxes whose colours show the quality of the underlying information. The granularity of the data display allows consequent drill-down, i.e. the user is able to focus the observation on a single data point. In addition, value frequency bars are available to present compact overviews. It also offers the possibility to define categories using context-sensitive rules and to assign colours to classes. The direct implementation of a broad range of descriptive and advanced statistics together with a hierarchical sorting feature permits user-defined exploration of the data. Thus CVGdb with d-matrix have been highly valuable tools to store and mine phenotype data of patients analyzed in the following, as well as to integrate these with obtained molecular data.

BMC Bioinformatics

BioMed Central

Software

Open Access

# d-matrix – database exploration, visualization and analysis

Dominik Seelow[1], Raffaello Galli[1], Siegrun Mebus[2], Hans-Peter Sperling[1,2], Hans Lehrach[1] and Silke Sperling*[1]

Address: [1]Vertebrate Genomics, Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany and [2]Pediatric Cardiology, German Heart Center, Augustenburger Platz 1, 13353 Berlin, Germany

Email: Dominik Seelow - dominik.seelow@web.de; Raffaello Galli - galli@molgen.mpg.de; Siegrun Mebus - mebus@dhzb.de; Hans-Peter Sperling - sperling@dhzb.de; Hans Lehrach - lehrach@molgen.mpg.de; Silke Sperling* - sperling@molgen.mpg.de

* Corresponding author

Published: 28 October 2004

*BMC Bioinformatics* 2004, **5**:168 doi:10.1186/1471-2105-5-168

Received: 23 August 2004
Accepted: 28 October 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/168

© 2004 Seelow et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** Motivated by a biomedical database set up by our group, we aimed to develop a generic database front-end with embedded knowledge discovery and analysis features. A major focus was the human-oriented representation of the data and the enabling of a closed circle of data query, exploration, visualization and analysis.

**Results:** We introduce a non-task-specific database front-end with a new visualization strategy and built-in analysis features, so called d-matrix. d-matrix is web-based and compatible with a broad range of database management systems. The graphical outcome consists of boxes whose colors show the quality of the underlying information and, as the name suggests, they are arranged in matrices. The granularity of the data display allows consequent drill-down. Furthermore, d-matrix offers context-sensitive categorization, hierarchical sorting and statistical analysis.

**Conclusions:** d-matrix enables data mining, with a high level of interactivity between humans and computer as a primary factor. We believe that the presented strategy can be very effective in general and especially useful for the integration of distinct data types such as phenotypical and molecular data.

## Background

d-matrix, originally designed with cardiovascular clinical and molecular genetic data in mind, is a generic database front-end that can be used to explore, visualize and analyze different typologies of datasets.

Both the generation and the analysis of genome, transcriptome and proteome data are becoming increasingly widespread, and these data must be merged to generate a molecular phenotype. Moreover, the correlation between molecular and phenotypical data requires acquiring both with comparable profoundness leading to the development of large and small scale databases holding both information [1-3]. In the same line, we developed a CardioVascular Genetic database (CVGdb), storing the detailed clinical phenotype of patients with congenital heart diseases as well as molecular data such as gene expression analysis results [4] and genotypes. However, querying and analyzing the stored data to uncover the valuable information hidden in the databases are difficult tasks. With some exceptions, these are approached by a two-step procedure, in which a database specific front-end serves the query and extraction of data, which are subsequently imported in stand-alone analysis tools for visualization, mining and statistics [5-11]. Moreover, the visualization and mining tools frequently focus on presenting overall views of data sets for a specific task and seldom permit single-case addressability or have drill-down capability. In today's systems, the perceptual abilities of human users are only used to a limited extend. We believe that it is essential to make users part of the overall process through computer support of their intelligence, creativity and perceptual abilities. Hence, a major research challenge is to find human-oriented forms of representing information and enabling rapid interaction between humans and computers in the query, visualization and analysis process [12].

It is not the purpose of this paper to survey the various solutions available to query, visualize and mine data, but rather to illustrate how such concepts could be combined usefully within one software tool. Here, the layout should not only preserve the structure of the information, it should also convey the quality of the distribution of the values contained in the database. The features of the display should then be designed to highlight those regularities, patterns or dependencies that are not easily detectable with an ordinary front-end.

One visual representation, which motivated the graphical display of the tool we describe here, is the data matrices handled in microarray studies, in which rows in the matrices typically represent genes and columns individual samples [13]. Rather than showing a numerical 'spreadsheet', it is convenient to display microarray data in such matrices, which indicate varying expression levels in a grid of varying colors.

With d-matrix we propose a generic front-end solution capable of extracting, exploring, visualizing and analyzing complex data. The software can be interfaced with the most common relational database management systems without any intervention on the schema or pre-processing phase. As the name suggests, the visual model proposed has the form of a matrix. Its elements are boxes whose colors show the quality of the underlying information. The granularity of the data display allows consequent drill-down, i.e. the user is able to focus the observation on a single data point. In addition, value frequency bars are available to present compact overviews. It also offers the possibility to define categories using context-sensitive rules and to assign colors to classes. The direct implementation of a broad range of descriptive and advanced statistics together with a hierarchical sorting feature permits user-defined exploration of the data.

## Implementation

### *Data Model*

The process of developing a uniform web interface for disparate data sources is a complex task because of the variability in the data models that underlie each source. To enable an effective two-dimensional display, the d-matrix model consists of a three-level tree. For the representation of a large database schema requiring a higher number of levels, several d-matrix instances can be built on the same database.

Within the proposed model, the main table addressing the objects of a study is considered as the root, the first level of the tree. The second level consists of tables that are joined with the root by means of its primary key and the third level consists of tables that are further joined with the ones at level two. In particular, the dependency of the root table with the second level tables can be either one-to-many or one-to-one, while the dependency of the second level table with the third level ones can be either many-to-one or one-to-one. To apply different query and visualization rules each branch of this tree is defined as a data group characterized by the same storing strategy. In cases where Entity-Attribute-Value (EAV) tables are interfaced, the Entity must correspond to the main ID.

As an example, we can refer to the CardioVascular Genetics database (CVGdb) schema set up by our group (Figure 1). Here, we selected the table Patients as the root of the tree, so that the CVGdb instance main ID is the Patients primary key. This selection is arbitrary and one could also choose Clones or Hybridizations, thereby focusing on different aspects of the overall dataset. In Figure 2, data groups and tree levels are represented. The data groups 2 to 4 address the EAV tables Invasive_Treatments, Medications and Samples; the groups 5 to 6 contain the same table Clones joined with different tables containing gene expression analysis results [4]; whereas the last group is built by two tables describing sequence variations (SV).

### *Data selection and query*

The schema is presented to the user in a structure recalling a file system selector (Figure 3A). Nodes represent attributes or value attributes that can be optionally divided further into more folders without any depth limitation. Collecting nodes in visually distinct entities becomes a necessity when coping with a large number of attributes. To obtain a quantitative measure of the information that is contained within groups of nodes, a *summary node* can be included in the query form. For each value of the x-axis, the values of the summary nodes are computed by counting the nonempty nodes in the respective folders.

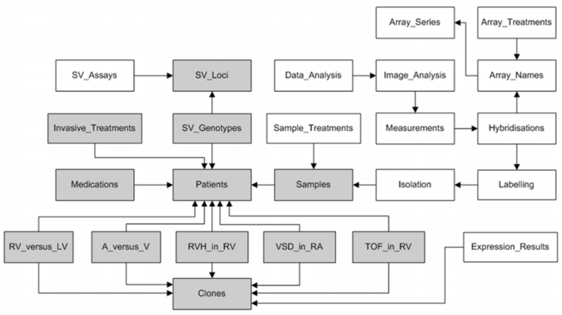


**Figure 1**
**Relational database schema of the CardioVascular Genetics database (CVGdb)** Tables are represented as boxes and foreign keys constraints as arrows. Grey boxes mark the schema subset interfaced in d-matrix. SV – Sequence Variations; RV_versus_LV, A_versus_V, RVH_in_RV, VSD_in_RA and TOF_in_RV are tables containing gene expression results [4].

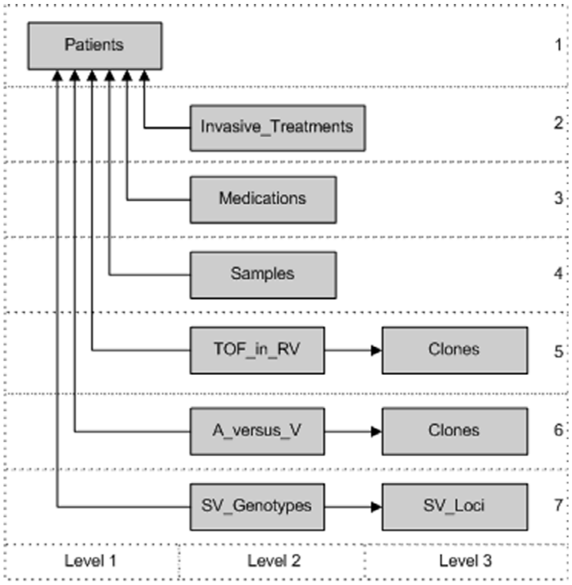


**Figure 2**
**Data Model** Shown is an excerpt of the three-level tree for interfacing d-matrix with CVGdb. The table "Patients" defines the root of the tree. Each branch refers to a defined data group consisting of one or two tables, respectively.

The query process consists of two steps. First, the users select all nodes they want to be included in the query (Figure 3A); second, these are listed in a query form where conditions and analysis features can be specified (Figure 3B). To visually distinguish between nodes referring to data belonging to single-table data group and two-table data groups, single-table group nodes are represented as sheet-like-icons, whereas two-tables data group nodes are represented by double-arrow-like icons: diagonally oriented for the nodes that belong to the second level tables, and vertically oriented for the nodes that belong to the third level tables (Figure 3A). The attribute on which the query display shall be focused can be selected by means of the three-banded icons placed on the right side of the nodes. For each of the nodes, the query form permits the definition of sorting order and direction (ascendant/ descendent), values and operators for query conditions, display order and parameters for statistical evaluations. The value cell is not shown if the node itself is an attribute value. Alternatively to the matrix view of the query result, the user can optionally export the resulting dataset in form of text or XML (Figure 3B).

### *Data visualization*

The graphical output of d-matrix consists of two-dimensional matrices, whose colored boxes code the meaning of the underlying information, the description of the chosen nodes and a prospect of statistical evaluations (Figure 4). The display of the data is determined by the data dimensionality. The main ID corresponds always to the x-axis of the matrix. To permit the display of single and multiple dependencies with the main ID, the y-axis shows either node descriptions or node values.

In cases of single dependency each data point is represented by one box of the matrix. If there is a multiple dependency (two-table data groups), subsequently more rows for each value of a single node are displayed. EAV data groups can lead to both single or multiple dependency; in the second case the entries are aggregated in one matrix box. In Figure 4 the tuples of the data group "PHENOTYPES" addressing the table Patients are displayed in the first matrix. Each tuple corresponds to a column whereas row headers are node descriptions. The tuples of the data group "SEQUENCE VARIATIONS" addressing the tables SV_Genotypes and SV_Loci are aggregated column-wise and grouped by the main ID. Here, there is more than one tuple for each column whereas row headers are values of the node Locus ID. Hence, each column of boxes on the matrix display represents an aggregation of more than one tuple of the query result. Following data mining terminology, we can say that in d-matrix *cases* (and aggregations of them) are represented column-wise.

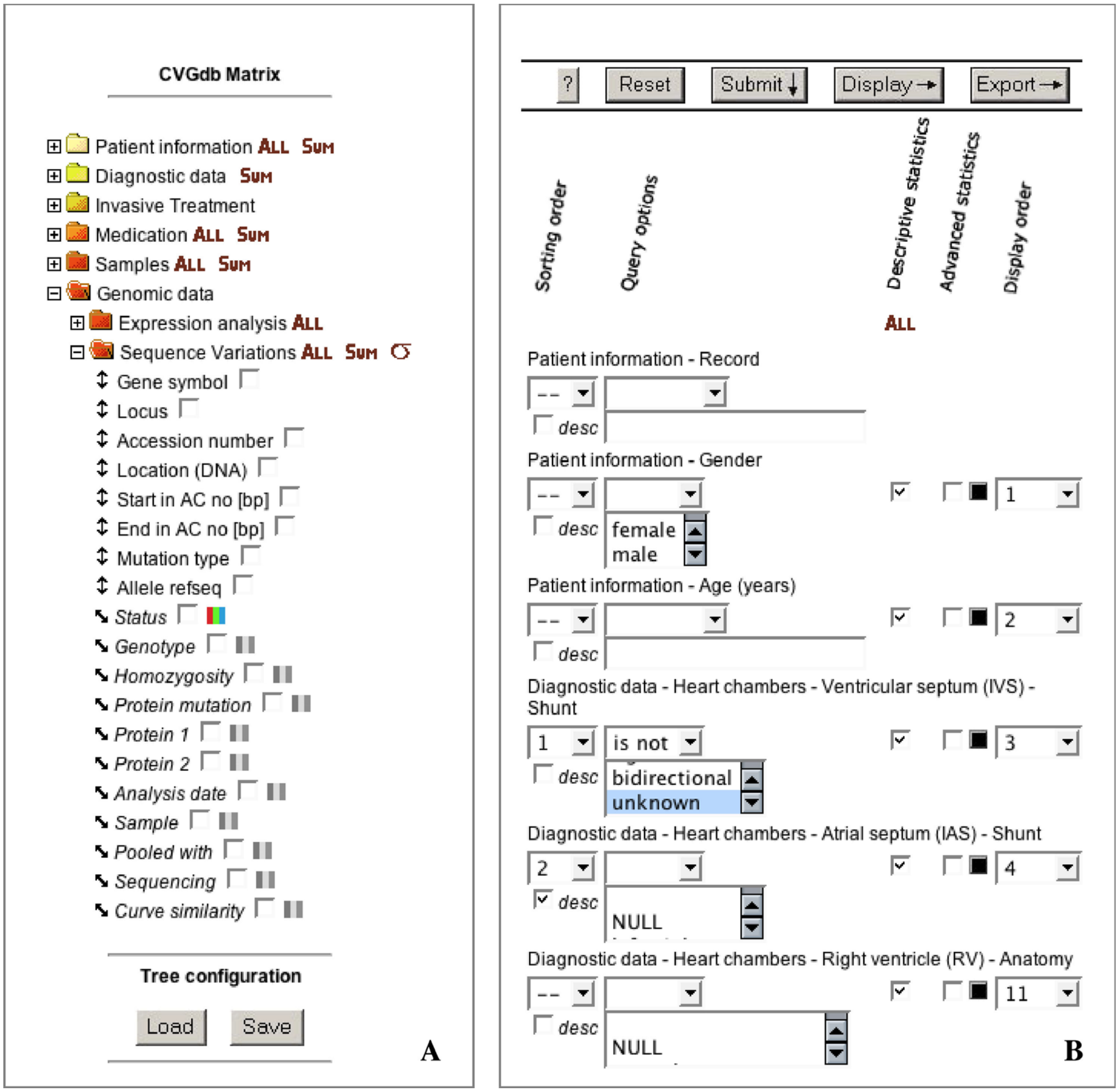
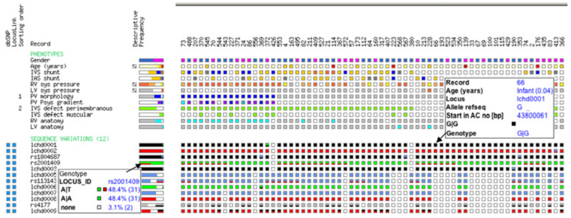


**Figure 3**

**Data selection and query** Within the data selection schema (A) users can choose all nodes they want to be included in the query. If a data group consists of two tables, the nodes are represented by vertical arrows for the first table and diagonal arrows for the second. The attribute on which the query display is focused can be selected by the three-banded icons, which switch from black-white to color and vice versa upon selection. Furthermore, trees can be saved and reloaded for subsequent analysis. Upon selection all nodes are listed in a secondary form (B), where query conditions, display and sorting order as well as the implementation of descriptive and advanced statistic can be specified. In addition to the graphical output, the query can be exported as a text of XML file.

When the matrix oversize the available space, the use of two distinct scrollbars lets the user move the data matrix horizontally and vertically. The general overview is given together with the advantage of single-case addressability,



| | |
|---|---|
| Record | 66 |
| Age (years) | Infant (0.04) |
| Locus | lchd0001 |
| Allele refseq | G |
| Start in AC no [bp] | 43800061 |
| G\|G | ■ |
| Genotype | G\|G |

| Genotype | | |
|---|---|---|
| LOCUS_ID | | rs2001409 |
| A\|T | 🟩🟥 | 48.4% (31) |
| A\|A | 🟩 | 48.4% (31) |
| none | ☐ | 3.1% (2) |

**Figure 4**

**Graphical output of d-matrix** The graphical output consists of the matrices itself, the description of the nodes displayed, a prospect of statistical evaluations and hyperlinks to external resources. Each matrix corresponds to a single data group (Phenotypes; Sequence variations). The x-axis of the matrix is defined by the main ID (Record) and the y-axis by the nodes displayed. The terms like "Gender", "Age (Years)" and "IVS Shunt" are descriptive names for the respective column names GENDER, AGE_YEARS and IVS_SHUNT of table PATIENTS; terms like "lchd0001" and "lchd0002" refer to locus names, values of the column LOCUS_ID of table SEQ_VAR_LOCI. The matrix is built by colored boxes coding for the meaning of the information itself, which is further described in the pop-up window (as shown for Record 366 and lchd0009). Frequency bars and boxes for descriptive statistics are displayed. Numbers are reflecting the sorting order, whereas blue boxes at the left border hold the hyperlinks.

i.e. each case (tuple) representation is entirely visible and its components clearly distinguishable.

The display is obtained as a group of images (generated using the Perl GD module and stored as temporary files), each in a separate HTML DIV container, which can be moved independently.

### *Drill-down*

The matrix display represents a summarized view of the query. Each box holds three levels of detail: first, the coordinates that uniquely identify the box position and represent two units of information; second, the color that corresponds to either a single value or a category; third, the hidden content of the box obtained by drill-down, which gives all remaining information for that box.

In the d-matrix display the drill-down can be obtained for each box in form of a pop-up window (Figure 4). The content structure of this new window varies according to the data group to which the box belongs, although it always contains the value that is substituted by its color code together with the underlying node description. Further supplementary data can be included from attributes of the same data group.

It is possible to add further detail by the mean of hyperlinks to grant access to remote databases, external analysis results and multimedia documents (Figure 4), or even to trigger further analysis processes.

### *Schema interface and configuration*

The software requires four configuration files: the data definitions file that is needed to connect d-matrix with the relational schema, a database settings file storing the information to access the database, a color file for the definition of the colors used in the matrix and a general server settings file. Every configuration file is maintained as plain text to permit easy access and modification.

The structure of the data definitions file must reflect the hierarchy in which the metadata (relational schema definition) have to be organized on the screen, while its textual content depicts a level of abstraction (*definitional abstraction*) [14] between the database physical representation and the human-comprehensible view of the data. Therefore, the data definitions file reflects the subdivision of the database schema in data groups. For each group the table attributes, information about identifiers, joining conditions as well as aggregation (where needed), display settings and the content of the pop-up window have to be defined. User-defined human-intelligible terms can be assigned for any term used in the database. Besides the attributes' names, types and descriptions, it is possible to define categories, orderings and associations with colors. It is important to notice that the rules that define categories can even involve other attributes of the same data group. This context-sensitive categorization, intended as a *qualitative abstraction* [14], allows the concurrent representation of two layers of information.

For each attribute value, value range or defined category, rules can be given to assign its respective color. This leads to a common method to visualize both discrete and continuous variables. In addition, categorized numeric values can be treated as categorical in specific contexts like sorting and statistics. Furthermore, colored boxes can be composed by combining the values of two nodes, which enables, for example, the visualization of both Alleles within horizontally split boxes for sequence variations (Figure 4).

Several data definitions files (each defining a separate d-matrix instance) can independently coexist on the same server for the same or different database systems and schemata.

### *Visual data mining and statistical analysis*

d-matrix permits consecutive data-filtering operations that – as a whole – can be seen as a single user-driven data mining session. A compact and information-dense graphical outcome, context-sensitive categorization, hierarchical sorting and drill-down enable this mining process. Frequency bars give an overview of the overall queried dataset whereas box plots improve the visual perception of the data distribution. A key feature within the mining process is the opportunity to obtain different views of a single data set rapidly in parallel using different browser windows. Here, the interactivity becomes a primary factor and is supported by the human-oriented representation.

A wide range of descriptive statistics and statistical tests is directly accessible. This permits statistical evaluation of the correlation between attributes and determination whether it is reasonable or not to assume that a sample fits to a specific distribution. For numerical values it is possible to perform up to ten different statistical tests, while for non-numerical entities (Boolean and categorical data) the Chi-square and Fisher exact tests are available. The user interface automatically performs a selection of attributes and tests according to their respective compatibility.

In addition to directly implemented tests, external data analysis environments like R [15] or user defined routines can be easily interfaced. The results of the tests, together with the descriptive statistics, are displayed at the side of the matrix and colors of the boxes reflect the results (e.g. significance) of the tests.

### CardioVascular Genetics database (CVGdb)

For interfacing d-matrix with the CVGdb, we assigned categories if appropriate and colors to more than 700 nodes. Figure 5 shows an example of a single user-driven data mining session, which was initiated with the aim to discover cardiac phenotype features associated with shunts abroad the interventricular septum (IVS shunt). Therefore, the only query condition specified is that "IVS shunt" is not "NULL". This condition is fulfilled by 211 out of 560 IDs stored to date. In addition, a subset of nodes referring to phenotype descriptions physically surrounding the interventricular septum has been chosen to be displayed. To structure the display, hierarchical sorting has been applied to the 'IVS shunt' and an arbitrary selection of other nodes. Viewing the entrance matrix (Figure 5A), one could easily recognize data clusters such as the relation of the category 'bidirectional' of the 'IVS shunt' (blue boxes) to categories of interatrial septum shunts (IAS shunt) and right ventricular systolic pressure 'RV sys pressure'. Almost all patients with a bidirectional 'IAS shunt' are also characterized by a bidirectional 'IVS shunt'. Furthermore, the majority of bidirectional 'IVS shunt' is associated with severe 'RV sys pressure', whereas the non-sorted nodes pulmonary valve morphology (PV morphology), pulmonary valve systolic pressure gradient (PV Psys gradient) and right ventricular anatomy (RV anatomy) are distributed in a questionable co-occurrence to each other in this first matrix. For further evaluation, we focus on the 'RV anatomy' or the 'PV morphology' chosen as the first sorted nodes in the second and third matrix (Figure 5B,5C), respectively. By using the tree save/reload option to retrieve these new matrices, only the sorting criteria needed to be modified to obtain different views on the same data set in parallel using three browser windows. Hence, the frequency bars remain the same in all visualization sessions. Now it becomes clear that more than half of the patients with infundibular stenosis (RV anatomy) show a stenotic 'PV morphology', which by itself is highly associated with an extreme 'PV Psys gradient'. Applying the correlation analysis implemented in d-matrix, the significance of the correlation of the 'PV Psys gradient' with the 'RV sys pressure' could be verified (Figure 5D). The described data are available for the exploration using d-matrix at the web supplement.

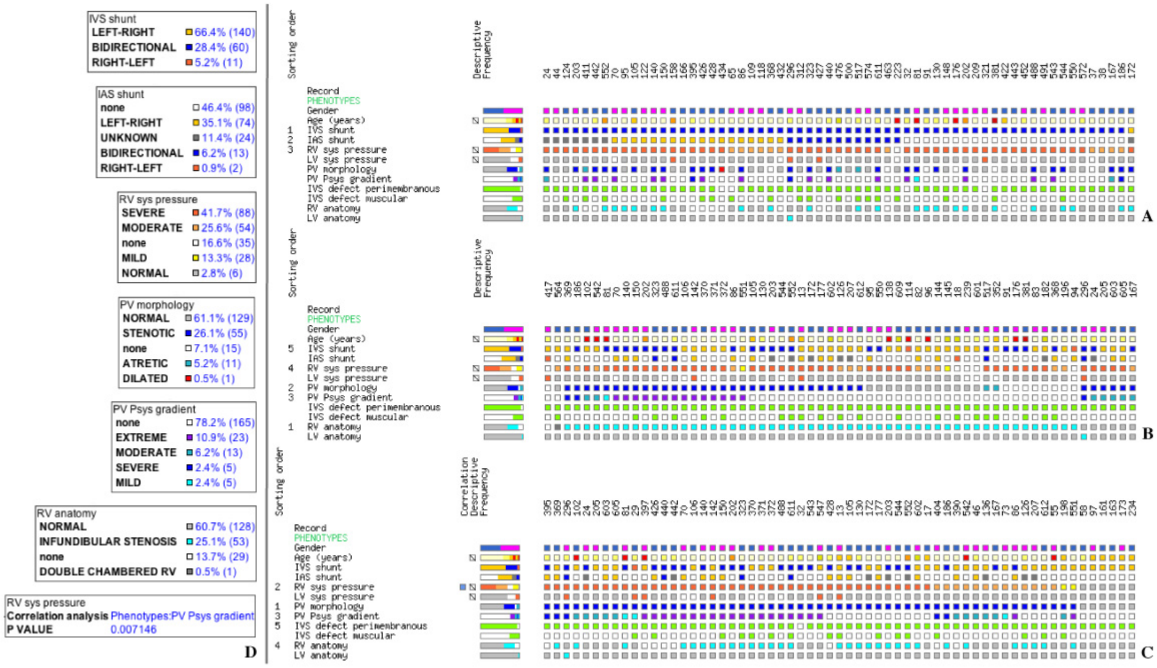
Finally, the session explained is just one out of several examples in which d-matrix proved to be highly effective for the visualization of regularities and dependencies within the CVGdb data. Moreover, based on the general visualization concept, d-matrix provides an integration between clinical and genetic information that is crucial for the correlation of phenotypical and molecular data (Figure 4).

### Other applications

With respect to an ongoing project on gene regulation, we found it very convenient to visualize potential transcription factor binding sites (TFBS) in promotor sequences by interfacing d-matrix [16]. Here, the nucleotides are used as the main ID (x-axis) and the TFBS are consequently displayed at the y-axis. This allows a much higher level of interactivity than a usual figure output. One could easily have different views of the data set by sorting or parallel display of different information, like color coded core or matrix match similarities.

To demonstrate the versatility of the software, we further interfaced d-matrix with a database that represents the periodic table of the elements [16]. Although we did not expect unusual or unexpected regularities in such a simple case, it was easy to obtain a matrix that shows the well-known dependency between Atomic Number, Atomic Mass and Energy Levels and the obvious lack of available information about elements with seven energy levels, which are the most unstable and rare.

The interfacing with both dataset required only one working day for each.



| IVS shunt | |
|---|---|
| LEFT-RIGHT | 66.4% (140) |
| BIDIRECTIONAL | 28.4% (60) |
| RIGHT-LEFT | 5.2% (11) |

| IAS shunt | |
|---|---|
| none | 46.4% (98) |
| LEFT-RIGHT | 35.1% (74) |
| UNKNOWN | 11.4% (24) |
| BIDIRECTIONAL | 6.2% (13) |
| RIGHT-LEFT | 0.9% (2) |

| RV sys pressure | |
|---|---|
| SEVERE | 41.7% (88) |
| MODERATE | 25.6% (54) |
| none | 16.6% (35) |
| MILD | 13.3% (28) |
| NORMAL | 2.8% (6) |

| PV morphology | |
|---|---|
| NORMAL | 61.1% (129) |
| STENOTIC | 26.1% (55) |
| none | 7.1% (15) |
| ATRETIC | 5.2% (11) |
| DILATED | 0.5% (1) |

| PV Psys gradient | |
|---|---|
| none | 78.2% (165) |
| EXTREME | 10.9% (23) |
| MODERATE | 6.2% (13) |
| SEVERE | 2.4% (5) |
| MILD | 2.4% (5) |

| RV anatomy | |
|---|---|
| NORMAL | 60.7% (128) |
| INFUNDIBULAR STENOSIS | 25.1% (53) |
| none | 13.7% (29) |
| DOUBLE CHAMBERED RV | 0.5% (1) |

| RV sys pressure | |
|---|---|
| Correlation analysis | Phenotypes:PV Psys gradient |
| P VALUE | 0.007146 |

**Figure 5**

**Example of a data exploration session for CVGdb** Shown are the first 61 of 211 records that meet the query condition "IVS shunt" is not "Null" focusing on different views of the data given by different sorting options (A, B, C). To provide information about the color code as well as the overall query output, pop-up windows for frequency bars of sorted nodes are shown (D). Further, the pop-up window for the correlation analysis between 'RV sys pressure' and PV Psys gradient' is displayed (D). See text for detailed description of the observed cluster.

## Results and discussion

We have presented d-matrix, a non-task-specific database front-end with a new visualization strategy with embedded analysis features.

The graphical outcome of d-matrix consists of colored boxes arranged in matrices; it permits single-case addressability with further drill-down capability. Together with the hierarchical sorting and statistical feedbacks, d-matrix enables consecutive data-filtering operations that – as a whole – can be considered as a single data mining session. Also, the result of such a session can be exported for further study. For a qualitative evaluation of d-matrix, one should not only focus the attention on the final display, which only represents the end product of a sequence of user-driven data exploration sessions. The high level of interactivity that our approach offers is indeed a primary factor; with d-matrix, the communication between human and computer is a rapid interaction.

The future development of d-matrix will focus on the implementation of clustering algorithms to be executed before display. Furthermore, we envisage the design of instruments to inquire metadata to maximize the quantity of information that will be eventually displayed and analyzed [17]. In addition, a user-friendly way to interact with configuration files will be granted by specific CGI scripts leading to a further reduction of the time to interface d-matrix with relational schemata.

An inquiry of the solutions reported to date for data exploration, visualization and analysis resulted in an approximate distinction between reports about efforts for database development with their task specific front-end solutions and stand-alone data analysis, visualization and mining tools. In our view, d-matrix stands in between those two groups and aims to combine features of both efforts, which we believe can be very effective and useful in general and especially for the association of distinct data types such as phenotypical and molecular data. As a front-end, it does not require complex installation processes or maintenance, and it is suitable for multi-user remote access. As a visual data mining tool, it gives an effective display that allows the detection of exceptions, trends, regularities, clusters and dependencies, as well as incomplete or erroneous data.

## Availability and requirements

**Project name:** d-matrix

**Project home page:** http://www.molgen.mpg.de/~heart/index_dmatrix.html

**Operating system(s):** Platform independent

**Programming language:** Perl

**Other requirements:** d-matrix was successfully interfaced to Oracle 8i, MySQL, Microsoft Access and text-based databases and is compatible with recent JavaScript-enabled browsers.

**License:** d-matrix is available on request from the author. To academic institutions d-matrix is available for a fee of 250 Euro that is intended to cover our costs of distribution and maintenance.

## Authors' contributions

DS developed the first generation of d-matrix and carried out the main programming work. RG is the current maintainer and carried out the main implementation. SM and HPS participated in the design, testing and quality control. HH participated in the conceptual design. SS conceived the development of d-matrix, managed and participated in its design and implementation.

## Acknowledgements

We thank Ulrike Klehmet for testing d-matrix at CVGdb as a non-biomedical scientist and Anja von Heydebreck for insightful discussions. We gratefully acknowledge the support of this work from the Max-Planck-Society and the BMBF grant 01GR0155.

## References

1. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ: **HGVbase: a curated resource describing human DNA variation and phenotype relationships.** *Nucleic Acids Res* 2004, **32(Database issue):**D516-519.
2. **Genome Web** [http://www.hgmp.mrc.ac.uk/GenomeWeb/]
3. Nadkarni PM: **The challenges of recording phenotype in a generalizable and computable form.** *Pharmacogenomics J* 2003, **3(1):**8-10.
4. Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S: **Genome-wide array analysis of normal and malformed human hearts.** *Circulation* 2003, **107(19):**2467-2474.
5. Walker AJ, Cross SS, Harrison RF: **Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique.** *Lancet* 1999, **354(9189):**1518-1521.
6. Falkman G: **Information visualisation in clinical Odontology: multidimensional analysis and interactive data exploration.** *Artif Intell Med* 2001, **22(2):**133-158.
7. Shao Q, Li Y, Campbell E, De Boer ES, Laginestra E, Statzenko A: **Statistical visualization for data exploration: a case study on Sydney Olympic Park.** *Chemosphere* 2003, **52(9):**1601-1614.
8. Gilbert DR, Schroeder M, van Helden J: **Interactive visualization and exploration of relationships between biological objects.** *Trends Biotechnol* 2000, **18(12):**487-494.
9. Grinstein G, Trutschl M, Cvek U: **High-Dimensional Visualizations.** In: *7th Data Mining Conference-KDD 2001: San Francisco, California* 2001.
10. Wegman EJ: **Visual data mining.** *Stat Med* 2003, **22(9):**1383-1397.
11. Rost U, Bornberg-Bauer E: **TreeWiz: interactive exploration of huge trees.** *Bioinformatics* 2002, **18(1):**109-114.
12. Keim D, Kriegel HP: **VisDB: Database Exploration using Multidimensional Visualization.** In: *IEEE Computer Graphics and Applications: 1994* 1994:40-49.
13. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95(25):**14863-14868.
14. Lavrac N, Keravnou E, Zupan B: **Intelligent Data Analysis in Medicine.** *Volume 42*. New York: Marcel Dekker; 2000.
15. Ihaka R, Gentleman R: **Language for Data Analysis and Graphics.** *J of Comp and Graphical Stats* 1996, **5:**299-314.
16. **d-matrix web supplement** [http://www.molgen.mpg.de/~heart/index_dmatrix.html]
17. Weiner M, Sherr M, Cohen A: **Metadata tables to enable dynamic data modeling and web interface design: the SEER example.** *Int J Med Inform* 2002, **65(1):**51-58.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

### 2.1.2 Genome-wide expression profiling of congenital heart disease

> Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, **Sperling S**. Genome-wide array analysis of normal and malformed human hearts. *Circulation* 2003;**107**:2467-2474.

Normal cardiac development in humans leads to a four chambered heart, defining the basis of normal cardiac physiology. The right heart belongs to the low-pressure pulmonary system and the left heart to the high-pressure body circulation. As a consequence of heart malformations, abnormal hemodynamic features can occur due to volume or pressure overload and lead to an adaptation process of the heart. So far, knowledge about the molecular pathways involved in this process has been mainly gained by animal studies (Bauer et al., 1998; Nediani et al., 2000; Baumgarten et al., 2002).

A genome-wide gene expression study of congenitally malformed hearts in human was performed, with the purpose to ultimately identify genes associated with dysdevelopment as well as genes involved in the adaptation processes of the heart. The defined collection of patient material in cooperation with the Deutsches Herzzentrum Berlin, allowed the selection of a balanced patient population enabling the separation of disease or tissue specific expression patterns. Considering known confounding factors like age and gender, 40 individuals were selected for the analysis. All patients were clinically studied, and their hemodynamic state was worked up by cardiac catheterisation and echocardiography prior to surgery.

In total, 6 normalization samples and 55 patient samples have been studied. The latter belonged to four categories: (A) Right atrial (n=2) and ventricular (n=9) samples of patients with classical Tetralogy of Fallot (TOF), exposed to pressure overload resulting in right ventricular hypertrophy. (B) Right ventricular samples of a patient population (RVdis, n=7) with right ventricular hypertrophy in response to pressure overload and a variety of cardiac malformations. These samples were analyzed together with right ventricular samples of TOF as one sample population (RVH) to elucidate the common biomechanical adaptation processes. (C) Right atrial samples of hearts with ventricular septal defect (VSD, n=4), which have not been exposed to biomechanical stress. (D) To identify chamber-specific genes and to provide a framework of interpretation, normal right and left atrial (RA, n=4; LA, n=3), ventricle (RV, n=6; LV, n=6) and interventricular septum (IVS, n=6) samples were profiled as well as right atrial samples of patients with atrial septal defects (ASD, n=8). The normal RV, LV and IVS samples were obtained from 6 different individuals.

Pursuing a genome-wide approach, ~9 million measurements of gene expression were made. Characteristic expression patterns were obtained from a linear model analysis for the normal atria and ventricle, TOF, RVH and VSD. The comparison between normal right and left ventricular tissue (RV vs LV), using paired t-tests, was based on matched samples from 6 individuals.

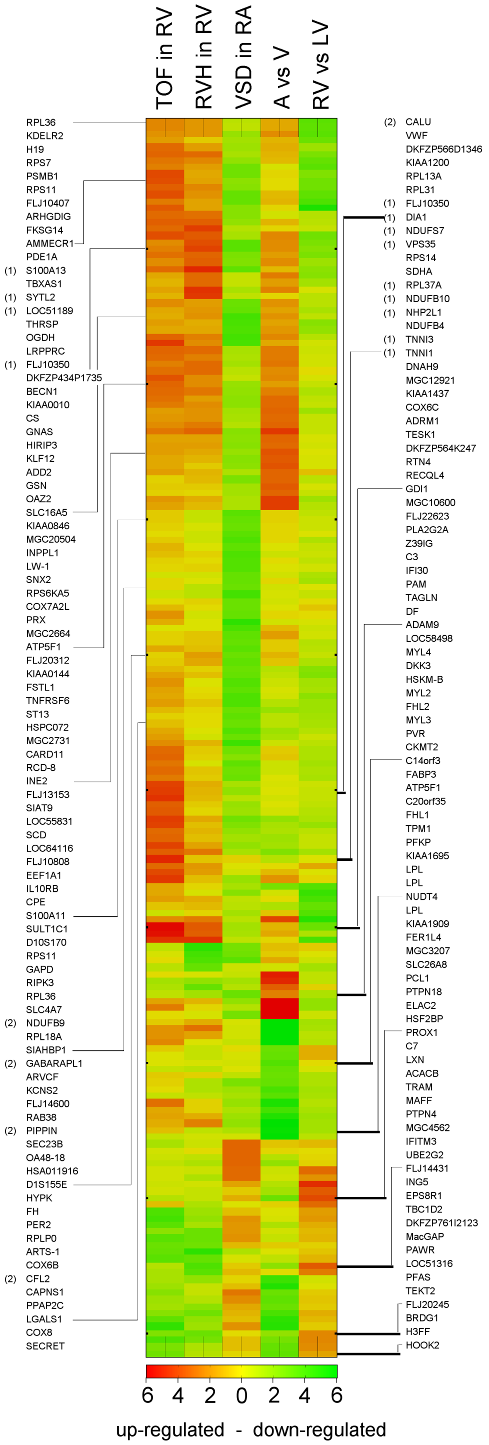


**Figure 2. Overview of probability values and expression levels for different phenotype comparisons.** Each comparison is represented by one column and each gene by one row. The –log10(P) of each gene in the particular comparison is colour-coded in yellow to red for upregulated and yellow to green for downregulated genes. Comparison Tetralogy of Fallot (TOF in RV), Ventricle Septal Defect (VSD in RA) and right ventricular hypertrophy (RVH in RV) versus normal heart tissue of the same location resp. right ventricle (RV) or atrium (RA). Comparison atrium versus ventricle (A vs V) and right versus left ventricle (RV vs LV)

### *Gene expression pattern and cardiac phenotype*

**Figure 2** provides an overview of the molecular signatures. It appears that about 25% of the CHD-associated genes (green or red in one of the first 3 columns) are not chamber-specific in the normal heart (yellow in the last two columns). Further inspection shows a partially similar expression dynamic of the TOF portrait with that of RVH, but an opposite expression dynamic compared to VSD. In addition, a large amount of genes characteristic for the molecular signature of VSD are not differentially expressed in any other analysis. Comparison of the expression profile of genes characteristic for TOF and RVH provides the possibility of subtracting both from each other and identifies genes specific for either TOF or RVH. A correspondence analysis (Fellenberg et al., 2001) using Gene Ontology categories was performed to provide a global view of the association of functional gene categories with particular phenotypes (**Figure 3**).

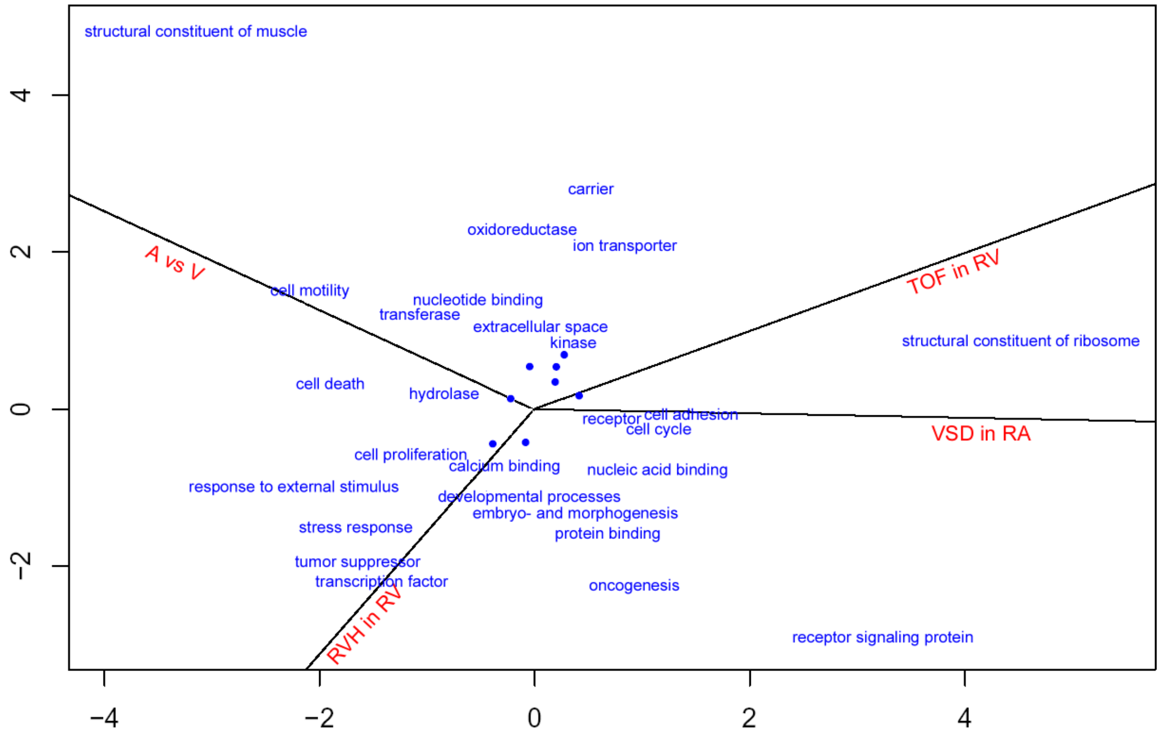


**Figure 3. Association of functional gene categories (blue) to studied phenotypes (red)** with respect to differential gene expression received by app. 9 million measurements. Biplot obtained from correspondence analysis. Categories that are not specifically associated with any phenotype are represented by dots.

### *Chamber-specific expression*

As the human heart consists of two general compartments - the atria and ventricle -, the molecular repertoire that each of these expresses was analyzed first. In addition to well known chamber-specific genes, like atrial and ventricular myosin light chains, this signature includes diverse previously unknown chamber-specific genes for muscle contraction, extracellular components, cell growth and differentiation and energy metabolism.

The less force-developing atria were characterized by higher expression of genes encoding proteins associated with extracellular matrix or actin-modulation like CST3 and PCOLCE. The translation factor EEF1A and the DNA helicase REQL4 were highly significant up-regulated in atria. EEF1A-2/S1 protein is activated upon myogenic differentiation and delays myotube death after apoptotic stress induction (Chambers et al., 1998). KCNIP2, a potential target for ventricular tachycardia (Kuo et al., 2001; Guo et al., 2002), is even higher expressed in the human atria than in the ventricle ($P$ = 0.01), pointing to a potential role also in atrial arrhythmia. Genes with expression levels higher in ventricle than in atria mainly belong to three major functional classes: cytoskeleton - contraction, metabolism - energy turnover and cell cycle - growth. Several of these genes are involved in ventricular myocardial disorders: TMP1 (Karibe et al., 2001) is mutated in human cardiomyopathies, FHL1 is down-regulated in failing human hearts (Yang et al., 2000), and ANKRD2 is involved in the process of cardiac hypertrophy (Pallavicini et al., 2001).

In addition to atrial and ventricular specifications, the molecular differences between the normal high-pressure left and low-pressure right ventricle was analyzed. In general, genes encoding proteins involved in cell cycle, cell differentiation and energy metabolism were discovered to be downregulated in the right ventricle compared to the left ventricle (e.g ALK3) (Gaussin et al., 2002). No significant difference between left ventricle and interventricular septum were observed.

### *Tetralogy of Fallot and RVH*

Distinct molecular portraits of TOF and RVH were observed with genes of various functional classes. Even though the right ventricular hypertrophy is part of TOF, a clear distinction between these two gene expression profiles could be made applying the statistical analysis once TOF and once TOF combined with the RVdis samples as one phenotype (RVH). Therefore, TOF reveals the molecular signature of the malformation in addition to the adaptation portrait.

Beside genes involved in cell cycle, a characteristic feature of the TOF signature is the up-regulation of ribosomal proteins: S6, L37a, S3A, S14 and L13A. A specific role of ribosomal proteins during cardiac development has only been described for the chick ribosomal protein L10, which is down-regulated in the cardiac outflow tract of chick embryos lacking neural crest cells (Kirby et al., 1995).

The expression data reveal a TOF-specific dysregulation of potential targets that could be involved in pathways leading to cardiac maldevelopment, for example SNIP, A2BP1 and are upregulated. SNIP interacts with Smad4, a mediator of TGF-ß, activin, and BMP-signaling, which are essential for normal cardiac development. A2BP1 belongs to a novel gene family sharing RNA-binding motifs expressed at the developing heart during mouse embryogenesis (Kiehl et al., 2001). KIAA1437 binds k-ras, where k-ras-deficient mice develop a thin ventricular wall and die until term (Garcia et al., 2000).

In RVH we observed a hypertrophy-specific gene expression pattern of genes mainly involved in stress response, cell proliferation and metabolism. Intriguing is the up-regulation of ADD2, whose relative ADD1 was recently shown to be associated with hypertension in human (Morrison et al., 2002). As the expression of several genes of the RVH signature was similar to their expression levels in LV, an analysis focusing on whether the molecular adaptation to pressure overload could lead to a molecular transition from right to left ventricular characteristics was performed. A significantly positive correlation coefficient of 0.27 ($P < 0.0001$, permutation test) was found indicating that the genes dysregulated in RVH have a tendency to behave similarly in the disease state as in normal LV tissue.

### *Ventricular septal defect*

To obtain a molecular portrait that is not influenced by biomechanical adaptation processes, right atrial samples of patients with VSD were studied, intact tricuspid valve and normal RA pressure. A VSD-specific molecular signature dominated by downregulated genes with respect to the other RA samples was observed. As seen in TOF, several ribosomal proteins (S11, L18A, L36, LP0, L31, and MRPS7) are differentially expressed, but here they are downregulated. Other VSD-specific genes encode ion transporters or function during vertebrate development. The differential expression of ion channels was restricted to solute and potassium channels (SLC26A8, SLC16A5, SLC4A7, KCNS2, KCNN3). A thorough literature study of downregulated genes in VSD revealed that a major part is involved in cell proliferation and differentiation during embryogenesis as well as apoptosis. Examples are AMD1 (Nishimura et al., 2002), RIPK3 (Sun et al., 1999), EGLN1 (Taylor, 2001) and ARVCF deleted in velo-cardio-facial syndrome (Sirotkin et al., 1997). The question whether the observed downregulated transcription mirrors a reduction of essential proteins leading to incomplete fusion, an underdeveloped atrium or an unknown physiological process, will have to be elucidated in the future.

Overall the findings demonstrate that the analysis of malformed human hearts using powerful techniques like microarrays combined with statistical methods opens a new window to understand cardiac adaptation and development. The disease-specific expression profiles point to disturbances in the underlying transcription network and prompted it further exploration as described in the following.

# Genome-Wide Array Analysis of Normal and Malformed Human Hearts

Bogac Kaynak, MSc; Anja von Heydebreck, PhD; Siegrun Mebus, MD; Dominik Seelow, MSc;
Steffen Hennig, PhD; Jan Vogel, BSc; Hans-Peter Sperling, MD; Reinhard Pregla, MD;
Vladimir Alexi-Meskishvili, MD, PhD; Roland Hetzer, MD, PhD; Peter E. Lange, MD, PhD;
Martin Vingron, PhD; Hans Lehrach, PhD; Silke Sperling, MD

***Background*****—**We present the first genome-wide cDNA array analysis of human congenitally malformed hearts and attempted to partially elucidate these complex phenotypes. Most congential heart defects, which account for the largest number of birth defects in humans, represent complex genetic disorders. As a consequence of the malformation, abnormal hemodynamic features occur and cause an adaptation process of the heart.

***Methods and Results*****—**The statistical analysis of our data suggests distinct gene expression profiles associated with tetralogy of Fallot, ventricular septal defect, and right ventricular hypertrophy. Applying correspondence analysis, we could associate specific gene functions to specific phenotypes. Furthermore, our study design allows the suggestion that alterations associated with primary genetic abnormalities can be distinguished from those associated with the adaptive response of the heart to the malformation (right ventricular pressure overload hypertrophy). We provide evidence for the molecular transition of the hypertrophic right ventricle to normal left ventricular characteristics. Furthermore, we present data on chamber-specific gene expression.

***Conclusions*****—**Our findings propose that array analysis of malformed human hearts opens a new window to understand the complex genetic network of cardiac development and adaptation. For detailed access, see the online-only Data Supplement. **(*Circulation*. 2003;107:2467-2474.)**

**Key Words:** heart defects, congenital ■ hypertrophy ■ atrium ■ ventricle ■ molecular biology

Congenital heart defects (CHDs) account for the largest number of birth defects in humans, with an incidence of ≈8 per 1000 live births. Heart formation requires complex interactions among cells originating from different cell lineages that arise from 3 distinct origins: cardiogenic mesoderm, neural crest, and proepicardium. Only a minority of CHDs are caused by single-gene defects, whereas most are thought to be multigenetic disorders. The heterogeneity of CHDs associated with single-gene defects, as demonstrated for NKX2.5 or TBX5 mutations,[1,2] makes mechanistic understanding of gene function challenging and points to a complex genetic network with modifier genes, genetic polymorphism, and the influence of environmental factors.

Normal cardiac development in humans leads to a 4-chambered heart, defining the basis of normal cardiac physiology. The right heart belongs to the low-pressure pulmonary system and the left heart to the high-pressure body circulation. As a consequence of heart malformations, abnormal hemodynamic features can occur because of volume or pressure overload and lead to an adaptation process of the heart. So far, knowledge about the molecular pathways involved in this process has been mainly gained by animal studies.[3–5]

Taking the above into account, we attempt a genome-wide gene expression study of congenitally malformed hearts in humans, with the purpose to ultimately identify genes associated with dysdevelopment as well as genes involved in adaptation processes of the heart. Our present data suggest that patients with congenitally malformed hearts can serve as a model to study these transcriptional fingerprints.

We examined and compared genes dysregulated in defined congenitally malformed hearts with the molecular response to pressure overload leading to hypertrophy and the chamber-specific cardiac molecular portrait. Finally, we provide a cardiac-specific clone selection representing genes persistently expressed in normal and diseased myocardium.

American Heart Association
Circulation
JOURNAL OF THE AMERICAN HEART ASSOCIATION

---

Received December 9, 2002; revision received February 26, 2003; accepted February 27, 2003.

From the Departments of Vertebrate Genetics (B.K., D.S., S.H., J.V., H.L., S.S.) and Computational Molecular Biology (A.v.H., M.V.), Max-Planck-Institute for Molecular Genetics, Berlin; and the Departments of Pediatric Cardiology (S.M., H.-P.S., P.E.L.) and Cardiovascular Surgery (R.P., V.A.-M., R.H.), German Heart Center Berlin, Germany.

**Futher information on obtained gene expression profiles, gene names, accession numbers, and clone IDs is available in an online-only Data Supplement at http://www.circulationaha.org.**

Correspondence to Dr Silke Sperling, Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany. E-mail sperling@molgen.mpg.de

© 2003 American Heart Association, Inc.

***Circulation* is available at http://www.circulationaha.org** **DOI: 10.1161/01.CIR.0000066694.21510.E2**

Taken together, our data provide the first genome-wide transcriptional fingerprint of functionally known and unknown genes expressed in the human heart at different cardiac conditions. For detailed access, we provide a Data Supplement, in which the obtained gene expression profiles can be searched by gene names, accession numbers, and clone IDs.[5a]

# Methods

## Samples

All cardiac samples were obtained from the German Heart Center at cardiac surgery after short-term cardioplegia, after ethical approval by the institutional review committee and informed consent of the patient or parents. Tissue from normal human hearts was obtained from unmatched organ donors without cardiac diseases, where the hearts could not be transplanted because of organizational difficulties. All samples were directly snap frozen in liquid nitrogen after excision and stored at −80°C. HEK 293 cells were harvested and used as a normalization sample.

## Array

In total, 61 Human Unigene Set-RZPD 2 cDNA arrays from the German Genome Resource Center[6] were used, each represented by 3 Hybond $N^+$ membranes containing polymerase chain reaction (PCR) products of 74 695 different IMAGE clones spotted in duplicate. At least one IMAGE clone of 33 947 Unigene clusters has been resequenced (see the Data Supplement).

## Sample Preparation and Array Hybridization

Total RNA of all cardiac tissues and HEK 293 cells was extracted using TRIzol reagent (Gibco BRL) according to the manufacturer's protocol. Labeling was performed by reverse transcription of 8 $\mu$g total RNA with avian myeoblastosis virus reverse transcriptase (AMV-RT; Promega) in the presence of $pd(T)_{12\text{-}18}$ (Amersham Pharmacia Biotech) and $\alpha^{33}$P-dCTP (Amersham Pharmacia Biotech). Unincorporated nucleotides were removed using ProbeQuant G-50 micro columns (Amersham Pharmacia Biotech), and cDNA was added to the hybridization solution together with salmon sperm DNA (Gibco BRL), placenta DNA, and $pd(A)_{40}$ (MWG Biotech). Hybridizations were performed at 65°C for 16 hours. After washing, arrays were exposed for 24 hours and scanned using a Fuji Film Bas-1800 reader (Fuji photo film). Image analysis was carried out using the X-Digitize image processing software.[7]

## Array Data Preprocessing

Data normalization was principally performed as described previously.[8] After local background subtraction, all intensity values below 200 were reset to this level and the intensities of each hybridization were scaled to a constant sum. Visual inspection showed that the variance of the resulting log-transformed intensities was approximately constant across the intensity range. Because the hybridizations were performed on arrays from 2 different production batches, which significantly affected the measurements, we corrected the influence of the production batch as follows. We considered 2 virtual reference hybridizations, defined by the spot-wise medians of 10 hybridizations from the respective array batch. These 2 sets of hybridizations were performed with patient samples as well as normalization samples sharing equal phenotype profiles. The logarithms of ratios between the intensities of each hybridization of interest and those of the respective batch-specific virtual reference hybridization were shifted such that the median over the 40% of spots with highest average log-intensity became 0. Finally, the obtained log-ratio (base 10) values were averaged over duplicate spots per clone, resulting in values that are referred to as normalized expression levels in the text. To limit the analysis to meaningful data that did not show consistently low intensity or low variance across the samples, all additional statistical analysis was performed on the normalized expression levels of 8069 selected clones, whose intensities were among the 15% highest in at least 4 hybridizations and whose natural log ratios had a standard deviation across the samples of at least 0.5.

## Differential Gene Expression Analysis

To account for the influence of major confounding factors, we used linear models incorporating the factors age, gender, disease, and tissue type as a statistical framework for identifying genes that are affected in particular phenotypes.[9] We subjected the normalized expression levels of each clone to a linear model of the following form: $y_{hijk} = \mu + A_h + S_i + T_j + D_k + \epsilon$, where $A_h$ is the effect of age h (young or old), $S_i$ the effect of the patient's gender, $T_j$ the effect of tissue j (right ventricle [RV] or left ventricle [LV], right or left atrium), and $D_k$ the effect of disease status k (atrial septal defect [ASD] II, TOF, ventricular septal defect [VSD], right ventricular hypertrophy [RVdis], and normal). The models were fitted with the least-squares method using the function lm() in the statistical software R,[10] with the constraints that the coefficients for the different levels of each factor sum up to 0. For all coefficients of interest, the null hypothesis of their being equal to 0 was tested under the assumption of independent errors following a normal distribution. Thus we obtained a probability value quantifying the statistical significance of differential expression for each gene and each phenotype comparison of interest. The reliability of the obtained results was additionally assessed by estimating the rates of falsely significant genes with a permutation approach. For this purpose, the linear models were fitted for 500 random permutations of the sample labels. For each coefficient of interest, the expected proportion of false positives among the significant genes (false discovery rate) was estimated from the randomized data.[11]

For comparison between normal RV and LV tissue, paired samples from 6 individuals were analyzed using *t* tests for paired observations.

## Multivariate Analysis

To analyze which phenotypical differences among the tissue samples are most pronounced in terms of gene expression patterns, we used the class discovery method ISIS.[12] ISIS searches for binary class distinctions among the set of tissue samples that are characterized by clear differential expression of a corresponding subset of genes.

For the standard visual display of gene expression matrices, genes were grouped by average linkage hierarchical clustering using the Euclidean distance as dissimilarity measure.

## Real-Time PCR

Real-time PCR assays were carried out using SYBR Green PCR Master Mix (Applied Biosystems) on ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) according to the manufacturer's protocol. Intron spanning primers (MWG Biotech) were designed using the Primer Express software (Applied Biosystems). Primer sequences are available in the Data Supplement. RT reactions were carried out via AMV-RT (Promega) with random hexamers (Amersham Pharmacia Biotech) with 1 $\mu$g total RNA. We analyzed the significance of the normalized $\Delta\Delta$CT values with *t* tests and considered genes with $P < 0.05$ as confirmed to be significantly differentially expressed. Data normalization was performed using HPRT as a housekeeping gene.

# Results

## Study Design

To allow the selection of a balanced patient population enabling the separation of disease- or tissue-specific expression patterns, we collected 150 samples by a standardized procedure during cardiac surgery over a time period of 3 years at the German Heart Center Berlin. Considering known confounding factors like age and gender, we selected 40 individuals for our additional analysis. All patients were clinically studied, and their hemodynamic state was worked

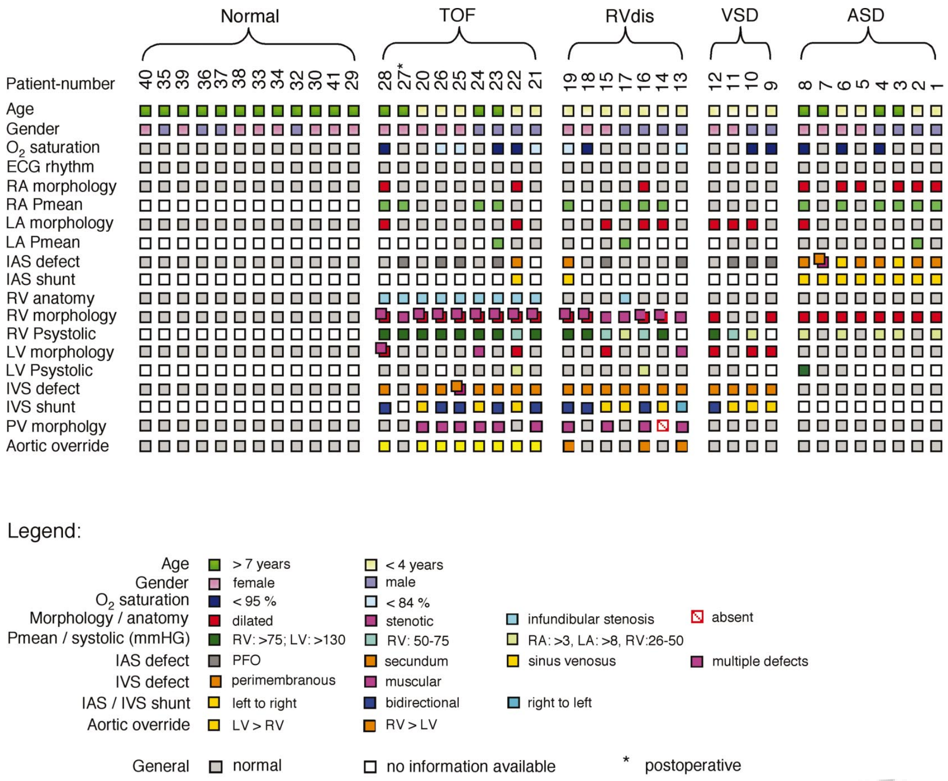


**Figure 1.** Phenotype matrix. Each individual is represented by one column. In addition to general information, all pathological features are indicated. See legend for color-code information. Double boxes are used for more than one information per row. IAS indicates intra-atrial septum; PV, pulmonary valve.

up by cardiac catheterization and echocardiography before surgery (Figure 1).

In total, 6 normalization samples and 55 patient samples have been studied. The latter belonged to 4 categories (Figure 1), as follows: (1) right atrial (RA, n=2) and ventricular (n=9) samples of patients with classical tetralogy of Fallot (TOF), exposed to pressure overload resulting in right ventricular hypertrophy; (2) right ventricular samples of a patient population (RVdis, n=7) with right ventricular hypertrophy in response to pressure overload and a variety of cardiac malformations, which we analyzed together with right ventricular samples of TOF as one sample population (RVH) to elucidate the common biomechanical adaptation processes; (3) RA samples of hearts with VSD (n=4), which have not been exposed to biomechanical stress; and (4) normal RA (n=4), left atrial (LA, n=3), RV (n=6), LV (n=6), and interventricular septum (IVS, n=6) samples as well as RA samples of patients with ASD (n=8); which we profiled to identify chamber-specific genes and to provide a framework of interpretation. The normal RV, LV, and IVS samples were obtained from 6 different individuals.

**TABLE 1. Overview of Differentially Expressed Genes ($P<0.01$)**

| Analyzed Samples | Data Set | TOF in RV* (9 of 22) | RVH in RV* (16 of 22) | VSD in RA* (4 of 18) | A vs V* (20 vs 20) | RV vs LV† (6 pairs) |
|---|---|---|---|---|---|---|
| Clones | | | | | | |
| Total (resequenced) | 8069 (3524) | 323 (142) | 198 (81) | 215 (105) | 438 (158) | 149 (59) |
| Estimated false discovery rate | ... | 14% | 25% | 23% | 12% | 34% |
| Unigene clusters | | | | | | |
| Total (resequenced) | 6059 (3523) | 264 (142) | 154 (81) | 178 (105) | 315 (158) | 117 (58) |
| Known genes | 1880 | 86 | 49 | 86 | 100 | 53 |

Total numbers of cDNA clones differentially expressed in different phenotype comparisons together with the represented nonredundant numbers of Unigene clusters and known genes. The estimated false discovery rate and the number of resequenced clones are presented. The used statistical method is indicated:

*Linear model; †Paired *t* test.

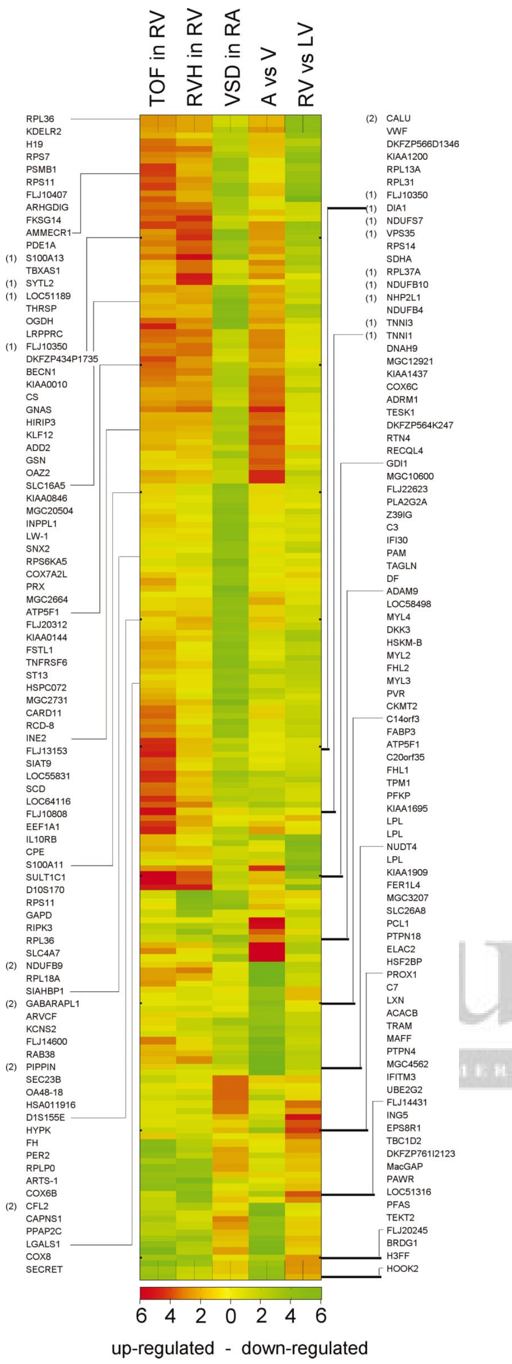


**Figure 2.** Overview of probability values and expression levels for different phenotype comparisons. Shown are all sequence-confirmed and annotated genes with $P<0.01$ in at least 1 comparison. Gene names are listed. Each comparison is represented by 1 column, and each gene by 1 row. The $-\log_{10}(P)$ of each gene in the particular comparison is color-coded in yellow to red for upregulated and yellow to green for downregulated genes. For example, a value of 2 stands for $P=0.01$ and is color-coded in red if the gene is upregulated. The differentially expressed genes confirmed by real-time PCR are indicated with 1 if tested for TOF in RV and 2 if tested for VSD in RA.

**TABLE 2. Comparison of Array Analysis and Real-Time PCR**

| Clone ID | Gene Name | Array Analysis<br>P | <br>Fold-Change | Real-Time PCR<br>P | <br>Fold-Change |
|---|---|---|---|---|---|
| TOF in RV | | | | | |
| IMAGP956P1758 | DIA | 0.001 | 1.6 | 0.01 | 1.5 |
| IMAGP956I2251 | FLJ10350 | 0.001 | 1.7 | 0.0007 | 2.7 |
| IMAGP956A2060 | LOC51189 | 0.006 | 1.6 | 0.0002 | 4.5 |
| IMAGP956N1316 | NDUFB10 | 0.005 | 1.5 | 0.02 | 1.5 |
| IMAGP956F1812 | NDUFS7 | 0.0006 | 1.6 | 0.003 | 2.3 |
| IMAGP956I042 | NHP2L1 | 0.005 | 1.7 | 0.04 | 1.4 |
| IMAGP956F0258 | RPL37A | 0.001 | 1.7 | 0.04 | 1.7 |
| IMAGP956K2159 | S100A13 | 0.002 | 1.6 | 0.01 | 2.4 |
| IMAGP956D2115 | SDHA | 0.0008 | 1.8 | 0.06 | 1.7 |
| IMAGP956H0722 | SYTL2 | 0.009 | 1.6 | 0.01 | 2.2 |
| IMAGP956G2255 | TNNI1 | 0.0004 | 2.7 | 0.01 | 24.5 |
| IMAGP956K2055 | TNNI3 | 0.004 | 1.7 | 0.0002 | 3.0 |
| IMAGP956E1912 | VPS35 | 0.004 | 1.5 | 0.001 | 2.9 |
| IMAGP956L2013 | VWF | 0.005 | 1.5 | 0.005 | 3.5 |
| VSD in RA | | | | | |
| IMAGP956G1159 | CALU | 0.001 | 0.4 | 0.01 | 0.6 |
| IMAGP956D0729 | CFL2 | 0.002 | 0.5 | 0.002 | 0.4 |
| IMAGP956D0830 | COX6B | 0.002 | 0.5 | 0.1 | 0.7 |
| IMAGP956C0130 | GABARAPL1 | 0.001 | 0.2 | 0.001 | 0.3 |
| IMAGP956F1716 | GSN | 0.009 | 0.4 | 0.07 | 0.6 |
| IMAGP956N0631 | NDUFB9 | 0.007 | 0.5 | 0.04 | 0.6 |
| IMAGP956A1215 | PIPPIN | 0.002 | 0.3 | 0.0006 | 0.1 |

Results of real-time PCR verification for 21 randomly chosen genes with $P<0.01$. Shown are $P$ values and fold-changes of genes associated with TOF and VSD compared with other samples from the same cardiac chamber. Clone-IDs and gene names are given.

Pursuing a genome-wide approach, we used the Human Unigene Set-RZPD 2 cDNA arrays. The set contains 74 695 different IMAGE clones belonging to 49 255 different Unigene clusters with 12 657 known genes. In total, ≈9 million measurements of gene expression were made. All gene expression data are available for download in the Data Supplement.

## Gene Expression Pattern and Cardiac Phenotype

Characteristic expression patterns were obtained from a linear model analysis for the following phenotype comparisons: (A versus V), the comparison of atria and ventricle, where we based our analysis on 40 samples from different individuals (20 atria and 20 ventricle); (TOF in RV), where we analyzed all 22 RV tissue samples for genes associated with TOF (9 samples); (RVH in RV), where the same 22 samples were studied for effects of right ventricular hypertrophy (16 samples); and (VSD in RA), where we analyzed 18 RA samples regarding the effects of VSD (4 samples) (Table 1). The comparison between normal right and left ventricular tissue (RV versus LV), using paired *t* tests, was based on matched samples from 6 individuals.

To verify our observed gene expression results, we investigated expression levels of 21 randomly chosen genes with

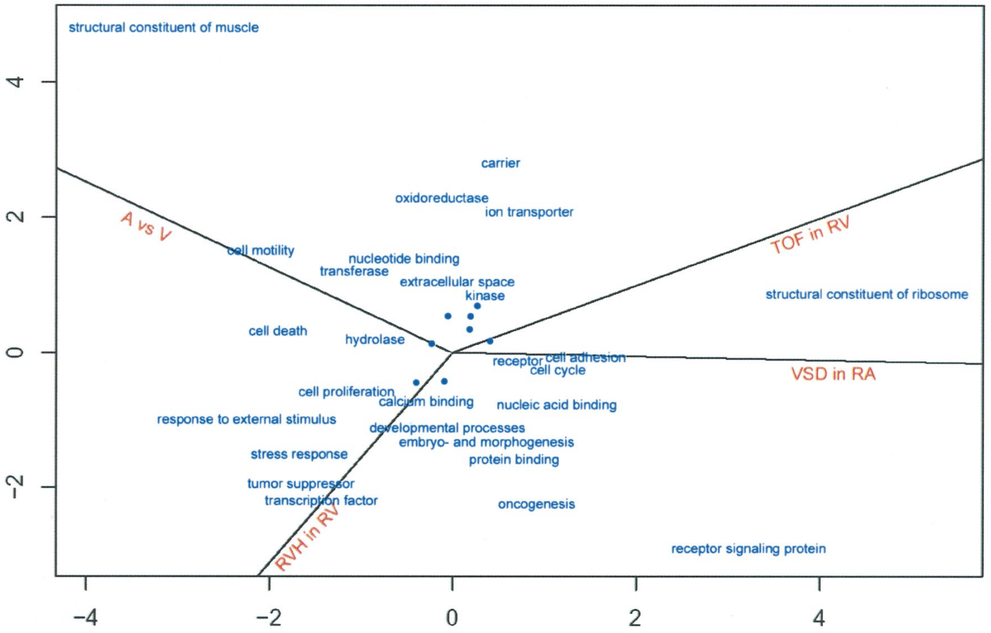


**Figure 3.** Biplot obtained from correspondence analysis. The plot shows the association between gene ontology categories (blue) and phenotypes (red) with respect to differential gene expression (see Results). The gene ontology categories belong to the 4th level of granularity, allowing a detailed but not diverse functional description. Categories that are not specifically associated with any phenotype are represented by dots.

$P<0.01$ by real-time PCR and confirmed 18 of them to be differentially expressed (Figure 2). Comparison of the obtained fold-changes by array analysis versus real-time PCR confirmed our approach as conservative, revealing higher expression differences in the latter (Table 2).

To provide an overview of the molecular signatures, we hierarchically clustered all known genes represented by resequenced clones that appeared significantly differential ($P<0.01$) in at least one of the above comparisons [Figure 2, $-\log_{10}(P)$ color-coded]. This overview allows a comparison between the transcriptional fingerprint of each of these genes in the different phenotypes. It appears that ≈25% of the CHD-associated genes (green or red in one of the first 3 columns) are not chamber-specific in the normal heart (yellow in the last 2 columns). Additional inspection shows a partially similar expression dynamic of the TOF portrait with that of RVH but an opposite expression dynamic compared with VSD. In addition, a large amount of genes characteristic for the molecular signature of VSD are not differentially expressed in any other analysis. Comparison of the expression profile of genes characteristic for TOF and RVH provides the possibility of subtracting both from each other and identifies genes specific for either TOF or RVH.

## Association of Functional Gene Categories to Specific Phenotypes

To provide a global view of the association of functional gene categories with particular phenotypes, we used the statistical method of correspondence analysis[13] (Figure 3). Clones were assigned to gene ontology categories according to the annotations in the LocusLink database.[14] We applied correspondence analysis to the contingency table of numbers of differentially expressed genes ($P<0.05$) per gene ontology category and phenotype comparison. Gene categories with similar patterns of regulation across phenotypes are mapped close to each other. Furthermore, the biplot shows the associations between gene categories and phenotypes. A gene category lies far in the direction of a certain phenotype, if disproportionally many genes are differentially expressed in this phenotype. For instance, many genes contributing to the structural integrity of a muscle fiber (structural constituent of muscle) are differentially expressed between atria and ventricle.

## Class Discovery Using ISIS

We used the class discovery method ISIS to see which class distinctions among the tissue samples are most pronounced in terms of gene expression profiles. The most outstanding binary class distinction identified, irrespective of sample annotations, was exactly the distinction between atria and ventricle samples.

## Chamber-Specific Expression

Because the human heart consists of 2 general compartments, the atria and ventricle, we analyzed the molecular repertoire that each of these expresses (Data Supplement). In addition to well-known chamber-specific genes, like atrial and ventricular myosin light chains, this signature includes diverse previously unknown chamber-specific genes for muscle contraction, extracellular components, cell growth and differentiation, and energy metabolism.

Circulation

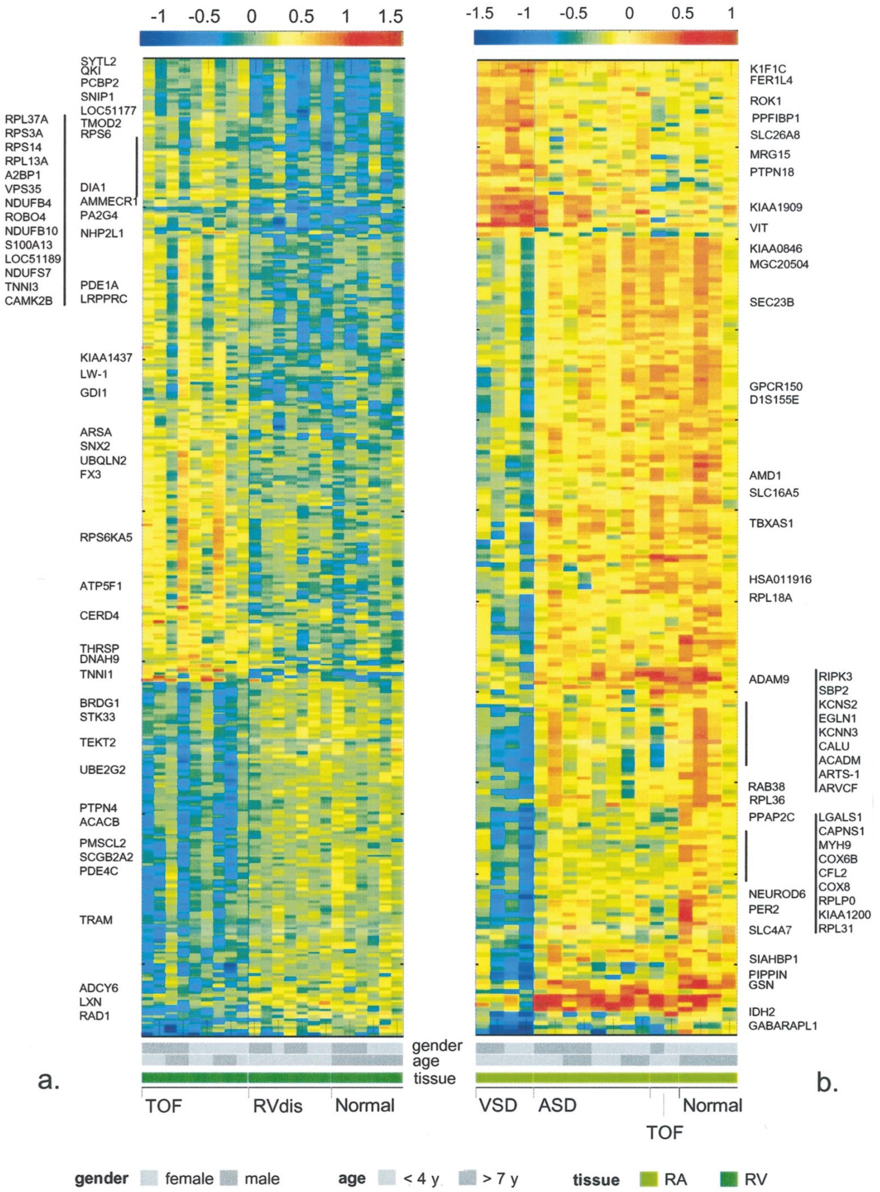


**Figure 4.** Hierarchical clustering of differentially expressed genes in TOF in RV (a) and VSD in RA (b). Shown are the expression patterns of genes associated with TOF and VSD compared with the other samples from the same cardiac chamber. Each column represents a single patient and each row a single clone. Normalized expression levels of clones with $P<0.01$ are color-coded for each sample. For examples of annotated genes, gene symbols are listed. The phenotype information for gender, age, tissue, and disease state is indicated.

The less force-developing atria were characterized by higher expression of genes encoding proteins associated with extracellular matrix or actin modulation, like CST3 and PCOLCE. The translation factor EEF1A and the DNA helicase REQL4 were highly significantly upregulated in atria. EEF1A-2/S1 protein is activated on myogenic differentiation and delays myotube death after apoptotic stress induction.[15] KCNIP2, a potential target for ventricular tachycardia,[16,17] is even more highly expressed in the human atria than in the ventricle ($P=0.01$), pointing to a potential role also in atrial arrhythmia. Genes with expression levels higher in ventricle than in atria belong mainly to 3 major functional classes: cytoskeleton-contraction, metabolism–energy turnover, and cell cycle–growth. Several of these genes are involved in ventricular myocardial disorders: TMP1[18] is mutated in human cardiomyopathies, FHL1 is downregulated in failing human hearts,[19] and ANKRD2 is involved in the process of cardiac hypertrophy.[20]

The combination of computational analysis of protein similarity together with our genome-wide transcription footprint of the heart points to so far functionally unknown genes. For example, we found Unigene cluster Hs.355815, which has a 59% protein sequence similarity to the myotonic dystrophy-associated protein kinase $\beta$ in rat, specifically expressed in ventricle.

In addition to atrial and ventricular specifications, we analyzed molecular differences between the normal high-pressure LV and low-pressure RV (Table 1, Figure 2). In general, we discovered genes encoding proteins involved in cell cycle, cell differentiation, and energy metabolism to be downregulated in the RV compared with the LV. An example of new information derived from our study is the LV-specific expression of Unigene cluster Hs.323099 with a sequence similarity of 93% to ALK3, which is essential for cardiac development.[21] Hs.323099 maps to chromosome 10q22, the genomic region that was previously linked to autosomal cardiomyopathy.[22] Furthermore, we did not observe any significant difference between LV and IVS.

### Tetralogy of Fallot and RVH

We observed distinct molecular portraits of TOF and RVH with genes of various functional classes. Even though the right ventricular hypertrophy is part of TOF, we could clearly distinguish between these 2 gene expression profiles by regarding in the statistical analysis once TOF and once TOF combined with the RVdis samples as one phenotype (RVH). Therefore, TOF reveals the molecular signature of the malformation in addition to the adaptation portrait.

Beside genes involved in cell cycle, a characteristic feature of the TOF signature is the upregulation of ribosomal proteins S6, L37a, S3A, S14, and L13A (Figure 4A). A specific role of ribosomal proteins during cardiac development has been only described for the chick ribosomal protein L10, which is downregulated in the cardiac outflow tract of chick embryos lacking neural crest cells.[23]

Our expression data reveal a TOF-specific dysregulation of potential targets that could be involved in pathways leading to cardiac dysdevelopment (Figure 4A); for example, SNIP, A2BP1, and KIAA1437 are upregulated. SNIP interacts with Smad4, a mediator of TGF-$\beta$, activin, and BMP signaling, which are essential for normal cardiac development.[25] A2BP1 belongs to a novel gene family sharing RNA-binding motifs expressed at the developing heart during mouse embryogenesis.[26] KIAA1437 binds k-ras, where k-ras–deficient mice develop a thin ventricular wall and die until term.[27] Genes markedly downregulated in TOF include STK33, BRDG1, and TEKT2.

In RVH we observed a hypertrophy-specific gene expression pattern of genes mainly involved in stress response, cell proliferation, and metabolism. Intriguing is the upregulation of ADD2, whose relative ADD1 was recently shown to be associated with hypertension in human.[28] Because the expression of several genes of the RVH signature was similar to their expression levels in LV (Figure 2), we analyzed whether the molecular adaptation to pressure overload could lead to a molecular transition from right to left ventricular characteristics. For the RVH-associated genes ($P<0.01$), we compared the difference between the mean intensity of each gene in RVH and that in normal RV samples to the mean intensity of the gene in normal LV samples. We found a significantly positive correlation coefficient of 0.27 ($P<0.0001$, permutation test), indicating that the genes dysregulated in RVH have a tendency to behave similarly in the disease state as in normal LV tissue.

To separate in TOF the malformation from the adaptation-specific molecular signature, we subtracted the RVH-specific genes from the TOF molecular portrait. All genes with $P<0.1$ in RVH were subtracted from all genes with $P<0.01$ in TOF, resulting in 88 clones highly specific for TOF with regard to the primary changes underlying the malformation process (Data Supplement).

### Ventricular Septal Defect

To obtain a molecular portrait that is not influenced by biomechanical adaptation processes, we studied RA samples of patients with VSD, intact tricuspid valve, and normal RA pressure. We observed a VSD-specific molecular signature dominated by downregulated genes with respect to the other RA samples (Figure 4B). As seen in TOF, several ribosomal proteins (S11, L18A, L36, LP0, L31, and MRPS7) are differentially expressed, but here they are downregulated. Other VSD-specific genes encode ion transporters or function during vertebrate development. The differential expression of ion channels was restricted to solute and potassium channels (SLC26A8, SLC16A5, SLC4A7, KCNS2, and KCNN3). A thorough literature study of downregulated genes in VSD revealed that a major part is involved in cell proliferation and differentiation during embryogenesis as well as apoptosis. Examples are AMD1,[29] RIPK3,[30] EGLN1,[31] and SIAHBP1.[32] It is noteworthy to mention the significant downregulation of ARVCF deleted in velo-cardio-facial syndrome.[33]

### Genome-Wide Compendium of Persistent Human Heart Transcripts

In addition, we selected a set of 6075 clones (4340 different Unigene clusters) that appeared to be persistently transcribed in our samples (Data Supplement).

## Discussion

Most microarray studies in human published to date, including the present one, have been observational studies, and it should be born in mind that effects related to bias and confounding have to be considered in advance.[34] Facing these statistical problems, we started this study by the collection of a large amount of samples, enabling a final selection of a patient population that is as balanced as possible with respect to known confounding factors. After visual inspection of a commonly used *t* test analysis, we observed age and tissue to be major confounding factors and therefore chose a linear model incorporating these factors as our statistical framework. Nevertheless, the findings of our study should be regarded as preliminary, and additional studies will be necessary to elucidate the role of the identified genes in normal and diseased human hearts.

We are able to show disease-specific molecular portraits for TOF and VSD as well as genes involved in the biome-

American Heart Association. Fighting Heart Disease and Stroke
Circulation
JOURNAL OF THE AMERICAN HEART ASSOCIATION

chanical adaptation process due to pressure overload. The combination of statistical analysis and our study design enabled the separation of secondary adaptation-specific expression patterns from effects attributable to primary cardiac dysdevelopment. In addition, we provide a chamber-specific genome-wide expression portrait of the normal human heart.

For a global overview of the disease and tissue types, we applied a variety of computational tools, such as the class discovery method ISIS and the statistical method of correspondence analysis, which could associate functional gene categories with particular phenotypes. We were able to confirm known genes and pathways involved in the developmental process and the molecular difference between the atria and ventricle. Furthermore, we obtained a global molecular portrait of the normal and diseased heart using a genome-wide approach including functionally unknown genes.

Unexpectedly, ribosomal proteins were found as a characteristic part of the TOF and VSD portrait. Taking into account that most human ribosomal genes were expressed at similar levels throughout our sample population, it is likely that these specific differentially expressed genes play an extraribosomal role during cardiac development.[33]

The global analysis of the VSD molecular signature showed a general transcriptional downregulation compared with TOF as well as with all other sample categories. The question of whether the observed downregulated transcription mirrors a reduction of essential proteins leading to incomplete fusion, an underdeveloped atrium, or an unknown physiological process will have to be elucidated in the future.

Our findings suggest that the analysis of malformed human hearts using powerful techniques like microarrays combined with statistical methods opens a new window to understanding cardiac adaptation and development.

## Acknowledgments

This project was supported by a grant from The Max Planck Society for the Advancement of Science. We acknowledge the support of the German Genome Resource Center Berlin. We thank M. Steinfath for image analysis; I. Dunkel for technical assistance, and J. Kreuder, D. Rawer, W. Huber, and L. Pauliks for helpful discussions.

## References

1. Benson DW, Silberbach GM, Kavanaugh-McHugh A, et al. Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways. *J Clin Invest.* 1999;104:1567–1573.
2. Basson CT, Bachinsky DR, Lin RC, et al. Mutations in human TBX5 cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet.* 1997;15:30–35.
3. Nediani C, Formigli L, Perna AM, et al. Early changes induced in the left ventricle by pressure overload: an experimental study on swine heart. *J Mol Cell Cardiol.* 2000;32:131–142.
4. Bauer EP, Kuki S, Zimmermann R, et al. Upregulated and downregulated transcription of myocardial genes after pulmonary artery banding in pigs. *Ann Thorac Surg.*1998;66:527–531.
5. Baumgarten G, Knuefermann P, Kalra D, et al. Load-dependent and -independent regulation of proinflammatory cytokine and cytokine receptor gene expression in the adult mammalian heart. *Circulation.* 2002;105:2192–2197.

5a. Kaynak B, von Heydebreck A, Mebus S, et al. Genome-wide array analysis of normal and malformed human hearts. Available at: http://www.molgen.mpg.de/~chd/arraywebsup. Accessed April 29, 2003.

6. Resource Center and Primary Database (RZPD), GmbH. Available at: http://www.rzpd.de. Accessed April 4, 2003.
7. Steinfath M, Wruck W, Seidel H, et al. Automated image analysis for array hybridisation experiments. *Bioinformatics.* 2001;17:634–641.
8. Beißbarth T, Fellenberg K, Brors B, et al. Processing and quality control of DNA array hybridization data. *Bioinformatics.* 2000;16:1014–1022.
9. Jin W, Riley RM, Wolfinger RD, et al. The contributions of sex, genotype and age to transcriptional variance in *Drosophila* melanogaster. *Nat Genet.* 2001;29:389–95.
10. Ihaka R, Gentleman R. A language for data analysis and graphics. *J Comput Graph Stat.* 1996;5:299–314.
11. Storey JD, Tibshirani R. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: Parmigiani G, Garrett ES, Irizarry RA, et al, eds. *The Analysis of Gene Expression Data: Methods and Software.* New York: Springer. In press.
12. von Heydebreck A, Huber W, Poustka A, et al. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics.* 2001;17(suppl 1):107–114.
13. Fellenberg K, Hauser NC, Brors B, et al. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A.* 2001;98:10781–10786.
14. LocusLink database. Available at: http://www.godatabase.org. Accessed April 4, 2003.
15. Chambers DM, Peters J, Abbott CM. The lethal mutation of the mouse wasted (wst) is a deletion that abolishes expression of a tissue-specific isoform of translation elongation factor 1alpha, encoded by the Eef1a2 gene. *Proc Natl Acad Sci U S A.* 1998;95:4463–4468.
16. Kuo HC, Cheng CF, Clark RB, et al. A defect in the Kv channel-interacting protein 2 (KChIP2) leads to a complete loss of I(to) and confers susceptibility to ventricular tachycardia. *Cell.* 2001;107:801–813.
17. Guo W, Li H, Aimond F, et al. Role of heteromultimers in the generation of myocardial transient outward $K^+$ currents. *Circ Res.* 2002;90:586–593.
18. Karibe A, Tobacman LS, Strand J, et al. Hypertrophic cardiomyopathy caused by a novel $\alpha$-tropomyosin mutation (V95A) is associated with mild cardiac phenotype, abnormal calcium binding to troponin, abnormal myosin cycling, and poor prognosis. *Circulation,* 2001;103:65–71.
19. Yang J, Moravec CS, Sussman MA, et al. Decreased SLIM1 expression and increased gelsolin expression in failing human hearts measured by high-density oligonucleotide arrays. *Circulation.* 2000;102:3046–3052.
20. Pallavicini A, Kojic S, Bean C, et al. Characterization of human skeletal muscle Ankrd2. *Biochem Biophys Res Commun.* 2001;285:378–386.
21. Gaussin V, Van de Putte T, Mishina Y, et al. Endocardial cushion and myocardial defects after cardiac myocyte-specific conditional deletion of the bone morphogenetic protein receptor ALK3. *Proc Natl Acad Sci U S A.* 2002;99:2878–2883.
22. Bowles KR, Gajarski R, Porter P, et al. Gene mapping of familial autosomal dominant dilated cardiomyopathy to chromosome 10q21–23. *J Clin Invest.* 1996;98:1355–1360.
23. Kirby ML, Cheng G, Stadt H, et al. Differential expression of the L10 ribosomal protein during heart development. *Biochem Biophys Res Commun.* 1995;212:461–465.
24. Deleted in proof.
25. Kiehl TR, Shibata H, Vo T, et al. Identification and expression of a mouse ortholog of A2BP1. *Mamm Genome.* 2001;12:595–601.
26. Garcia JM, Gonzalez R, Silva JM, et al. Mutational status of K-ras and TP53 genes in primary sarcomas of the heart. *Br J Cancer.* 2000;82:1183–1185.
27. Morrison AC, Bray MS, Folsom AR, et al. ADD1 460W allele associated with cardiovascular disease in hypertensive individuals. *Hypertension.* 2002;39:1053–1057.
28. Nishimura K, Nakatsu F, Kashiwagi K. et al. Essential role of S-adenosylmethionine decarboxylase in mouse embryonic development. *Genes Cells.* 2002;7:41–47.
29. Sun X, Lee J, Navas T, et al. RIP3, a novel apoptosis-inducing kinase. *J Biol Chem.* 1999;274:16871–16875.
30. Taylor MS. Characterization and comparative analysis of the EGLN gene family. *Gene.* 2001;275:125–132.
31. Liu J, Akoulitchev S, Weber A, et al. Defective interplay of activators and repressors with TFIH in xeroderma pigmentosum. *Cell.* 2001;104:353–363.
32. Sirotkin H, O'Donnell H, DasGupta R, et al. Identification of a new human catenin gene family member (ARVCF) from the region deleted in velo-cardio-facial syndrome. *Genomics.* 1997;41:75–83.
33. Potter JD. At the interfaces of epidemiology, genetics and genomics. *Nat Rev Genet.* 2001;2:142–147.
34. Wool IG. Extraribosomal functions of ribosomal proteins. In: Green R, Schroeder R, eds. *Ribosomal RNA and Group I Introns.* Austin: RG Landes; 1996:153–178.

Circulation
JOURNAL OF THE AMERICAN

### 2.1.3 Analysis of cardiac transcription factors and epigenetic marks at a global scale

Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, **Sperling S**. Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics* 2008;**91**:41-51.

Fischer JJ, Krueger T, Schueler M, Schlesinger J, Lange M, Toenjes M, **Sperling S**. The cardiac transcription network driven by Gata4, Mef2a, Nkx2.5 and Srf and epigenetic marks. *in preparation* 2008.

To understand molecular and developmental pathways in eukaryotic cells, TFs must be viewed within their regulatory context including other TFs and cofactors. Moreover, the ability of TFs to bind to DNA is highly influenced by the accessibility of their binding sites. In eukaryotic cells, DNA is packaged into chromatin by association with histone proteins. A high compaction of chromatin renders the DNA inaccessible to TF binding, silencing the genes in these regions. Consequently, the networks directing gene expression not only include the interplay between different TFs and co-regulatory elements but also epigenetic factors such as histone modifications. Thus, a major aim is the inference of epigenetic and genetic effects on transcription and finally the construction of regulatory networks in cardiomyocytes. Using chromatin immunoprecipitation with array detection (ChIP-chip) two major studies were performed, firstly, the analysis of histone modifications and secondly, the analysis of DNA binding sites of particular key transcription factors (Nkx2.5, Gata4, Srf, Mef2). Furthermore, knockdown of the respective transcription factors by siRNA enabled the dissection of direct and indirect targets.

#### ***Histone modifications in transcription and muscle differentiation***

The role of histone modification marks is currently a matter of a vigorous debate. According to the histone code hypothesis “distinct histone modifications, on one or more tails, act sequentially or in combination to form a ‘histone code’ that is read by other proteins to bring about distinct downstream events” (Strahl and Allis, 2000). This hypothesis has been much debated; in particular, if modifications encode distinct read-outs (Jenuwein and Allis, 2001; Turner, 2002; Cosgrove and Wolberger, 2005; Margueron et al., 2005) and about whether histone modifications serve as instructive signalling marks or rather appear as a consequence of transcription (Kurdistani and Grunstein, 2003).

To investigate the relationship between transcript levels and four histone modifications (H3K4me2, H3K4me3, H3ac, H4ac), two murine muscle cell lines were used, the cardiomyocyte cell line HL-1 cell and the skeletal muscle cell line C2C12. The later enabled the monitoring histone modification conversion, as cells were analyzed in the undifferentiated stage as myoblasts as well as in the differentiated stage as myotubes. The use of cell lines had the advantage of higher homogeneity and clearly defined cell states. Histone modifications were analyzed by chromatin immunoprecipitation followed by microarray analysis on custom-made oligonucleotide arrays (NimbleGen) that represented upstream (5kb) and transcribed regions of a comprehensive set of muscle expressed genes (8.585 genes corresponding to 10.976 transcripts). Expression levels for the same set of transcripts were determined using expression arrays. In total, approximately 3.000 sites for each of the histone modification types and in each cell type were identified. **Figure 4** illustrates this for the example of Hand2, for which in HL-1 cells two locations with H4ac-H3acK4me2/3 were found.

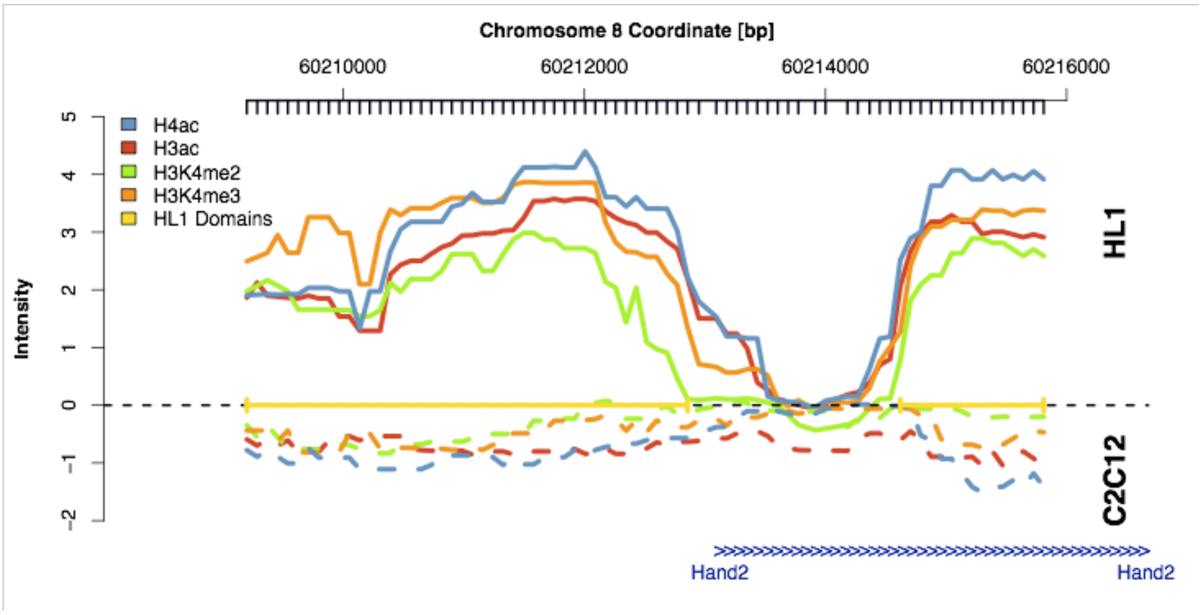


**Figure 4. Normalized and smoothed ChIP-chip intensities around the TSS of the Hand2 gene.** In C2C12 undifferentiated cells (dashed lines) no modifications are associated with the Hand2 gene. In HL-1 (solid lines) two domains with the modification code H4ac-H3acK4me2/3 were identified. Each domain consists of four modified sites.

In summary, the study showed that the average transcript levels associated with combinations of modifications are not simply (additively) related to those associated with individual modifications. The dynamics of histone modifications during muscle cell differentiation showed that the appearance of modifications are, by themselves, rarely associated with higher transcript levels. This is consistent with the view that these marks are a prerequisite, but not a sufficient driving force or a necessary result of transcription. Histone modifications may primarily function as signalling marks for specific effectors, and thus increase the combinatorial possibilities in the regulation of transcription. To gain a full understanding of the role of histone marks in transcriptional regulation, histone modifications need to be viewed in combination with their effectors, such as transcription factors.

### *The cardiac transcription network driven by Gata4, Mef2a, Nkx2.5 and Srf, and epigenetic marks*

The transcription factors Gata4, Mef2a, Nkx2.5, and Srf are known to be essential for the formation of the cellular structures required for a functional beating heart, by regulating the expression of structural genes such as Actin or Titin (Balza and Misra, 2006). The essential function of these TFs is most clearly demonstrated by severe phenotypes observed in mouse models. Mice lacking Gata4 die between 8.0 and 9.0 days postcoitum (dpc), because of failure of ventral morphogenesis and heart tube formation (Bhattacharya et al., 2006). Nkx2.5 is essential for normal heart morphogenesis, myogenesis, and function (Lyons et al., 1995). Targeted interruption of Nkx2.5 leads to abnormal heart morphogenesis, growth retardation and embryonic lethality at approximately 9-10 dpc. The majority of Mef2a(-/-) mice die within the first week of life and exhibit pronounced dilation of the right ventricle, myofibrillar fragmentation, mitochondrial disorganization and activation of a fetal cardiac gene program (Naya et al., 2002). Homozygous Srf-null mutations in mice result in lethality at gastrulation and severe defects in the contractile apparatus of the cardiomyocytes (Bhattacharya et al., 2006). Embryonic stem cells lacking Srf display defective formation of cytoskeletal structures, including actin stress fibers and focal adhesion plaques (Schratt et al., 2002).

The TFs Gata4, Mef2a, Nkx2.5, and Srf are evolutionarily highly conserved and known to play a role in the formation of congenital heart diseases in human patients. For example, more than ten disease-related mutations in NKX2.5 have been documented in patients with a spectrum of congenital heart diseases (Akazawa and Komuro, 2005). The most common phenotypes are secundum atrial septal defect and atrial-ventricular conduction disturbance, but other cardiac abnormalities have been reported as well.

Furthermore, the formation of heterodimers between some of these TFs has been reported and they might form a sub-network, in which they regulate each other's expression. The following activating binding events had been reported: Gata4 → Nkx2.5 (Searcy et al., 1998), Srf → Gata4 (Balza and Misra, 2006), Srf → Nkx2.5 (Spencer and Misra, 1996), Srf → Srf (Balza and Misra, 2006). In addition, the expression of Gata4 and Nkx2.5 was reduced in a cell culture model where the function of all Mef2 proteins had been abolished (Karamboulas et al., 2006).

Using ChIP-chip analysis 469 Gata4, 970 Mef2a, 392 Nkx2.5, and 1.510 Srf target genes were identified, including 42 known targets. Several genes previously known to be dysregulated in mutants/knock-outs of the respective transcription factor are direct targets. For example, the decrease of *Gata4* and *Nkx2.5* levels in cells depleted of all four Mef2s (Karamboulas et al., 2006) can now be explained by the observed binding of Mef2a at the corresponding promoters.

To gain insights into the transcription factor functionality, overrepresented gene ontology terms were analyzed among target genes and confirmed at a global scale the importance of the respective TFs for cardiac development and function. For example, among the Nkx2.5 targets identified in this study, the GO term 'heart looping' is significantly overrepresented; and in *Nkx2.5* hypomorphs looping of the linear heart tube is not initiated (Lyons et al., 1995).

As the expression of genes is typically coordinated by multiple transcription factors, the frequency of corregulation by different transcription factors was investigated. Genes are frequently bound by more than one TF and all possible combinations occur, suggesting combinatorial gene regulation. Gata4 and Nkx2.5 had the lowest number of targets (Gata4 469, Nkx2.5 392) but showed co-binding to 203 genes. Thus, their occurrence is highly correlated. Although Mef2a and Srf bind at 438 genes together, they each have a much higher number of target genes (Mef2a 970, Srf 1.510). Pairwise physical interaction has been described between several of the investigated TFs (Akazawa and Komuro, 2005; Clark et al., 2006).

The activating potential of a factor in governing gene expression is strongly influenced by the accessibility of its binding sites within the chromatin structure. As the four histone modifications H3ac, H4ac, H3K4me2, and H3K4me3 are considered to induce an open chromatin configuration, their cooccurrence with the analyzed TFs was investigated. Approximately 55-74% of the respective transcription factor binding sites are additionally marked by one or more histone modifications; in a randomized simulation only between 23% and 38% are expected to appear together. It is well known that the investigated TFs interact with a variety of histone modifying enzymes. The histone acetyl transferases (HAT) p300 not only acetylates lysine residues on histone 3 but also on Gata4, thereby enhancing the DNA-binding and activating potential of this TF. The Srf-cofactor Myocardin (Myocd) has been reported to recruit p300 to Srf binding sites whereby histone 3 acetylation is induced and gene expression enhanced (Cao et al., 2005). Consequently the presence of H3ac has an influence on the expression levels of direct target genes was investigated (**Figure 5**). Genes showing neither TF binding nor H3ac were used as reference. For Nkx2.5 and Mef2a the expression levels of bound genes were significantly higher than the reference group, independent of whether H3ac was present or not. In case of the p300-interacting proteins Gata4 and Srf the expression levels of bound genes were only significantly increased if the binding sites were additionally marked by H3ac (80% of Gata4 and 72% of Srf TFBS). This indicates that acetylation of histone 3, probably via p300, supports the activating function of Gata4 and Srf.

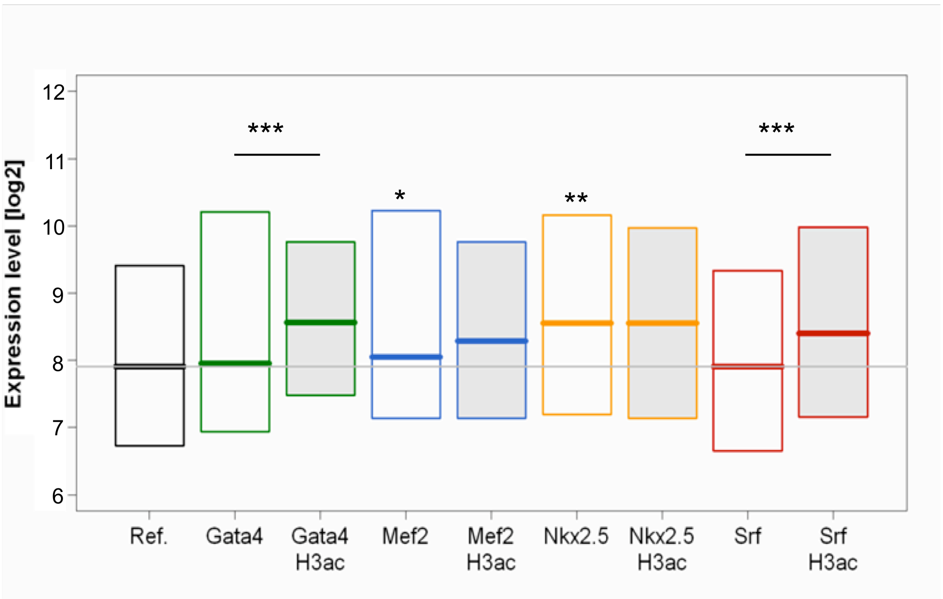


**Figure 5. The influence of histone 3 acetylation on expression of TF target genes.** For each TF the binding sites were categorized into two groups depending on whether the TF binds alone or co-occurred with H3ac. The expression levels are represented as box plots. The resultant *p*-values are indicated: $p \leq 0.005$ (***), $p \leq 0.01$ (**) and $p \leq 0.05$ (*). As reference the expression levels of genes showing neither binding of investigated TFs nor H3ac is given.

The approach used in this study gave global insights into the architecture of transcriptional regulatory networks in general and into the functions of the investigated transcription factors in particular. Together with the previous observation on histone modifications, a novel regulatory circuit linking the histone 3 acetylation with activation of gene expression through Gata4 and Srf could be delineated. Moreover, these data are a valuable basis for the effort to construct cardiac gene regulatory networks based on correlated gene expression in human malformed hearts and prediction of transcription factor binding sites.

Author's personal copy

Available online at www.sciencedirect.com

ScienceDirect

GENOMICS

Genomics 91 (2008) 41 – 51

www.elsevier.com/locate/ygeno

# Combinatorial effects of four histone modifications in transcription and differentiation ☆

Jenny J. Fischer [a], Joern Toedling [b], Tammo Krueger [a], Markus Schueler [a], Wolfgang Huber [b], Silke Sperling [a],*

[a] *Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany*
[b] *European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK*

Received 23 July 2007; accepted 31 August 2007
Available online 8 November 2007

## Abstract

Nucleosomes are involved in DNA compaction and transcriptional regulation. Yet it is unclear whether histone modification marks are primary or secondary to transcription and whether they interact to form a *histone code.* We investigated the relationship between transcription and four histone modifications (H4ac, H3ac, H3K4me2/3) using ChIP–chip and expression microarray readouts from two murine cell lines, one in two differentiation stages. We found that their association with transcript levels strongly depends on the combination of histone modifications. H3K4me2 coincides with elevated expression levels only in combination with acetylation, while H3ac positive association is diminished by co-occurring modifications. During differentiation, upregulated transcripts frequently gain H4ac, while most modification conversions are uncorrelated with expression changes. Our results suggest histone modifications form a code, as their combinatorial composition is associated with distinct readouts. Histones may primarily function as signaling marks for specific effectors rather than being a sufficient driving force for or a consequence of transcription.
© 2007 Elsevier Inc. All rights reserved.

*Keywords:* Cell differentiation; Chromatin immunoprecipitation; Histones; Microarray analysis; Muscles; Myocytes, cardiac; Nucleosomes; Transcription, genetic

Histones together with DNA form the nucleosome and are characterized by a central region and an exposed N-terminal tail. These tails provide a surface for interactions with other proteins [1] and are susceptible to a wide variety of posttranslational modifications [2]. These include acetylation [3] and methylation [4,5] at lysine residues. The type of histone modification contributes to the degree of DNA accessibility and gene transcription. Acetylation leads to a reduction of positive charges on the histone tails and thereby diminishes the interaction with the negatively charged DNA backbone, leading to an open chromatin state [6]. Methylations have been ascribed activating and repressive functions, depending on the position of the residue [5,7], and different effects may be associated with mono-, di-, or trimethylation of lysine residues.

According to the histone code hypothesis “distinct histone modifications, on one or more tails, act sequentially or in combination to form a ‘histone code’ that is read by other proteins to bring about distinct downstream events” [8]. This hypothesis has been much debated, in particular as to if modifications encode distinct readouts [9–12] and whether histone modifications serve as instructive signaling marks or rather appear as a consequence of transcription [13]. Although different interpretations of the concept of a code are possible, we suggest that the existence of such a code should manifest itself such that different modification combinations lead to distinct outcomes. Histone modifications can be recognized and subsequently translated into a functional consequence by specific effectors, such as bromo- and chromodomain proteins [1]. Central to the current debate is the question of the extent to which histone

---

*Abbreviations:* ChIP, chromatin immunoprecipitation; ChIP–chip, chromatin immunoprecipitation followed by microarray analysis; GO, Gene Ontology; H3ac, histone 3 acetylated at lysine residues 9 and 14; H3K4me2, histone 3 dimethylated at lysine 4; H3K4me3, histone 3 trimethylated at lysine 4; H4ac, histone 4 acetylated at lysine residues 5, 8, 12, and 16; TSS, transcription start site.

☆ Microarray data from this article have been deposited with the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) E-TABM-182 and E-TABM-183.

* Corresponding author. Fax: +49 30 8413 1699.
*E-mail address:* sperling@molgen.mpg.de (S. Sperling).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.08.010

Author's personal copy

modifications act in a combinatorial manner, wherein different combinations lead to different downstream effects. To address this issue, two questions need to be answered: which combinations occur in vivo, and what are their functional consequences? Recently, the joint occurrence of one activating and one repressive histone mark has been reported to result in weak or no transcription [14,15], which may indicate either the existence of a histone code or a predominant effect of repressive marks.

We therefore concentrated on modifications previously all described to be equally associated with higher transcript levels in yeast [16–18] and higher eukaryotes [19–21], enabling us to investigate combination effects without the possible confounding factor of a dominant repressive mark. As the investigation of combination effects is feasible only for modifications often occurring at the same chromosomal location we furthermore limited our investigations to modifications reported to colocalize frequently [16–21]. We studied acetylation of histones 3 and 4 (H3ac and H4ac) as well as di- and trimethylation of lysine 4 on histone 3 (H3K4me2, H3K4me3). Although the correlation between these modifications has been described to be high, possible functional consequences of colocalizations compared to single occurrence have so far not been investigated. To obtain a global picture, we investigated these modifications on a transcriptome-wide basis in two different murine cell lines, cardiomyocytes (HL-1) and skeletal muscle (C2C12), the latter as undifferentiated myoblasts as well as differentiated myotubes. This allowed us to characterize the combinatorial relationship between modifications and transcription in three different cell types and to follow changes during differentiation. If histone modifications were mainly a result of transcription, we would expect changes in expression level to be highly correlated with modification changes.

Considering each modification individually and, in a novel approach, their combinatorial occurrence, we showed that combinations of modifications were associated with transcription in a manner that was not simply related to their individual effects. This suggests that combinatorial information is encoded in the histone tails. During differentiation a large number of transcripts were associated with modification changes, although these were mostly not associated with changes in expression. This implies that histone modifications may primarily function as prerequisite signaling marks for, rather than being a consequence of, transcription.

## Results

### *Genes are characterized by a discrete number of modified domains and modification codes*

We investigated histone modifications in three different cell types (myotubes, myoblasts, and cardiomyocytes) and followed changes during differentiation from myoblasts to myotubes. The use of cell lines had the advantage of higher homogeneity and clearly defined cell states. Histone modifications were analyzed by chromatin immunoprecipitation followed by microarray analysis (ChIP–chip) on custom-made oligonucleotide arrays that represented upstream and transcribed regions of a comprehensive set of muscle-expressed genes (8,585 genes corresponding to 10,976 transcripts). Expression levels for the same set of transcripts were determined using expression arrays. We assigned modified sites to a transcription start site (TSS) of a gene if they were located within 5 kb upstream of the TSS or within the transcribed region. For genes with multiple TSSs, we considered each TSS and the respective transcribed region individually. In total, we identified, in each cell type, approximately 3,000 sites for each of the histone modification types (Fig. 1). The average number of sites per represented gene (~0.35) and median site sizes (~600 bp) are in good agreement with previous data [19] (Table S1).

To gain an understanding of the combinatorial co-occurrence of histone modifications, we analyzed how often genomic locations were enriched for one, two, three, or all four modification types. Four histone marks can occur in 15 different combinations at one position; however, we observed that *in vivo* only a few of these occurred frequently. While modifications on histone 3 predominantly appeared together, histone 4 acetylation occurred mainly either by itself or in conjunction with all three other modifications (Fig. 2A). The distributions of combinations were similar for the three cell types (Fig. 2B). To analyze further the combinatorial occurrence of modifications, we assigned a modification code to each location and termed this a *modified domain,* or more concisely, a *domain.* Fig. 2C illustrates this for the example of *Hand2*, for which in HL-1 cells two domains with the modification code H4ac–H3acK4me2/3 were found. In general, we found about half of the analyzed genes to be associated with modified domains (Fig. 1). Two-thirds of these genes were marked by only one modified domain and approximately one-fourth by two domains. In cases in which one domain per TSS was identified, these were predominantly multicode H3acK4me2/3, H4ac–H3acK4me2/3 or single-code H4ac. If two modified domains per TSS occurred, these generally had different modification codes, one of them characteristically H3acK4me2/3. Single-code H4ac was found to be typical for genes marked by two or more identical modified domains. The distribution and composition of domains showed little variation between the cell types.

### *Modified domains provide insights into the localization of modifications*

We asked whether particular modifications, or combinations thereof, showed a preference for particular genomic locations relative to the TSS. Considering modified sites (Fig. 3A), we observed enrichment of histone modifications within ±1 kb, although few modifications were found directly at the TSS. Histone 4 acetylation was predominantly localized upstream, and the three modifications on histone 3 more frequently downstream of the TSS.

This picture was refined by consideration of the combinatorial nature of histone modifications (Figs. 3B and C). We found that although multicode domains containing H3K4me2 peaked close to the TSS, domains marked by H3K4me2 alone were distributed throughout the transcribed region. Domains charac-

Author's personal copy

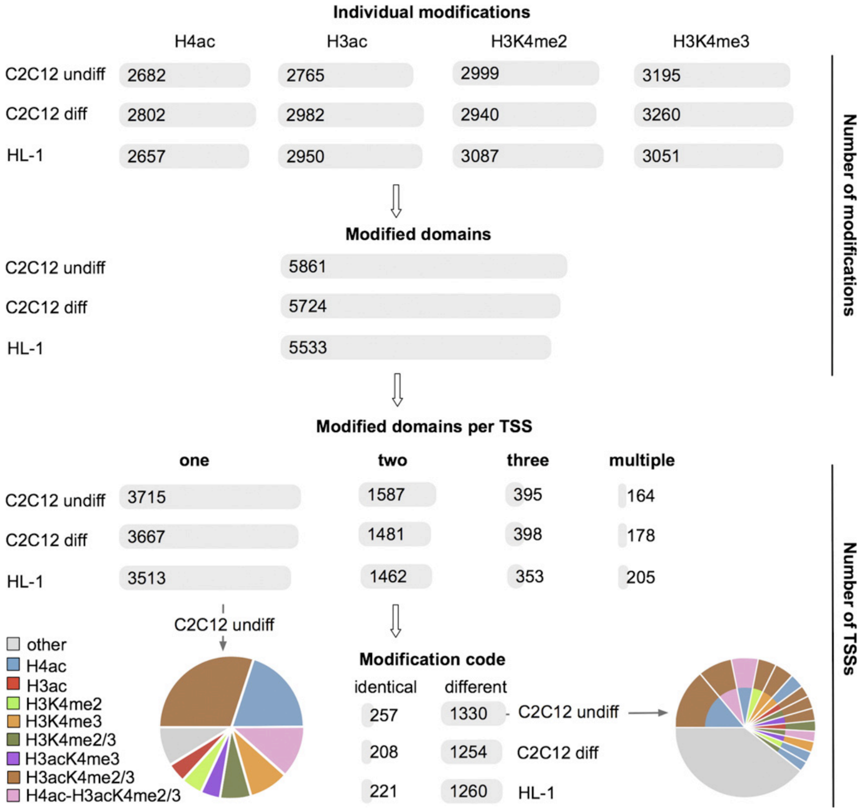


Fig. 1. The distribution of modifications shows little variation over the cell types. The horizontal bars show the number of individual modifications (top row) and of modified domains (second row) for the three cell types. The third row shows the distribution of the number of modified domains per TSS. TSSs with one or two domains are most frequent. The pie charts show the distributions of modification codes for one-domain transcripts and for two-domain transcripts with different codes in C2C12 undifferentiated cells.

terized by H4ac alone occurred almost exclusively upstream, while in conjunction with other modifications (H4ac–H3acK4-me2me3) it was found as often up- as downstream.

### *Modified domains show combinatorial effect of modifications*

We asked how the presence of histone modifications is related to transcript expression levels. Even though modified domains showed a preference for certain positions relative to the TSS, we did not observe a significant association between domain position and transcript levels (Fig. S1). Therefore, the following analysis considers presence and absence of domains irrespective of their position. The data from each of the three cell types were evaluated separately (data not shown), as the various modification patterns between the cell types differed strongly. Of the approximately 5,000 TSSs per cell type found to be associated with domains only 1,267 TSSs were marked by the same modifications in myoblasts, myotubes, and cardiomyocytes, while around 1,500 TSSs showed cell-type-specific modification patterns. Although we find this high degree of cell-type specificity in domains, the overall results linking modification patterns to transcript levels were found to be highly comparable and therefore a composite data set that includes domains and transcript levels from all three cell types was used in the following to illustrate the results.

We investigated the relationship between transcript levels and histone modifications from two angles. First, transcripts were classified according to their expression levels and the frequency of domains within these classes was compared. Second, we classified transcripts based on their associated domains and investigated the expression levels. Transcripts were divided into four different expression categories and the occurrence of modifications within these classes was analyzed, once based on modified sites and once considering modified domains. Regarding modified sites (i.e., without taking combinatorics into account), we observed that all four histone modifications, in particular the two types of acetylation, coincided with elevated transcript levels (Fig. 4A). The 10% most highly expressed transcripts, however, showed no further increase in modifications compared to medium expression levels. Repeating the analysis with modified domains revealed strikingly different results (Fig. 4B). The fraction of single-code H3ac domains increased by 200% in the class of expressed transcripts

Author's personal copy

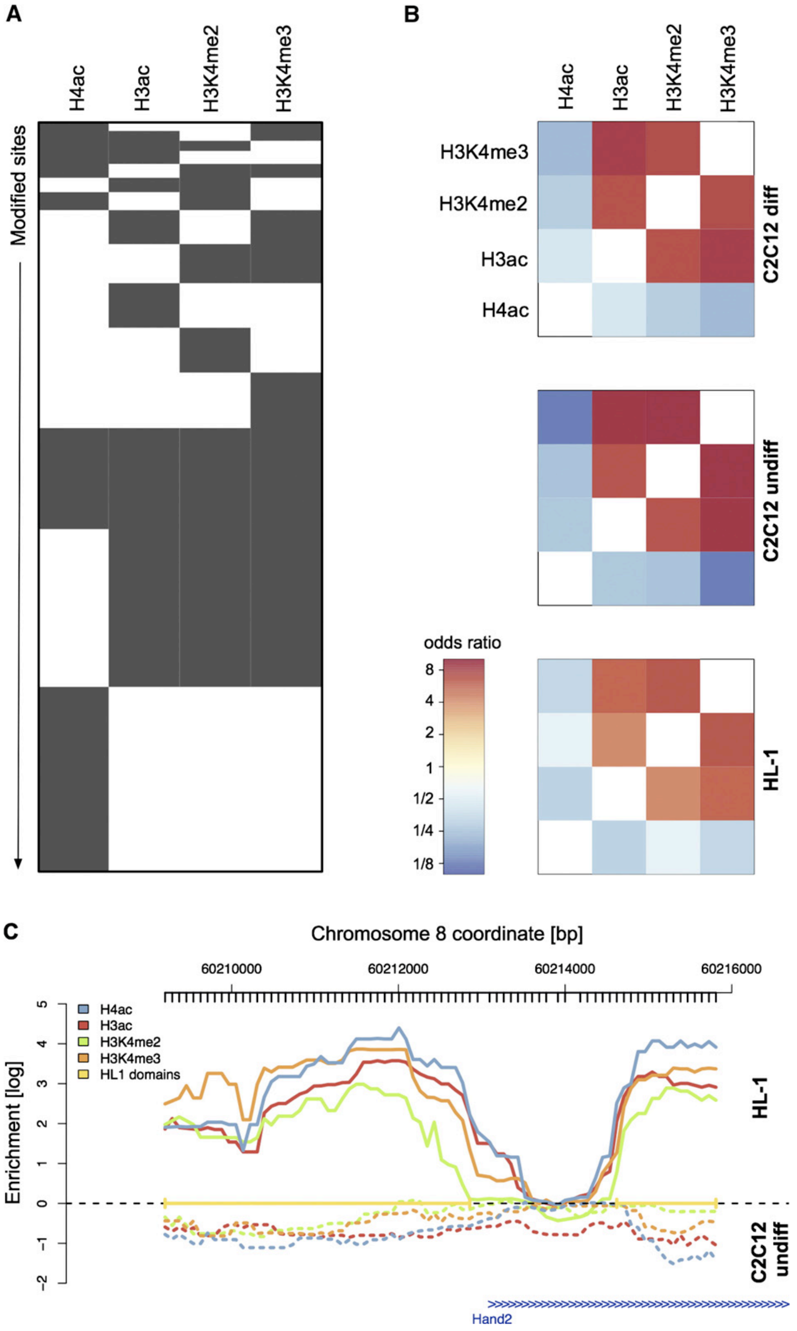


Fig. 2. Modifications are highly correlated. (A) The combinatorial occurrence of histone modified sites. Each row in the heat map corresponds to a combination of modified sites. Dark indicates presence, white absence. The height of the rows is proportional to the number of occurrences of a combination, summed over the three cell types. The most frequent cases are H4ac alone, H3acK4me2/3, and the combination of both. (B) Odds ratios of pair-wise contingency tables of the occurrence of modified sites in domains. The pattern is similar for all three cell types. Red indicates positively correlated occurrence, blue corresponds to anticorrelation. (C) Normalized and smoothed ChIP–chip intensities around the TSS of the *Hand2* gene. In C2C12 undifferentiated cells (dashed lines) no modifications are associated with the Hand2 gene. In HL-1 (solid lines) two domains with the modification code H4ac–H3acK4me2/3 were identified. Each domain consists of four modified sites.

Author's personal copy

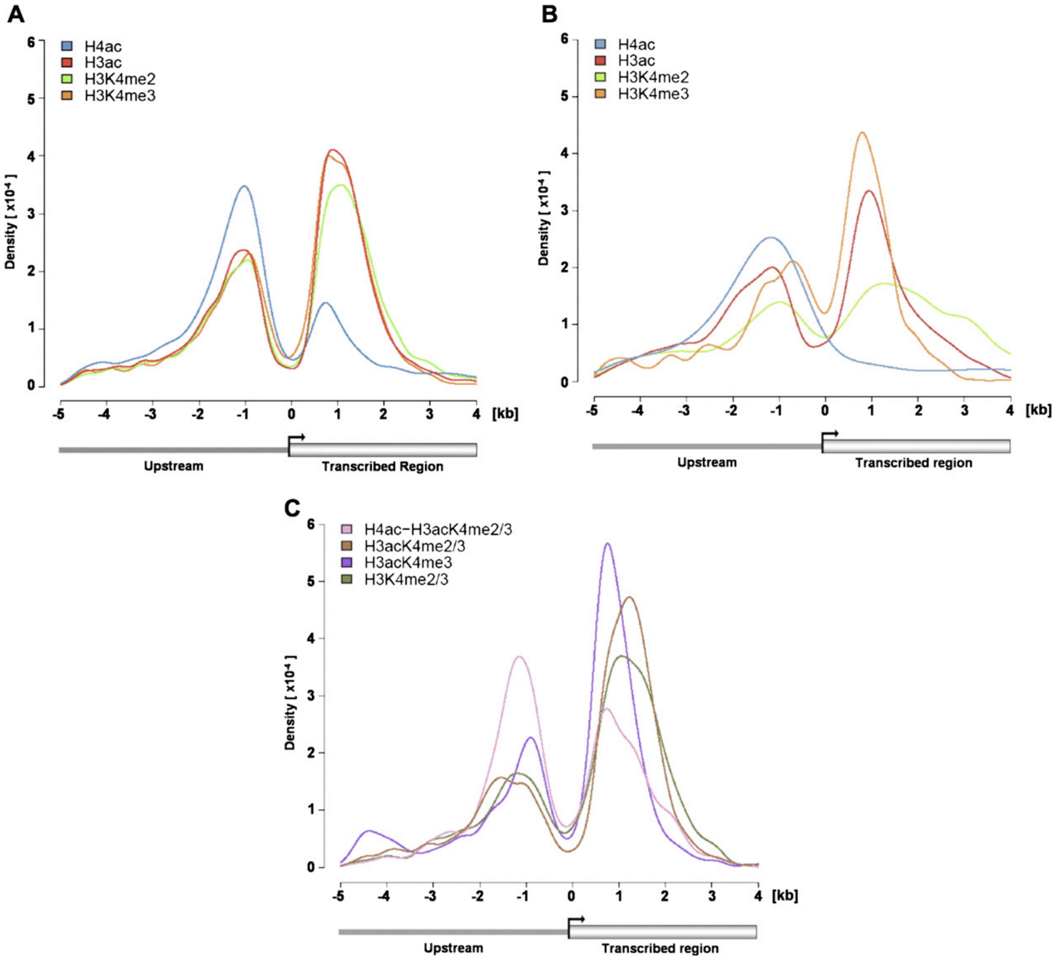


Fig. 3. Positional distribution of histone modifications relative to the TSS. Shown are data for C2C12 undifferentiated cells. The 5-kb upstream and 4-kb downstream regions of each TSS were aligned by the TSS (*x* axis). The *y* axis shows the densities. (A) Analysis based on modified sites (i.e., without taking into account combinatorics). Modifications on histone 3 show similar distributions. H4ac occurs predominantly up- but also downstream. (B and C) Analysis on the basis of modified domains (i.e., considering co-occurrence of modifications). (B) Localization of the four single-code domains is shown. Single-code H4ac is positioned almost exclusively upstream, single-code H3K4me2 is distributed throughout the transcribed region. (C) Localization of the four most frequent multicode domains is shown. Domains containing all four modifications occur similarly often up- and downstream. Domains coding for histone 3 modifications are predominantly positioned in the transcribed region.

compared to the nonexpressed class. Domains containing further modifications in addition to H3ac did not show such a pronounced increase. The proportion of domains solely containing one or both methylation modifications varied only slightly between three expression groups but was decreased for the most highly expressed transcripts.

Next, we classified all transcripts represented on the ChIP arrays according to the modifications in the vicinity of the TSSs. Genes represented on the array but without significant enrichment of probes for any of the investigated modifications were categorized as nonmodified. Without considering combinations, all four modifications clearly, and to approximately the same degree, coincided with higher expression values (Fig. 5A, Table S2). Repeating the analysis on the basis of modified domains, we obtained a ranking of the effect of combinations on expression levels (Fig. 5B and Table S3). Single-code H3K4me2 and multicode H3K4me2/3 modified transcripts' mean expression levels were not significantly higher expressed than nonmodified transcripts, whereas H3K4me3 alone was associated with slightly higher expressed transcripts (mean fold change 1.22, $p=6\times10^{-12}$). TSSs associated with single-code H3ac coincided with the highest expression levels (mean fold change 1.53, $p=1\times10^{-30}$), and in case of co-occurrence of further modifications the respective transcripts showed lower expression ($p\leq3\times10^{-2}$). None of the single-code domains were repressive. Co-occurrence of modifications generally was associated with lower expression levels than expected from the sum of the individual modifications effects (linear model analysis, Table S6). These findings demonstrate the combinatorial, nonadditive effect of histone modifications.

Author's personal copy

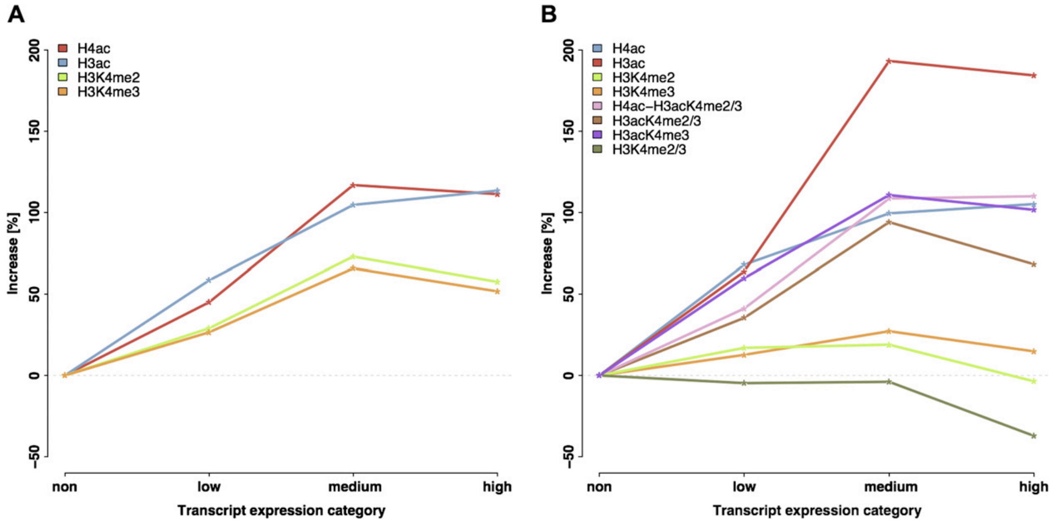


Fig. 4. Higher expression levels of transcripts are associated with stronger histone acetylation. Transcripts were classified into four groups according to expression level: non-, low-, medium- and high-expressed (*x* axis). The *y* axis shows the percentage increase in the frequency of modifications in each group, compared to the nonexpressed group. The curves were obtained by taking the median across the three cell types. (A) Analysis based on modified sites. Each individual modification increases over the expression groups. The frequencies of the two acetylation states and of the two methylation states, respectively, show similar increases. (B) Analysis on the basis of domains. Shown are the eight most frequent modified domains. The frequency of single-code H3ac shows the strongest increase. Other acetylation-containing domains behave similarly and increase by approximately 100%. Domains coding for one or both methylations show basically no change.

### *Change of modification state during differentiation is not generally associated with change of expression level*

To complete the static picture gained from the analysis of independent cell types, we examined histone modifications during skeletal muscle differentiation and investigated the dynamic of changes of modifications and their correlation to changes in expression. If histone modifications were a consequence of transcription or, alternatively, a sufficient driving force, we would expect modification

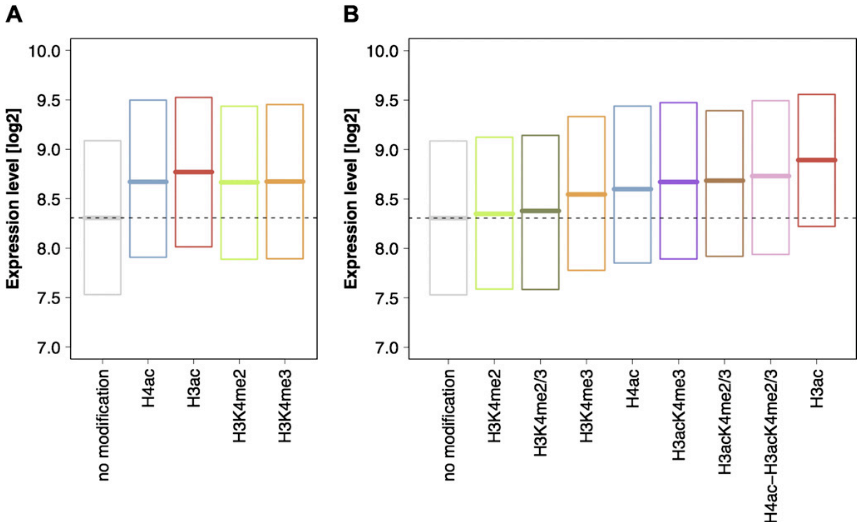


Fig. 5. The combinatorial nature of the relationship between histone modifications and expression. Transcripts were grouped by their associated histone modifications. The expression level distribution in each group is represented by box plots. (A) Analysis based on modified sites. All four individual modifications are similarly associated with elevated expression levels compared to the no-modification group ($p<1\times10^{-30}$, Table S2). (B) Analysis on the basis of domains. The highest levels are seen with H3ac alone; H3ac combined with the other three modifications shows comparatively lower levels ($p\le3\times10^{-2}$, Table S3). The levels with single-code H3K4me3 are comparatively low and even reduced in combination with H3K4me2. H3K4me2-associated transcripts are not significantly different from the no-modification group.

Author's personal copy

changes to be highly associated with changes of expression levels.

We found 299 transcripts to be significantly differentially expressed between myoblasts and myotubes. Gene ontology annotations overrepresented for these transcripts were in good agreement with published results [22] (Table S4). In comparison to the amount of differentially expressed transcripts, a much higher number of TSSs (3,498) were associated with modification changes between myoblasts and myotubes. Although the overall number of modified domains remained fairly constant, a high number of modifications were lost or gained during differentiation (Fig. 6). Strikingly, the majority of modification conversions involved H4ac both as a singly occurring modification and in the context of multicode domains: 727 TSSs showed H4ac in myoblasts and were associated with no modification at all in myotubes. Domains coding for the three modifications on histone 3 gained additional histone 4 acetylation in 299 cases, whereas 134 domains showing all four modifications lost H4ac while retaining H3acK4me2/3. In comparison, only 25 domains coding for the three modifications on histone 3 were lost and 29 gained. A change from no modification to single-code H4ac in differentiation was found for 854 TSSs and was associated with the presence of Mef2 binding sites in the surrounding sequence (group specificity score $4.5 \times 10^{-4}$). Mef2 is a histone acetyl transferase recruiting factor and is essential for skeletal muscle differentiation [23].

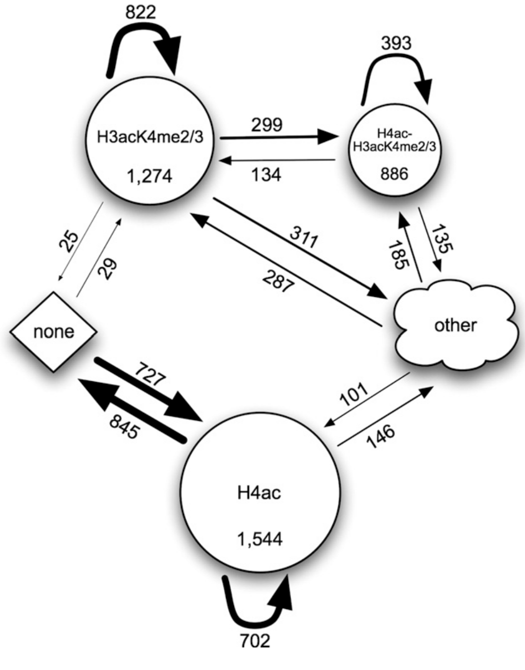


Fig. 6. The dynamics of domain codes in C2C12 differentiation. Shown is a schematic representation of the loss and gain of histone modifications in the process of differentiation. Each node is the representation of the number of modified domains found upstream or within the transcribed region of genes in myotubes. Only the three domain codes showing the highest number of changes are shown. Self-loops indicate number of domains retaining both position and code identity in differentiation. Straight arrows represent conversions of domains in the transition from undifferentiated (myoblasts) to differentiated cells (myotubes). The number next to each arrow denotes the absolute frequency of that conversion type. For example, 886 domains containing all four modifications are present in myotubes. 393 of these domains were present at the identical position in myoblasts, while 299 originated from H3acK4me2/3 by the addition of H4ac. On the other hand 134 of the domains present in myoblasts lost H4ac while retaining H3acK4me2/3.

Most of the changes in histone modification status between undifferentiated and differentiated C2C12 cells were not associated with differential expression. For example, among the 854 transcripts associated with a gain of H4ac, only 29 showed significantly higher expression levels (Table S6A). However, among the total of 126 upregulated transcripts, gain of H4ac was seen significantly more often than among not differentially expressed transcripts ($p = 1 \times 10^{-7}$, logistic regression model, Table S7). Other modification changes were not significant. For the downregulated transcripts, we did not observe significant preferences for modification changes (Table S8).

## Discussion

We present the occurrence and combinatorial effect of two histone acetylation and two histone methylation states in three cell types in a transcriptome-wide investigation. A genomic location where the probes showed significant enrichment for one of the investigated modifications was called a modified site. The number of sites for each histone mark were similar in myotubes, myoblasts, and cardiomyocytes. Approximately half of the genes contributing to each transcriptome were associated with modified histones and multiple modifications frequently occurred together. We introduced the term *modified domain* to represent the combinatorial modification code observed at a particular sequence. The majority of TSSs were associated with only one modified domain, since modifications were typically clustered together. However, we found that although correlations of the modifications on histone 3 were strong, acetylation of histone 4 frequently occurred as a single-code domain that is without enrichment of any of the other modifications. This picture may be refined in the future by technological advances that provide higher spatial resolution and allow the characterization of modifications on the level of single nucleosomes or histones [2,24,25].

We found enrichment for histone marks within ±1 kb of the TSS, as described previously. However, only low levels of modifications occurred directly at the TSSs. This indicates that these regions are either rarely modified or contain few nucleosomes. From studies in yeast, nucleosome-depleted regions are well known [26,27] and have been recently shown to exist in higher eukaryotes as well [28]. Different modified domains show positional preferences relative to the TSS, above and beyond those of the individual modifications. In particular, single H4ac appears nearly always upstream. For yeast it has been reported that H4K3me2 occurs throughout transcribed regions. In higher eukaryotes, however, only a general

Author's personal copy

association of this mark with transcribed regions has been observed so far [19,29]. We could now show that in mouse the sites exclusively marked by H4K3me2 are distributed throughout transcribed regions, as in yeast.

We investigated the relationship between the occurrences of histone modifications upstream or within the transcribed region of a gene and the expression levels of the corresponding transcript. Instead of, or in addition to, expression levels, occupancy by RNA polymerase II has been used to determine the transcriptional status of a gene [16,30]. However, such results are, in contrast to expression array intensities, difficult to quantify. Indeed, generally only present/absent calls were reported. Furthermore, published data suggest a high number of false negatives in such experiments, as a substantial number of transcripts were detected on mRNA level but no binding of RNA Pol II could be found [30].

Our results indicate that histone tail modifications form a combinatorial code, in the sense that different combinations lead to distinct outcomes. Looking at the modifications one at a time, we found them to be approximately equally associated with higher transcript levels, as has been previously reported [16,18–21,31]. However, when the combinatorial nature of the modifications is considered, we find that the different modifications influence each other in such a way that the effect of one modification changes when a different modification is also present. We obtained consistent results when conducting the analysis both from the perspective of transcript levels and from the perspective of modification associations. We found H3ac to have the strongest association with higher transcript levels, while additional activating modifications reduced this effect. Modified domains consisting of single H3K4me2 or its combination with H3K4me3 showed no positive correlation with transcript levels. H3K4me3 has been employed as a marker to identify actively transcribed genes [30]. However, several studies have reported that a substantial percentage of transcripts associated with this modification were not expressed (31% [30], 24% [15,20,32]). Our results offer an explanation for these findings. It seems that H3K4me3 is not an optimal marker to identify transcribed genes and that the activating effect ascribed to it is mainly a result of its frequent colocalization with acetylations. Based on our analysis we would suggest histone acetylations, in particular H3ac, to be better predictors of elevated expression levels. We made similar observations in two independent cell lines, one of which was considered in two differentiation stages. This suggests that our results have general applicability.

Finally, we investigated the role of these modifications in the process of differentiation. We found histone 4 acetylation to be highly dynamic and that upregulated transcripts were associated with a gain of this modification. However, a large number of regions showed modification alterations, and these were mostly not associated with changes in expression. A possible explanation is that modifications precede transcription, but are not by themselves sufficient for its regulation.

It is conceivable that the histone modifications that are gained and lost during differentiation could function as signaling marks. Single-gene studies have shown that significant levels of modifications are present in the vicinity of genes prior to transcription, resulting in a poised chromatin state [29,33]. Furthermore, histone modifications may serve as recognition sites for the recruitment of effector modules. Several bromo-, PHD- and chromodomain proteins have been reported to bind to specifically acetylated and methylated lysine residues [3,9,34,35]. Mef2 is a key transcription factor in differentiation known to be inhibited by histone deacetylases in myoblasts. During differentiation to myotubes, Mef2 recruits histone acetyl transferases essential for differentiation, such as p300 and GRIP [23]. Sequences associated with gain of H4ac in differentiation showed overrepresentation of Mef2 binding sites, relative to sequences losing H4ac.

## Conclusions

In summary, our results agree with the *histone code hypothesis,* as defined in [8]. We demonstrated that the average transcript levels associated with combinations of modifications are not simply (additively) related to those associated with individual modifications. Investigating the dynamics of modifications within one cell line during differentiation, we showed that the appearance of modifications is, by itself, rarely associated with higher transcript levels. This is consistent with the view that these marks are a prerequisite for, but not a sufficient driving force or a necessary result of, transcription. Histone modifications may primarily function as signaling marks for specific effectors and could tremendously increase the combinatorial possibilities in the regulation of transcription. A full understanding of the role of histone marks in transcriptional regulation will emerge only when the interactions of multiple modifications with their effectors are considered as a whole.

## Materials and methods

### *Cell culture*

As histone modifications are known to be cell-type specific, we used cultured cell lines to minimize heterogeneity and to achieve clearly defined cell states. C2C12 cells were obtained from Professor Jakob Schmidt (Department of Biochemistry and Cell Biology, State University of New York, Stony Brook, NY, USA) and cultivated at 5% $CO_2$ and 37°C in Dulbecco's modified Eagle's medium (Gibco) supplemented with 1% penicillin/streptomycin (Gibco) and 10% fetal calf serum (Biochrom). Mononucleate C2C12 myocyte cells were harvested before reaching 70% confluence. To induce differentiation, cells were cultured with Dulbecco's modified Eagle's medium and 2% horse serum (Biochrom) and maintained for 48 h, when more than 90% of the cells had fused into myotubes [36]. HL-1 cells were provided by Professor William C. Claycomb (Departments of Biochemistry and Molecular Biology and Cell Biology and Anatomy, Louisiana State University Medical Center, New Orleans, LA, USA) and cultured as described [37]. HL-1 cells were harvested for experiments when showing maximum contraction.

### *Real-time PCR*

All real-time PCRs were measured on an ABI Prism 7700 (Applied Biosystems) in 10 μl reaction volume with 2 × Sybr Green I master mix (ABgene) and 100 nM primer in duplicate. Primers were designed using PrimerExpress software (Applied Biosystems) to amplify 100- to 150-bp fragments. Fold changes were calculated using the relative quantification method of $\Delta\Delta C_t$. Fold changes for expression analysis were normalized to

Author's personal copy

Hprt1. Primers used for expression analysis are given in Table S9. Fold change enrichments of ChIP samples were measured relative to input. Primers used for ChIP–chip confirmation are given in Table S10.

### Expression analysis

For each cell type (myoblasts, myotubes, and cardiomyocytes) six samples of total RNA were isolated from three different cell passages using Trizol (Invitrogen) and subsequently DNase digested (Promega) according to the manufacturer's instructions. RNA from three independent isolations was pooled, giving two biological replicates per cell type. RNA quality was confirmed by Bioanalyzer (Agilent) analysis, and RNA was subsequently labeled and hybridized by NimbleGen. Array intensities and calculated fold changes were confirmed for 15 transcripts in each cell type in replicate by real-time PCR (Figs. S2A and S2B).

### Chromatin immunoprecipitation

ChIP experiments were performed in duplicate for five different antibodies in parallel as described [38] with the following modifications: cells were cross-linked for 10 min at 37°C; sonication was carried out with a Branson 250 Sonifier with 12 pulses at power setting of 6 and 100% duty cycle for 30 s and 2 min on ice between pulses. Magnetic protein A/G beads were obtained from Invitrogen. For immunoprecipitation the following rabbit antibodies were used: anti-H3K9K14ac (Upstate 06-599, Lot 29505), anti-K5K8K12K16 ac (Upstate 06-866, Lot 29532), H3K4me2 (Abcam ab7766, Lot 66726), H3K4me3 (Abcam ab8580, Lot 77499), and rabbit normal IgG (Santa Cruz Biotechnology sc-2027, Lot K0304). These antibodies were previously shown to be specific in ChIP–chip experiments [17]. ChIP quality before linear amplification was confirmed by real-time PCR for three histone modified sites and four sites showing no enrichment, per cell line and modification as described above (in total 210 single verifications, Fig. S3A). Normal rabbit ChIPs gave no enrichment over input for any of these sites and yielded less than 1% DNA compared to specific antibodies and therefore did not yield enough DNA to amplify for ChIP–chip applications.

### Amplification of DNA

Amplification of ChIPed DNA and input control was carried out as described [39] except only one round of amplification with 20 cycles was performed. Amplified samples were purified using Wizard SV PCR purification kits (Promega) according to the manufacturer's instructions. DNA quality was confirmed by Bioanalyzer (Agilent) measurements. Samples were labeled and hybridized according to NimbleGen standard procedures. After bioinformatic analysis of the ChIP–chip data, 12 modified sites and 9 sites showing no enrichment were validated using amplified material by real-time PCR for each modification per cell type in duplicate (in total 504 single verifications, Fig. S3B).

### Design of arrays

Human or mouse transcripts expressed in heart, skeletal, or smooth muscle were selected from several sources as listed in Table S11. All identifiers were mapped to Ensembl version 26, human–mouse orthologs were identified, and redundant entries were removed. These transcripts were selected to be represented on the expression arrays. For the respective set of genes, the human–mouse conserved noncoding blocks (CNBs) in the 5-kb region upstream of annotated TSSs and in the first intron up to 10 kb downstream of each TSS were considered. For genes with less than 10% CNB sequence, a fixed region of 2.2 kb upstream and 0.8 kb of the first intron was selected. Additionally, the first exon of each gene was represented. The selected regions were repeat-masked and probes were designed by NimbleGen.

### ChIP microarray preprocessing

Probes were mapped to the mouse genome assembly mm8 using BLAT [40], allowing up to one mismatch per 50-mer probe, resulting in 389,918 genomic positions. Intensities of each channel were normalized and log-transformed using VSN. Log-ratio enrichment levels for each probe were calculated by subtraction of Cy3 (input) from Cy5 (ChIP sample).

### Identification of modified sites

Normalized probe levels for biological replicates were averaged and smoothed along chromosomal coordinates using a sliding window method. For each probe position the smoothed probe level was computed as the median over the probe levels in an 800 bp window centered at that position. To allow for different efficiencies of antibodies, a cutoff was defined for each type of histone modification separately, by repeating the above smoothing procedure on data where probe positions were randomly permuted and calculating the 99% quantile. We called a probe enriched if it had a smoothed probe level greater than the cutoff. Enriched probes were merged into enriched regions if less than 600 bp apart. Resultant regions of at least three probes were called modified sites.

### Identification of domains

A modified domain is a combination of different modified sites within one cell type. The genomic locations were required to overlap by at least 75% of the length of the smallest contained modified site. We mapped a domain to a TSS if its middle position was located up to 5 kb upstream or within the respective transcribed region.

### Expression microarray preprocessing

Probes were mapped to the mouse genome assembly mm8 using BLAT [40], allowing up to one mismatch per 24-mer probe. Only probes matching annotated Ensembl transcripts (version 39, June 2006) were further analyzed. Probe intensities were background corrected, quantile normalized, and summarized into transcript expression levels using the median-polish procedure [41].

### Expression level differences between transcript categories

Transcripts were categorized according to modified sites or modified domains. Pair-wise Wilcoxon rank sum tests were employed to assess differences in expression between transcript categories. The resulting $p$ values were corrected for multiple testing using the Bonferroni procedure. A corrected $p < 0.05$ was interpreted as evidence of transcripts in one category having expression levels significantly different from those in the other category.

### Transcript expression categories

Transcripts were categorized according to expression levels as non- (<7 arbitrary scanner units, $\log_2$ scale), low- (7–8.5), medium- (8.5–10.9), and high-expressed (>10.9). The first three cutoffs were determined by RT-PCR for transcripts of known transcriptional status. The cutoff for high expression is the 90% quantile of all data above 8.5.

### Identification of differentially expressed genes

Differential expression of transcripts between cell types was assessed using the moderated $t$ statistic of the empirical Bayes approach in the Bioconductor [42] package Limma [43]: $p$ values for differential expression were adjusted for multiple testing using Benjamini and Yekutieli's method for control of the false discovery rate [44]. Transcripts with an adjusted $p$ value smaller than or equal to 0.05 were considered to be differentially expressed.

### Linear model: expression levels

For the regression of the normalized transcript expression levels on histone modifications, we used the data of undifferentiated and differentiated C2C12 cells. We fitted a linear regression model to include, for each transcript, expression levels in each cell type, presence of modified sites, and median GC content of the microarray probe that mapped to the transcript. A $t$ test was

Author's personal copy

computed for the coefficient of each effect to assess whether it was significantly different from 0 (Table S5).

### *Logistic regression model*

For the logistic regression, the response was an indicator variable for differential upregulation or downregulation of a transcript, while the predictors were indicator variables for transcripts gaining or losing modifications during differentiation of C2C12 cells (Tables S7 and S8, respectively). We used the data of undifferentiated and differentiated C2C12 cells.

### *Gene Ontology (GO) associations to gene groups*

To analyze the association of differentially expressed transcripts with GO categories, the transcripts were mapped to genes. The association of gene groups to GO [45] terms was assessed according to Alexa et al. [46] through a conditional hypergeometric test for overrepresentation using a $p$ value threshold of 0.001. In a prefiltering step, we excluded from this analysis all transcripts whose expression levels across samples had an interquartile range below 0.5 $\log_2$ units.

### *Implementation*

The computational methods were implemented in the R programming language, using packages from the Bioconductor project [42], and have been published elsewhere [47].

### *Mef2 binding site analysis*

We compared the occurrence of the TRANSFAC [48] Mef2 binding site V $MMEF2 in a 1.5 kb window between 816 domains with H4ac gain and 790 domains with H4ac loss during differentiation. We used MAST [49] with a minimum sequence $p$ value of 0.1 and an $E$ value of 160. The group specificity score [50] was used to identify significant overrepresentation.

## Acknowledgments

S.S. and W.H. acknowledge support from the European Union (HeartRepair Grant 018630). S.S. is supported by the German government (BioProfile Grant 0313040B). We thank Steffen Grossmann for assistance with the array design, Jenny Schlesinger and Stefanie Hammer for their laboratory assistance, Matt Richie for bioinformatic support, and Richard Bourgon and Andreas Ladurner for helpful comments on the manuscript.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.08.010.

## References

[1] X. de la Cruz, S. Lois, S. Sanchez-Molina, M.A. Martinez-Balbas, Do protein motifs read the histone code? BioEssays 27 (2005) 164–175.

[2] H.C. Beck, et al., Quantitative proteomic analysis of post-translational modifications of human histones, Mol. Cell. Proteomics 5 (2006) 1314–1325.

[3] L. Verdone, M. Caserta, E. Di Mauro, Role of histone acetylation in the control of gene expression, Biochem. Cell. Biol. 83 (2005) 344–353.

[4] C. Martin, Y. Zhang, The diverse functions of histone lysine methylation, Nat. Rev., Mol. Cell Biol. 6 (2005) 838–849.

[5] I. Sims, J. Robert, K. Nishioka, D. Reinberg, Histone lysine methylation: a signature for chromatin function, Trends Genet. 19 (2003) 629–639.

[6] M. Shogren-Knaak, et al., Histone H4-K16 acetylation controls chromatin structure and protein interactions, Science 311 (2006) 844–847.

[7] A.J. Bannister, T. Kouzarides, Reversing histone methylation, Nature 436 (2005) 1103–1106.

[8] B.D. Strahl, C.D. Allis, The language of covalent histone modifications, Nature 403 (2000) 41–45.

[9] R. Margueron, P. Trojer, D. Reinberg, The key to development: interpreting the histone code? Curr. Opin. Genet. Dev. 15 (2005) 163–176.

[10] T. Jenuwein, C.D. Allis, Translating the histone code, Science 293 (2001) 1074–1080.

[11] M.S. Cosgrove, C. Wolberger, How does the histone code work? Biochem. Cell. Biol. 83 (2005) 468–476.

[12] B.M. Turner, Cellular memory and the histone code, Cell 111 (2002) 285–291.

[13] S.K. Kurdistani, M. Grunstein, Histone acetylation and deacetylation in yeast, Nat. Rev., Mol. Cell Biol. 4 (2003) 276–284.

[14] B.E. Bernstein, et al., A bivalent chromatin structure marks key developmental genes in embryonic stem cells, Cell 125 (2006) 315–326.

[15] T.Y. Roh, S. Cuddapah, K. Cui, K. Zhao, The genomic landscape of histone modifications in human T cells, Proc. Natl. Acad. Sci. USA 103 (2006) 15782–15787.

[16] C.L. Liu, et al., Single-nucleosome mapping of histone modifications in S. cerevisiae, PLoS Biol. 3 (2005) e328.

[17] D.K. Pokholok, et al., Genome-wide map of nucleosome acetylation and methylation in yeast, Cell 122 (2005) 517–527.

[18] B.E. Bernstein, et al., Methylation of histone H3 Lys 4 in coding regions of active genes, Proc. Natl. Acad. Sci. USA 99 (2002) 8695–8700.

[19] B.E. Bernstein, et al., Genomic maps and comparative analysis of histone modifications in human and mouse, Cell 120 (2005) 169–181.

[20] D. Schubeler, et al., The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote, Genes Dev. 18 (2004) 1263–1271.

[21] G. Liang, et al., Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome, Proc. Natl. Acad. Sci. USA 101 (2004) 7357–7362.

[22] K.K. Tomczak, et al., Expression profiling and identification of novel genes involved in myogenic differentiation, FASEB J. 18 (2004) 403–405.

[23] T.A. McKinsey, C.L. Zhang, E.N. Olson, Control of muscle development by dueling HATs and HDACs, Curr. Opin. Genet. Dev. 11 (2001) 497–504.

[24] C.M. Smith, Quantification of acetylation at proximal lysine residues using isotopic labeling and tandem mass spectrometry, Methods 36 (2005) 395–403.

[25] D.J. Clark, C.H. Shen, Mapping histone modifications by nucleosome immunoprecipitation, Methods Enzymol. 410 (2006) 416–430.

[26] G.-C. Yuan, et al., Genome-scale identification of nucleosome positions in S. cerevisiae, Science 309 (2005) 626–630.

[27] E. Segal, et al., A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.

[28] F. Ozsolak, J.S. Song, X.S. Liu, D.E. Fisher, High-throughput mapping of the chromatin structure of human promoters, Nat. Biotechnol. 25 (2007) 244–248.

[29] R. Schneider, et al., Histone H3 lysine 4 methylation patterns in higher eukaryotic genes, Nat. Cell Biol. 6 (2004) 73–77.

[30] T.H. Kim, et al., A high-resolution map of active promoters in the human genome, Nature 436 (2005) 876–880.

[31] F. Miao, R. Natarajan, Mapping global histone methylation patterns in the coding regions of human genes, Mol. Cell. Biol. 25 (2005) 4650–4661.

[32] A.B. Brinkman, et al., Histone modification patterns associated with the human X chromosome, EMBO Rep. 7 (2006) 628–634.

[33] A.L. Clayton, C.A. Hazzalin, L.C. Mahadevan, Enhanced histone acetylation and transcription: a dynamic perspective, Mol. Cell 23 (2006) 289–296.

[34] X. Shi, et al., ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression, Nature 442 (2006) 96–99.

Author's personal copy

[35] J. Wysocka, et al., A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling, Nature 442 (2006) 86–90.
[36] S. Liu, et al., Interaction of MyoD family proteins with enhancers of acetylcholine receptor subunit genes in vivo, J. Biol. Chem. 275 (2000) 41364–41368.
[37] W.C. Claycomb, et al., HL-1 cells: a cardiac muscle cell line that contracts and retains phenotypic characteristics of the adult cardiomyocyte, Proc. Natl. Acad. Sci. USA 95 (1998) 2979–2984.
[38] C.E. Horak, et al., GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp–chip analysis, Proc. Natl. Acad. Sci. USA 99 (2002) 2924–2929.
[39] V.R. Iyer, et al., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, Nature 409 (2001) 533–538.
[40] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[41] R.A. Irizarry, et al., Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Res. 31 (2003) e15.
[42] R.C. Gentleman, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (2004) R80.
[43] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004) (Article3).
[44] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Stat. 29 (2001) 1165–1188.
[45] M. Ashburner, et al., Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.
[46] A. Alexa, J. Rahnenfuhrer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, Bioinformatics 22 (2006) 1600–1607.
[47] J. Toedling, O. Sklyar, W. Huber, Ringo—an R/Bioconductor package for analyzing ChIP–chip readouts, BMC Bioinformatics 8 (2007) 221.
[48] V. Matys, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (2006) D108–D110.
[49] T.L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, Bioinformatics 14 (1998) 48–54.
[50] J.D. Hughes, P.W. Estep, S. Tavazoie, G.M. Church, Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, J. Mol. Biol. 296 (2000) 1205–1214.

### 2.1.4 Prediction of cardiac transcription networks

Toenjes M, Schueler M, Hammer S, Pape UJ, Fischer JJ, Berger F, Vingron M, **Sperling S**. Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes. *Mol Biosyst* 2008;**4**:589-598.

In this study, expression levels of a comprehensive set of 46 cardiac genes in 190 heart biopsies derived from healthy individuals and patients with a broad range of cardiac malformations were analyzed. Therefore the study design was inverse to the previous genome-wide array analysis, which focused on distinct cardiac phenotypes. The selected genes include transcription factors and potential downstream targets known from literature as well as those identified in the gene expression microarray analysis (Kaynak et al., 2003). To build the bridge between disease phenotypes and transcriptional networks, first a detailed phenotype ontology of the heart malformations was delineated. Next, the expression levels in normal and malformed hearts were placed within the context of the corresponding phenotype. Application of linear models that analyze gene expression integrating age and gender dependencies revealed transcriptional changes between distinct patient groups. Additionally, independent of corresponding phenotypes, groups of correlated genes were identified based on similar expression patterns of genes both in normal and malformed hearts. Combining these approaches, genes were identified that appeared to be specifically associated with certain phenotypes and showed correlated expression in general. Finally, based on correlated gene expression and transcription factor binding site prediction, which was optimized on the heart-specific ChIP data set described above (Fischer et al., 2008a), cardiac regulatory networks could be constructed. As proof of principle, these networks point out novel as well as known regulatory dependencies and moreover explain parts of the observed transcription patterns in diseased cardiac samples.

#### *Phenotype ontology*

To enable the selection of a balanced patient population allowing the separation of disease- or tissue-specific expression patterns, 190 human ventricular and atrial cardiac tissues were used. The clinical characterization comprised 250 features of morphological, hemodynamical and therapeutical information which are stored in the d-matrix database for detailed analysis and visualization (Seelow et al., 2004).

A phenotype ontology was delineated to compress the complex and partially overlapping disease characteristics. A list of 26 disease parameters in addition to tissue type, gender and age was compiled for each patient, including descriptors like “interatrial septal defect” and “right ventricle dilation” (**Figure 6**).

To define groups of patients with similar phenotypes, a complete linkage hierarchical clustering approach using this phenotype ontology was carried out. Patients were assigned to eight meta-phenotypes that represent specific clusters derived from cutting the dendrogram at a certain height as shown in **Figure 6**. *E.g.* the cluster *TOF-III* contains patients characterized by interatrial septal defects as well as stenosis and/or dilation of the main pulmonary artery in addition to the classical features of Tetralogy of Fallot (TOF), namely interventricular septal defect, overriding aorta, right ventricular hypertrophy and right ventricular outflow tract stenosis.

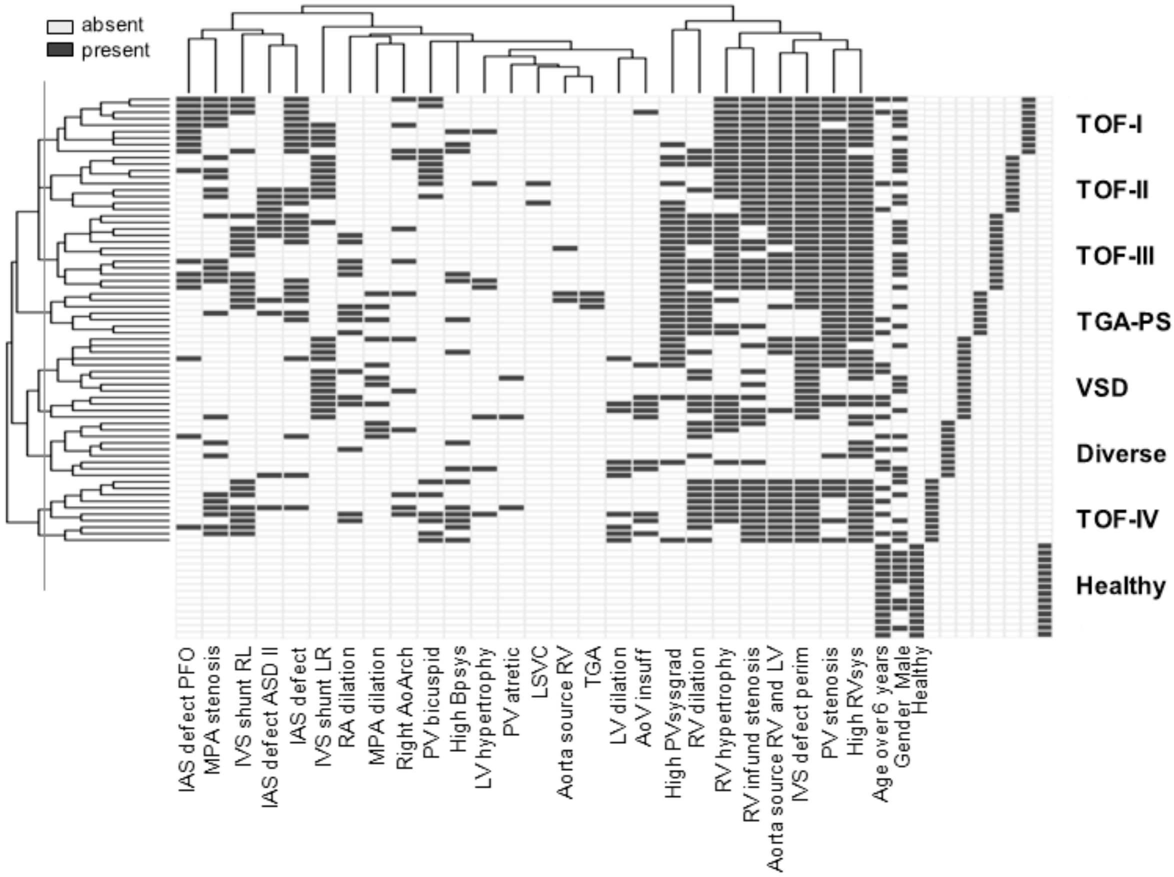


**Figure 6. Hierarchical clustering of cardiac disease phenotype criteria and assignment of patients with similar characteristics into meta-phenotype groups of ventricular samples.** The phenotype information for gender, age and disease state is indicated. Each row represents a single heart sample. The line indicates the used cut-off for assignment of meta-phenotypes.

### *Gene expression analysis*

To characterize the transcription patterns of the patient cohort, expression levels of 42 genes were measured by quantitative real-time PCR. For an initial overview of the expression data, hierarchical clustering using complete linkage was applied revealing clear differences between atrial and ventricular samples. Several of the genes displaying chamber-specific expression have already been described in studies of human and mouse myocardium. *E.g. NPPA*, *NR2F1*, *MYH6*, *MYL7* and *TAGLN* predominate in atria (Ellinghaus et al., 2005), whereas *Irx4* and *Myl2* are restricted to ventricles (Tabibiazar et al., 2003).

To extract the influence of phenotype clusters on gene activity considering known confounding factors such as age and gender (Kaynak et al., 2003), a linear modelling technique was used. A linear model $Y = \alpha_{meta-phenotype} + \beta_{age} + \gamma_{gender}$ was computed, where Y is the predicted expression value, $\alpha_{meta\text{-}phenotype}$ is the coefficient for each individual patient group sharing the same meta-phenotype, $\beta_{age}$ is the coefficient for our two age categories *young* (younger than 6 years) and *old* as well as $\gamma_{gender}$ determining gender specific effects. Deregulated genes were observed for almost all meta-phenotypes, except the cluster *Diverse,* which contains a mixture of different minor phenotypes excluding VSD and with a regular aortic source from the right ventricle. The other meta-phenotypes, characterized by distinct and moderate to severe abnormalities, have specific molecular portraits, such as *TBX20* and *MEF2C* being upregulated in patients with TOF and main pulmonary artery abnormalities (cluster TOF-III), whereas *TBX5* being only downregulated in patients with TOF and bicuspid pulmonary valve (cluster TOF-II). Some genes appear to be significantly deregulated in all diseased samples, indicated by an opposite regulation in the healthy cluster, *e.g. MEF2A* is upregulated in all disease meta-phenotypes. Based on the transcriptional profiles, previously not disease-associated candidate genes could be identified by this approach, like *TBX20* and *DPF3*, which eventually have been further investigated (Hammer et al., 2008; Lange et al., 2008).

### *Correlated gene expression*

To finally build transcription networks, groups of genes that show a correlated pattern of expression both in normal and diseased samples were dissected. To assess correlation between individual gene pairs, their Pearson correlation coefficient was computed over all samples in the dataset. Using random experiments the statistical significance of found correlation coefficients was evaluated. Subsequently, using hierarchical clustering, 19 clusters of significant distances between individual genes were derived. Centered expression vectors were sorted by the defined meta-phenotypes and similar expression patterns of genes can clearly be seen in normal and diseased tissue samples (**Figure 7**). The TFs *TBX20* and *MEF2C* displayed correlated expression patterns and strikingly, both are upregulated in patients with TOF-III analyzed with the linear model.

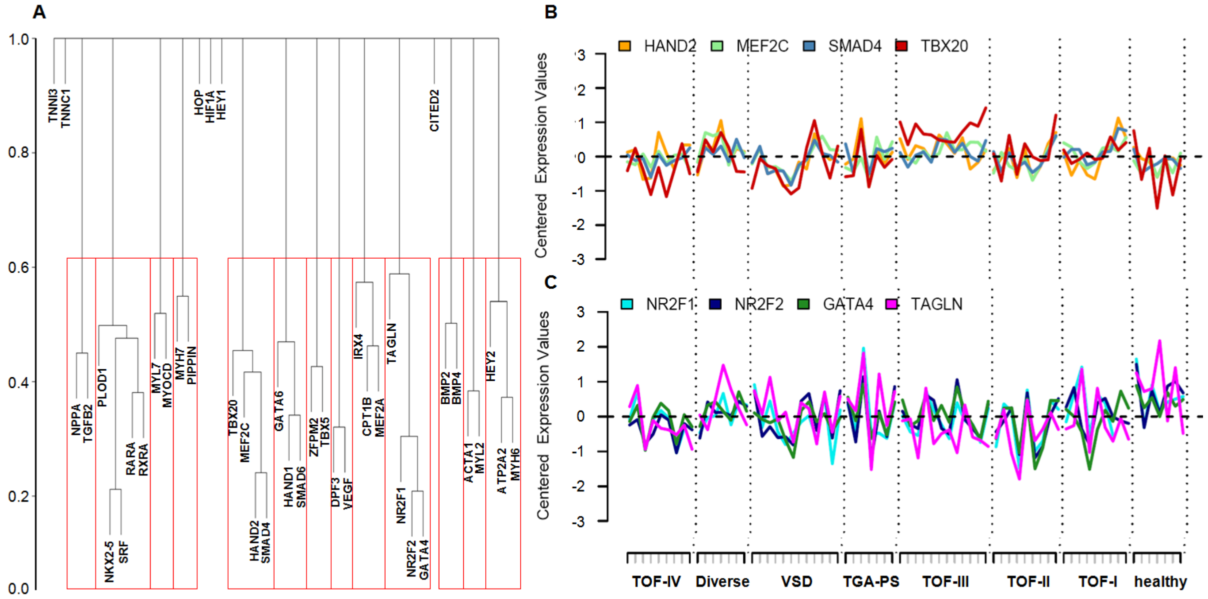


**Figure 7. Correlation of gene expression**. (A) Cluster dendrogram showing 13 correlated gene groups. Clustering was derived by cutting the cluster tree at the 1x10$^{-3}$ quantile of a random distribution. Y-axis indicates cluster distances. (B + C) Example of two correlated gene groups showing highly correlated patterns of expression in samples of healthy individuals and patients. Centered expression vectors were sorted by defined meta-phenotypes.

It might occur that two genes show correlated expression over a large set of samples but are strongly deregulated in a specific meta-phenotype *e.g.* due to a breakdown of TF networks. This would lead to a decreased correlation coefficient and loss of cluster assignment. Although, the found gene clusters could be a product of background noise. To consider the robustness of the correlated gene groups, the correlation analysis was repeated successively eliminating one meta-phenotype and taking the maximal correlation coefficient. While the resulting cluster dendrogram shows some changes in cluster association for single genes and subclusters, the majority of clusters stayed intact thereby confirming our found correlated gene groups. For example, the correlated gene group comprising *HAND2*, *MEF2C*, *SMAD4* and *TBX20* (**Figure 7**) was recovered and further enlarged by *DPF3* and *VEGF*, formerly building a separate correlated gene group, as well as *HIF1A* that had not been assigned to any correlated group beforehand. Even in the initial correlation analysis which considered all meta-phenotypes, *DPF3* showed significant correlation with all four genes and *HIF1A* and

*VEGF* with three and two, respectively. Finally, the significant maximal correlation coefficients over single meta-clusters were computed.

Showing strong correlation over the high number of different samples, it is likely that a correlated gene group is co-regulated by the same transcription factor. To discover TFs that have binding sites in the promoters of all of the genes belonging to one correlated gene group predictions of performing transcription factor binding site (TFBS) were performed. To find the best settings the prediction was optimized using wet lab data generated in the previous sections.

***Optimization of transcription factor binding site prediction using Chip data***

To predict possible binding of TFs to the promoter regions of the gene set, two different matching algorithms were used, one proposed by Rahmann *et al.* (Rahmann et al., 2003) (Rahmann-Matcher) and the Match algorithm provided by TRANSFAC (Kel et al., 2003).

The amount of promoter sequence as well as the use of conservation information taken for TFBS prediction varies among different studies (Nelander et al., 2005; Kim and Kim, 2006; Goff et al., 2007) and generally, the sequence length considered is positively correlated with an increase of noise (Rahmann et al., 2003). To make the TFBS prediction as biologically meaningful as possible with regard to these settings, the in vivo DNA-binding data obtained from ChIP-chip for Gata4, Mef2a and Nkx2-5 (Fischer et al., 2008a) were integrated to define the most suitable parameter settings. This approach was considered to be more applicable compared to arbitrarily chosen settings. To find an optimal balance between length of promoter sequence and noise level in the prediction of assigned binding sites, different upstream and downstream distances were used as an optimization criterion. Besides the amount of promoter sequence, the level of conservation was integrated as an optimization parameter. The third parameter optimized was the matching algorithm.

Finally the following scoring function was implemented:

$$S = A * B, \quad \text{where } A = \frac{true\ predictions}{all\ predictions} \quad \text{and} \quad B = \frac{predicted\ peaks}{all\ peaks}$$

The score $S$ comprises two factors ranging from 0 to 1 that measure different aspects of the TFBS predictions. $A$ measures the fraction of true among all predictions and $B$ measures the capability of predicting a ChIP peak. The product of both factors was used as a scoring function to reduce influences of extreme values in only one factor. The optimization process was performed for all three TFs and the average over the three individual scores computed for each setting was reported.

Applied to this scoring scheme, the TRANSFAC Match algorithm in general achieved higher scores than the Rahmann-Matcher. Furthermore, the fraction of true predictions decrease with the length of sequence used, which is likely due to an increase in noise level. However, TFBSs identified by ChIP can be observed at any distance from the transcription start sites. While the fraction of true predictions could be enhanced by using more stringent conservation settings, the amount of TF ChIP peaks predicted by the two algorithms heavily dropped at higher conservation levels. This finding is supported by observations that actual binding sites of TFs might be slightly modified during evolution for example to enable adaptation of TF binding affinity (Gerland et al., 2002; Copley et al., 2007). Using the new scoring function which incorporates both measures, the optimal setting was 1.250 bp upstream and 500 bp downstream together with a conservation level of 60 %. Subsequently, these settings and the TRANSFAC Match algorithm were used for the TFBS prediction.

### ***Regulatory cardiac networks***

Finally, regulatory networks based on identified *correlated gene groups* and predicted TFBSs representing the underlying potential regulatory dependencies were constructed. **Figure 8** displays two graphs representing predicted regulatory subnetworks for the *HAND2*, *MEF2C*, *SMAD4*, *TBX20* and *GATA4*, *NR2F1*, *NR2F2*, *TAGLN* correlated gene group. Comparing these predictions with the literature and observed ChIP-chip data, all except the two bindings to SMAD4 have been proposed in literature (Nkx2-5 → *Mef2C* (Skerjanc et al., 1998; Tanaka et al., 1999)), found in the ChIP data (Nkx2-5 → *Hand2*/*Tbx20*, Gata4 → *Tbx20*) or both (Gata4 → *Hand2 (McFadden et al., 2000)*/*Mef2C (Dodou et al., 2004)*). Interestingly, both *TBX20* and *MEF2C* are specifically upregulated in patients within the TOF-III cluster and the network sheds light on potential upstream regulators. The regulation of *TBX20* is not well known so far. The only described signalling molecule upstream of *Tbx20* are Bmp2 (Plageman and Yutzey, 2004), and TFAP2C as a direct regulator (Hammer et al., 2008). Identification of NKX2-5 and GATA4 as common regulators reveals them as interesting candidate genes to be responsible for the transcription pattern of the phenotype cluster. A causative connection is suggestive and mutations in both TFs have already been linked to TOF (Benson et al., 1999; Nemer et al., 2006). Measuring *Tbx20* levels in siRNA knockdown experiments of the respective TFs showed reduction of *Tbx20* mRNA levels by 20-50 % (Fischer et al., 2008a). These results demonstrate that binding of Nkx2-5 and Gata4 is indeed functional and activates *Tbx20* expression. Potentially, posttranscriptional modifications could explain why neither of the two TFs are part of this correlated gene group (Herrgard et al., 2003). Concerning the *GATA4*, *NR2F1*, *NR2F2* and *TAGLN* correlated gene group, several TFs were found that had predicted binding sites in all promoters of the four genes. Among them are TBX5, GATA6 as well as GATA4 and the two NR2F factors. Identification of the latter three TFs is quite remarkable as all the TFs present in this correlated gene group appear to show regulatory interactions with each other that could explain the identified correlation.

In order to substantiate the predicted TF regulations the transcription factor affinity prediction (TRAP) algorithm was integrated (Roider et al., 2007). In case of the *GATA4*, *NR2F1*, *NR2F2* and *TAGLN* correlated gene group, both GATA4 and GATA6 appeared to have all four gene promoters in their TOP-10 TRAP affinity tables. This underlines the results of the TFBS prediction in which they also showed binding to all group members. Furthermore, it highlights these GATA proteins as potential auto-regulatory key factors in the given subnetwork. In addition, SMAD6 showed high affinity to three of the four correlated genes, namely *NR2F1*, *NR2F2* and *TAGLN* and was predicted to be bound by GATA4 itself, which implies a functional role further downstream in the regulatory cascade.

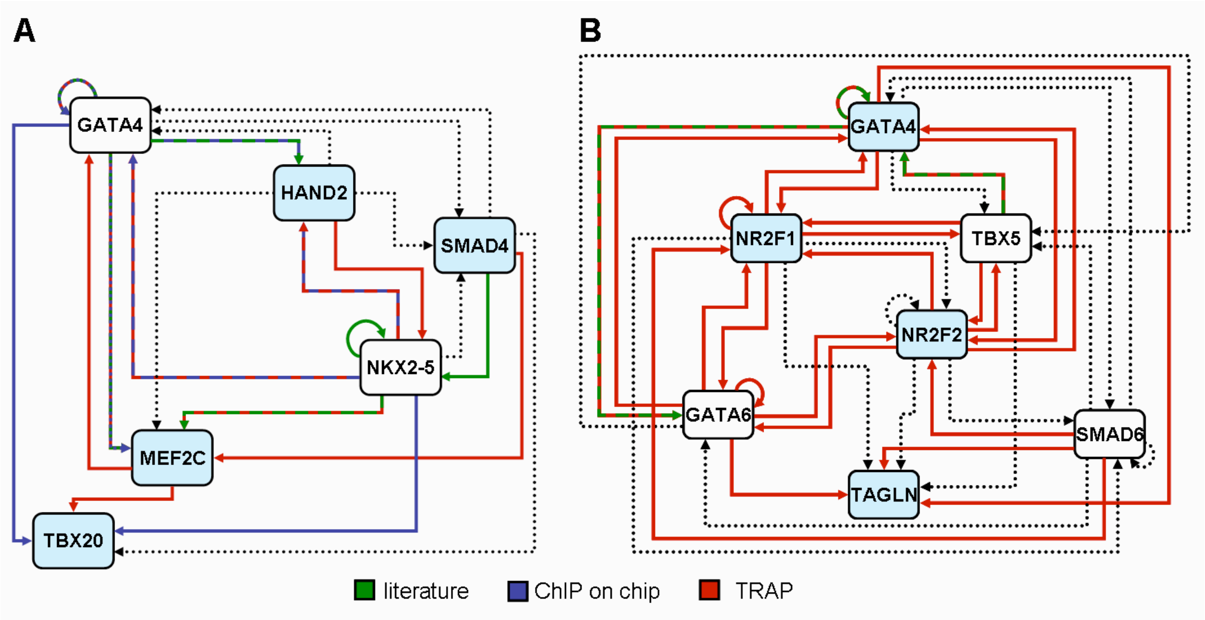


**Figure 8. Predicted regulatory networks for two correlated gene groups.** Genes comprising a group are marked gray and TFs predicted to bind all of them are highlighted by bold circles.

Taken together, a variety of methods to predict transcription networks had been proposed recently, however, integrative approaches combining complex clinical phenotype data with advanced bioinformatic and biochemical methods were still lacking. The presented first cardiac transcription networks are based on predicted transcription factor binding sites and gene expression profile disturbances in samples of congenital malformed hearts. The idea to use this complex phenotype was driven by the assumption that a broad panel of cardiac phenotypes associated with a range of genomic sequence variations and different modifiers, potentially underlying the phenotype, would lead to ranges of expression patterns rather than distinct profiles. This enabled the identification of transcriptional dependencies. Finally, several methods such as linear models, correlation analyses based on expression profiles as well as the prediction of *cis* regulatory elements to predict the transcription networks were combined. Furthermore, the obtained networks could be verified by data derived from literature and ChIP.

Though, one has to bear in mind that expression profiling detects only transcript abundance and not the activity of the encoded proteins. Posttranslational modifications, such as phosphorylation of MEF2 proteins, NKX2-5 and GATA4 (Molkentin et al., 1996; Kasahara and Izumo, 1999; Ornatsky et al., 1999; Charron et al., 2001), allow fine-tuning of gene activity independent of expression levels and add an additional layer of complexity to the network of transcription factors operating in the developing heart. In addition, several evolutionary conserved micro-RNAs function as regulators of target RNAs, one of which is *Mir1* that negatively regulates cardiac growth during mouse development by inhibiting translation of *Hand2* (Zhao et al., 2005).

# Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes†‡

Martje Toenjes,§[a] Markus Schueler,§[ab] Stefanie Hammer,[a] Utz J. Pape,[bc]
Jenny J. Fischer,[a] Felix Berger,[d] Martin Vingron[b] and Silke Sperling*[a]

*Received 8th January 2008, Accepted 28th February 2008*
*First published as an Advance Article on the web 2nd April 2008*
**DOI: 10.1039/b800207j**

We present an integrative approach combining sophisticated techniques to construct cardiac gene regulatory networks based on correlated gene expression and optimized prediction of transcription factor binding sites. We analyze transcription levels of a comprehensive set of 42 genes in biopsies derived from hearts of a cohort of 190 patients as well as healthy individuals. To precisely describe the variety of heart malformations observed in the patients, we delineate a detailed phenotype ontology that allows description of observed clinical characteristics as well as the definition of informative meta-phenotypes. Based on the expression data obtained by real-time PCR we identify specific disease associated transcription profiles by applying linear models. Furthermore, genes that show highly correlated expression patterns are depicted. By predicting binding sites on promoter settings optimized using a cardiac specific chromatin immunoprecipitation data set, we reveal regulatory dependencies. Several of the found interactions have been previously described in literature, demonstrating that the approach is a versatile tool to predict regulatory networks.

## Introduction

So far a variety of methods have been used to identify regulatory networks from gene expression data, often called 'reverse-engineering'.[1] The spectrum ranges from one-dimensional or two-dimensional (bi-)clustering approaches to techniques such as Bayesian network learning algorithms or ordinary differential equations.[2–4] Some methods thereby rely on the assumption that regulators and target genes show dependencies in their expression patterns (*e.g.* correlation).[5] Other approaches aim to identify functional *cis* regulatory sites pointing to binding of specific transcription factors (TFs).[6] Finally, there exist a number of biochemical techniques that identify regulatory networks from *in vitro* binding sites using chromatin immunoprecipitation (ChIP) or direct perturbations of TFs.[7] However, it is known that the performance of all these different techniques is dependent on the underlying dataset.[8] In this study, we present an integrative approach to identify regulatory networks comprising bioinformatic as well as biochemical techniques taking the human heart as a model. We combine several methods such as linear models, correlation analyses based on expression profiles as well as the prediction of *cis* regulatory elements and verify our predicted networks using data derived from literature and ChIP.

The heart is the first functional organ during embryogenesis and the one most susceptible to disease. A rapidly growing number of factors have been shown to be involved in regulating the pattern and timing of the expression of genes responsible for the cardiac lineage determination, heart chamber formation, valvulogenesis and conduction-system development.[9] Spatiotemporal and quantitative regulation of cardiac TFs must occur in a precise manner to ensure fine regulation of downstream targets. The complexity of these molecular cascades during development may explain the sensitivity of the heart to perturbations before birth and into old age. Congenital heart diseases (CHD) are the most common birth defects in humans. They arise during development of the embryo and affect 1 in every 100 live births and an even higher number in miscarriages.[10,11]

To gain insight into the formation of cardiac anomalies molecular genetic studies of human patient populations have been carried out. Linkage analysis and candidate-gene approaches have led to the identification of several gene mutations causing CHD (*e.g. CITED2*, *GATA4*, *NKX2-5* and *ZIC3*).[12–15] However, most heart malformations display variable expressivity and penetrance pointing to a multifactorial and multigenic basis. In humans and mice similar mutations can

---

[a] *Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany. E-mail: sperling@molgen.mpg.de; Fax: +49-30-84131699; Tel: +49-30-84131232*
[b] *Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany*
[c] *Department of Mathematics and Computer Science, Free University of Berlin, Berlin, Germany*
[d] *Department of Pediatric Cardiology, German Heart Center, Berlin, Germany*

† This article is part of a *Molecular BioSystems* 'Emerging Investigators' issue highlighting the work of outstanding young scientists at the chemical- and systems-biology interfaces.

‡ Electronic supplementary information (ESI) available: Hierarchical clustering of cardiac disease phenotype criteria for atrial samples; overview of measured correlations and assigned *p*-values; clustering tree of genes with correlated expression patterns for subsets of phenotype clusters; optimization of TF binding site prediction using TRANSFAC and Rahmann-Matching algorithms; information about selected genes for expression analysis; TRANSFAC TF binding matrices assigned to TFs selected for expression analysis. See DOI: 10.1039/b800207j

§ These authors contributed equally.

This journal is © The Royal Society of Chemistry 2008

cause a variety of phenotypes from one family, individual or inbred strain, respective to another. Heterozygous mutations in the homeobox transcription factor NKX2-5 in human can lead to such diverse abnormalities as atrial septal defects (ASDs), ventricular septal defects (VSDs), Ebstein's anomaly of the tricuspid valve, AV block, or tetralogy of Fallot (TOF), either alone or in combinations.[16] Haploinsufficiency is thought to be at the root of the malformations. A similar situation exists for the T-box factor TBX5, in which heterozygous mutations cause a variety of CHDs in the context of Holt–Oram syndrome.[17] The linkage to haploinsufficiency is supported by the occurrence of the same syndrome in mouse with one deleted copy of *Tbx5*.[18] The symptom severity of cardiac defects also depends on the type of mutation. Some missense mutations result in a non-functional protein, whereas others may lead to altered properties of unknown nature.[19] Certain mutations abolish binding of Tbx5 to its DNA-binding sites,[20] whereas others influence collaboration with other proteins.[21] For example, Nkx2-5 physically interacts with Tbx5 and Gata4 to synergistically activate downstream target genes.[22,23] Disruption of the stoichiometry of the TF interaction by a decreased amount of either protein may lead to similar effects on transcriptional targets. Intriguingly, mutations in human α-myosin heavy chain (MYH6), a direct target of NKX2-5, TBX5 and GATA4, also cause ASDs.[24] Additionally, the disease manifestation of decreased TF dosage may vary due to stochastic events of unknown nature or parameters comprising environmental influences and genetic modifiers.

This suggests that the regulatory context of TFs plays an important role and their function must be viewed in the context of transcriptional networks including the interplay between different TFs. For example, it has been shown that a decreased level of Tbx20 affects heart development *via* a breakdown of transcription factor networks.[25]

In this study, we analyzed expression levels of a comprehensive set of 46 cardiac genes in heart biopsies derived from healthy individuals and patients with a broad range of cardiac malformations. The selected genes include TFs and potential downstream targets known from literature as well as those identified in our previous microarray analysis.[26] To build the bridge between disease phenotypes and transcriptional networks, first a detailed phenotype ontology of the heart malformations was delineated. Next, the expression levels in normal and malformed hearts were placed within the context of the corresponding phenotype. Application of linear models that analyze gene expression integrating age and gender dependencies revealed transcriptional changes between distinct patient groups. Additionally, independent of corresponding phenotypes, groups of correlated genes were identified based on similar expression patterns of genes both in normal and malformed hearts. Combining these approaches, we were able to find genes that appeared to be specifically associated with certain phenotypes and showed correlated expression in general. Finally, based on correlated gene expression and transcription factor binding site prediction, which was optimized on a heart-specific ChIP data set, we constructed cardiac regulatory networks. As proof of principle, these networks point out novel as well as known regulatory dependencies and moreover explain parts of the observed transcription patterns in diseased cardiac samples.

## Results and discussion

### Phenotype ontology

To enable the selection of a balanced patient population allowing the separation of disease- or tissue-specific expression patterns, we collected 190 human ventricular and atrial cardiac tissues. The clinical characterization comprised 250 features of morphological, hemodynamical and therapeutical information which are stored in our *d*-matrix database for detailed analysis and visualization.[27]

To compress the complex and partially overlapping disease characteristics, we delineated a phenotype ontology. A list of 26 disease parameters in addition to tissue type, gender and age was compiled for each patient, including descriptors like "interatrial septal defect" and "right ventricle dilation" (Fig. 1).

To define groups of patients with similar phenotypes, a complete linkage hierarchical clustering approach using this phenotype ontology was carried out. Patients were assigned to eight meta-phenotypes that represent specific clusters derived from cutting the dendrogram at a certain height as shown in Fig. 1. *E.g.* the cluster *TOF-III* contains patients characterized by interatrial septal defects as well as stenosis and/or dilation of the main pulmonary artery in addition to the classical features of Tetralogy of Fallot (TOF), namely interventricular septal defect, overriding aorta, right ventricular hypertrophy and right ventricular outflow tract stenosis. In total, the ventricle and atrial samples were assigned to seven and four diseased groups in addition to healthy samples, respectively (Fig. 1 and Supplemental Fig. S1‡).

### Preliminary expression data analysis

To characterize the transcription patterns of our patient cohort, a set of 42 genes associated with heart development was selected and expression levels were measured by quantitative real-time PCR. For details regarding selected genes refer to Supplemental Table 1.‡ To normalize samples for different amounts of RNA, four house-keeping genes were measured additionally and normalization factors from the three most consistent house-keeping genes were calculated for each sample according to the method suggested by Vandesompele *et al.*[28] After the normalization process the housekeeping genes were excluded from subsequent analyses.

For an initial overview of the expression data, hierarchical clustering using complete linkage was applied revealing clear differences between atrial and ventricular samples (Fig. 2A). Several of the genes displaying chamber-specific expression have already been described in studies of human and mouse myocardium. *E.g. NPPA*, *NR2F1*, *MYH6*, *MYL7* and *TAGLN* predominate in atria,[29] whereas *Irx4* and *Myl2* are restricted to ventricles.[30] Correspondence analysis[31] supported the tissue-specific differences and demonstrated that diseased and healthy as well as aged and young individuals could be distinguished, implicating that the obtained data is biologically meaningful (Fig. 2B and C). Subsequent analyses were carried out for both cardiac tissues separately, whereof results of the ventricle are illustrated in this manuscript.

This journal is © The Royal Society of Chemistry 2008

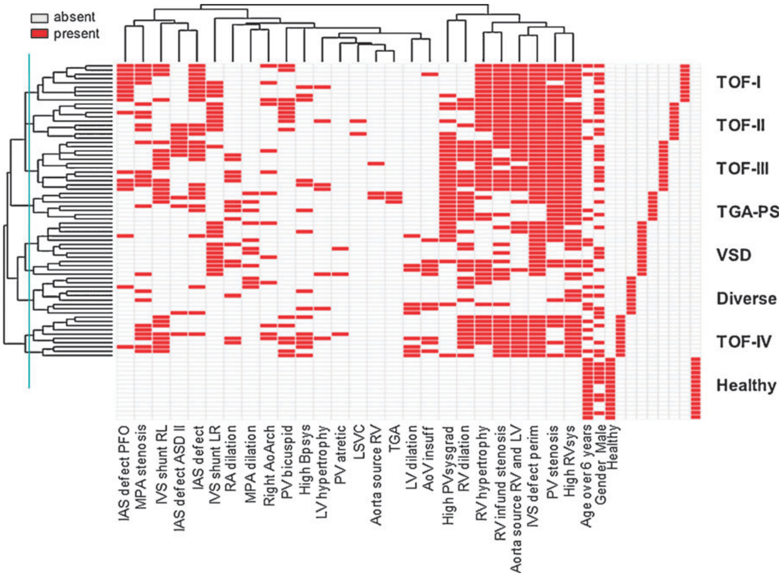


**Fig. 1** Hierarchical clustering of cardiac disease phenotype criteria and assignment of patients with similar characteristics into meta-phenotype groups of ventricular samples. The phenotype information for gender, age and disease state is indicated. Each row represents a single heart sample. The blue line indicates the used cut-off for assignment of meta-phenotypes.

### Linear model to detect differentially expressed genes

To extract the influence of phenotype clusters on gene activity considering known confounding factors such as age and gender,[26] we used linear modeling techniques. We computed the linear model $Y = \alpha_{\text{meta-phenotype}} + \beta_{\text{age}} + \gamma_{\text{gender}}$, where $Y$ is the predicted expression value, $\alpha_{\text{meta-phenotype}}$ is the coefficient for each individual patient group sharing the same meta-phenotype, $\beta_{\text{age}}$ is the coefficient for our two age categories *young* (younger than 6 years) and *old*, and $\gamma_{\text{gender}}$ determines gender specific effects. We did not use an additional intercept term because each individual expression vector was centered beforehand. After estimating each coefficient using a standard linear model, we tested whether it is significantly different from zero and has therefore a significant influence on gene expression. We used a significance level of 0.05 to determine relevant effects. Interestingly, we found deregulated genes for almost all meta-phenotypes, except the cluster *Diverse*, which contains a mixture of different minor phenotypes excluding VSD and with a regular aortic source from the right ventricle (Fig. 3). The other meta-phenotypes, characterized by distinct and moderate to severe abnormalities, have specific molecular portraits, such as *TBX20* and *MEF2C* being upregulated in patients with TOF and main pulmonary artery abnormalities (cluster TOF-III), whereas *TBX5* being only downregulated in patients with TOF and bicuspid pulmonary valve (cluster TOF-II). Some genes appear to be significantly deregulated in all diseased samples, indicated by an opposite regulation in the healthy cluster, *e.g.* *MEF2A* is upregulated in all disease meta-phenotypes. Based on the transcriptional profiles, previously not disease-associated candidate genes could be identified by this approach, like *TBX20* and *DPF3* which have been further investigated (ref. 32 and unpublished data, Lange and Sperling 2008).

### Correlated expression of genes

To finally build transcription networks, we were interested in groups of genes that show a correlated pattern of expression both in normal and diseased samples. To assess correlation between individual gene pairs, we computed their Pearson correlation coefficient over all samples in our dataset. Using random experiments we evaluated the statistical significance of found correlation coefficients. As a null model, we randomly assigned measurements to samples in the according expression vectors without replacement, and computed the correlation coefficients on the randomized expression vectors. This process was repeated 100 000 times and the extent of randomized coefficients exceeding the true coefficient was counted. We thereby derived an empirical *p*-value for the measured correlation coefficient of each individual gene pair. We applied a *p*-value threshold of $1 \times 10^{-3}$ to ensure a high level of significance. For a detailed overview of measured correlations and assigned *p*-values refer to Supplemental Fig. S2.‡ Subsequently, hierarchical clustering using complete linkage was performed only on significant correlation coefficients, while all non-significant coefficients were set to 0. The $1 \times 10^{-3}$ quantile of the overall random distribution was used to split the clustering tree to derive 19 clusters with significant distances between individual genes (Fig. 4A).

We call clusters comprising more than one gene *correlated gene groups* and two examples are shown in Fig. 4B and C. Centered expression vectors were sorted by the defined meta-phenotypes and similar expression patterns of genes can

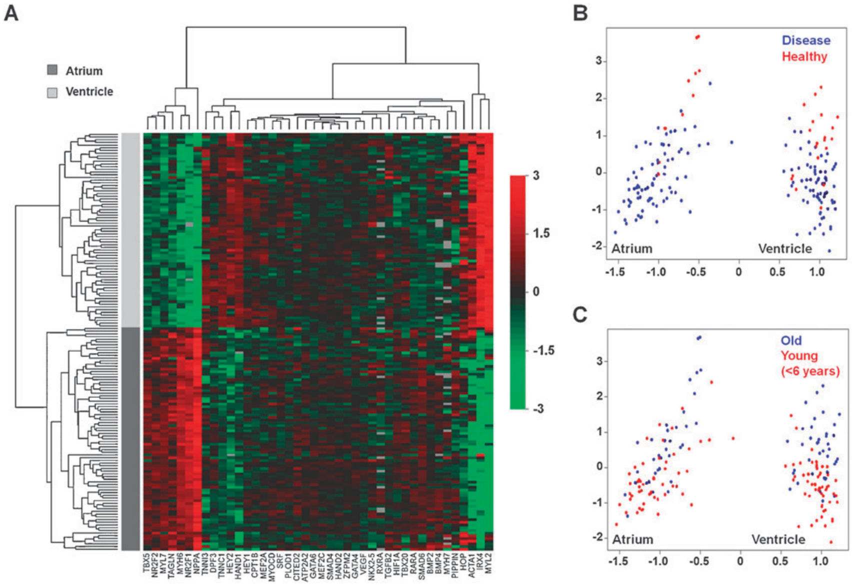


**Fig. 2** Preliminary gene expression data analysis. (A) Hierarchical clustering of gene expression levels measured by real-time PCR in cardiac tissue samples from patients with different heart malformations. Each column represents a gene and each row a single cardiac sample. Normalized and centered expression levels are color coded in red for upregulated and green for downregulated genes. Missing values are depicted in gray. (B + C) Biplot obtained from correspondence analysis. Each dot represents a single patient sample color coded by disease state (B) or age (C).

clearly be seen in normal and diseased tissue samples. The TFs *TBX20* and *MEF2C* displayed correlated expression patterns and strikingly, both are upregulated in patients with TOF-III analyzed with the linear model.

It might occur that two genes show correlated expression over a large set of samples but are strongly deregulated in a specific meta-phenotype *e.g.* due to a breakdown of TF networks. This would lead to a decreased correlation coefficient and loss of cluster assignment. On the other hand, the found gene clusters could be a product of background noise. To consider the robustness of our correlated gene groups, we repeated the correlation analysis successively eliminating one meta-phenotype and taking the maximal correlation coefficient. Random experiments were carried out as described above but comprising the meta-phenotype elimination. Significant maximal correlation coefficients were extracted and hierarchical clustering using complete linkage was performed (Supplemental Fig. S3‡). While the resulting cluster dendrogram shows some changes in cluster association for single genes and subclusters, the majority of clusters stayed intact thereby confirming our found correlated gene groups. For example, the correlated gene group comprising *HAND2*, *MEF2C*, *SMAD4* and *TBX20* was recovered and further enlarged by *DPF3* and *VEGF*, which formerly made up a separate correlated gene group, as well as *HIF1A* that had not been assigned to any correlated group before. Even in the initial correlation analysis which considered all meta-phenotypes, *DPF3* showed significant correlation with all four genes and *HIF1A* and *VEGF* with three and two, respectively (Supplemental Fig. S3‡). Finally, we computed significant maximal correlation coefficients over single meta-clusters only. Remarkably, using such a reduced set of samples, again many of the genes previously assigned to correlated gene groups retained the clustering (data not shown).

Showing strong correlation over the high number of different samples, it is likely that a correlated gene group is co-regulated by the same TF(s). Therefore, we tried to discover TFs that have binding sites in the promoters of all of the genes belonging to one correlated gene group by performing transcription factor binding site (TFBS) predictions. To find the best settings we optimized our prediction using wet lab data generated by us previously.

### Optimization of TFBS prediction using ChIP

To predict possible binding of TFs to the promoter regions of our gene set, we used two different matching algorithms, one proposed by Rahmann *et al.*[33] (Rahmann-Matcher) and the Match algorithm provided by TRANSFAC.[34] Matrices representing known TFBS for TFs in our gene set were retrieved from TRANSFAC.[35]

The length of promoter sequence as well as the use of conservation information taken for TFBS prediction varies among different studies[36–38] and generally, the sequence length considered is positively correlated with an increase of noise.[33] To make our TFBS prediction as biologically meaningful as possible with regard to these settings, we used data obtained from ChIP coupled with array based detection of enriched DNA-fragments in mouse cardiomyocytes for a subset of three TFs, namely Gata4, Mef2a and Nkx2-5 (unpublished data, Fischer and Sperling 2008). We consider this approach to be more applicable compared to arbitrarily chosen settings. To find an optimal balance between length of promoter sequence and noise level in the prediction of assigned binding sites, we

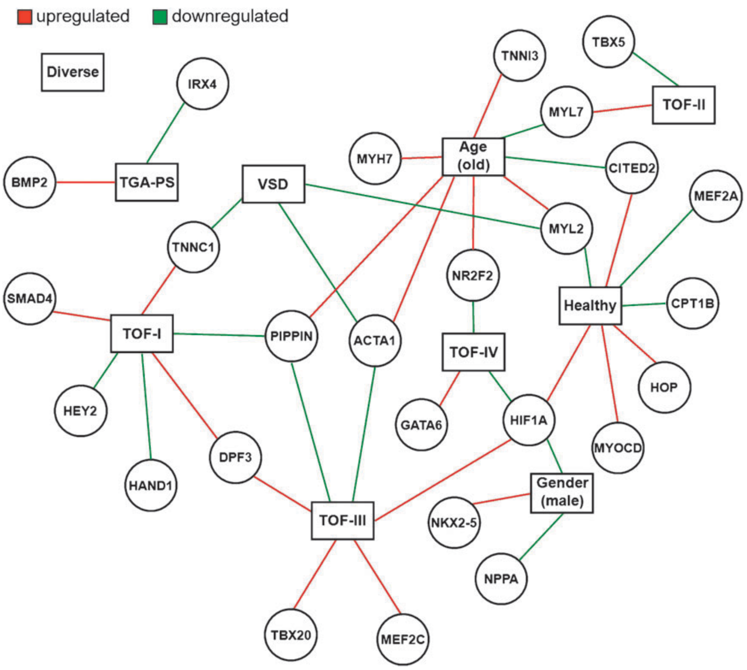


**Fig. 3** Network obtained from a linear model showing significantly deregulated genes in ventricular samples associated with meta-phenotypes as well as age and gender (marked as rectangles). Genes are depicted as circles. Green and red arrows indicate down- and upregulated genes respectively using a significance level of 0.05.

used different upstream and downstream distances as an optimization criterion. Besides the amount of promoter sequence, we used the level of conservation as an optimization parameter. The third parameter optimized was the matching algorithm.

To find an optimal TFBS prediction setting, we used the following scoring function, which was evaluated for each algorithm on each distance and conservation setting.

$$S = A \times B, \text{ where } A = \frac{\textit{true predictions}}{\textit{all predictions}} \text{ and}$$

$$B = \frac{\textit{predicted peaks}}{\textit{all peaks}}$$

The score $S$ comprises two factors ranging from 0 to 1 that measure different aspects of the TFBS predictions. $A$ measures the fraction of true among all predictions and $B$ measures the capability of predicting a ChIP peak. We used the product of both factors as a scoring function to reduce influences of extreme values in only one factor. The optimization process was performed for all three TFs and the average over the three individual scores computed for each setting was reported.

Applied to our scoring scheme, the TRANSFAC Match algorithm in general achieved higher scores than the Rahmann-Matcher (Supplemental Fig. S4‡). We further observed that the fraction of true predictions decreased with the length of sequence used, which is likely due to an increase in noise level. However, TFBSs identified by ChIP can be observed at any distance from the transcription start sites. While the fraction of true predictions could be enhanced by using more stringent conservation settings, the amount of TF ChIP peaks predicted by the two algorithms heavily dropped at higher conservation levels (Fig. 5). This finding is supported by observations that actual binding sites of TFs might be slightly modified during evolution for example to enable adaptation of TF binding affinity.[39,40] Using our scoring function which incorporates both measures, we found a setting of 1250 bp upstream and 500 bp downstream together with a conservation level of 60% to be optimal for our analyzed TFs. Subsequently, we used these settings and the TRANSFAC Match algorithm for our TFBS prediction.

### Regulatory cardiac networks

Finally, we constructed regulatory networks based on the identified *correlated gene groups* and the predicted TFBSs representing the underlying potential regulatory dependencies. For verification we compared the constructed networks with binding data derived by ChIP and known from literature (Fig. 6). Given that the overlap of literature and ChIP results is not complete it must be kept in mind that ChIP was performed in mouse cardiomyocytes and the literature describing TF binding is based on a variety of experimental setups.

Fig. 7 displays two graphs representing predicted regulatory subnetworks for the *HAND2*, *MEF2C*, *SMAD4*, *TBX20* and *GATA4*, *NR2F1*, *NR2F2*, *TAGLN* correlated gene groups (Fig. 7A and B). For the first correlated gene group, GATA4 and NKX2-5—known to interact with each other—were predicted to bind all four promoters. Comparing these predictions to the network in Fig. 6, all except the two bindings to

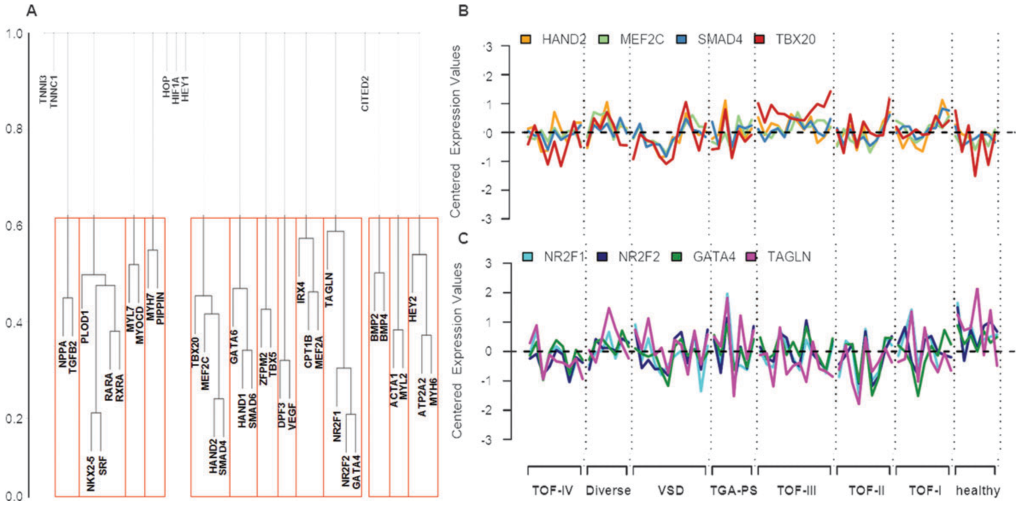


**Fig. 4** Correlation of gene expression. (A) Cluster dendrogram showing 13 correlated gene groups. Clustering was derived by cutting the cluster tree at the $1 \times 10^{-3}$ quantile of a random distribution. The *Y*-axis indicates cluster distances. (B + C) Example of two correlated gene groups showing highly correlated patterns of expression in samples of healthy individuals and patients. Centered expression vectors were sorted by defined meta-phenotypes.

SMAD4 have been proposed in literature (Nkx2-5 → *Mef2C*[41,42]), found in our ChIP data (Nkx2-5 → *Hand2*/*Tbx20*, Gata4 → *Tbx20*) or both (Gata4 → *Hand2*[43]/*Mef2C*[44]). Interestingly, both *TBX20* and *MEF2C* are specifically upregulated in patients within the TOF-III cluster and our approach sheds light on potential upstream regulators. The regulation of *TBX20* is not well known so far. The only described signaling molecule upstream of *Tbx20* is Bmp2,[45] and recently we could show that TFAP2C is a direct regulator.[32] Identification of NKX2-5 and GATA4 as common regulators reveals them as interesting candidate genes to be responsible for the transcription pattern of the phenotype cluster. A causative connection is suggestive and mutations in both TFs have already been linked to TOF.[16,46] Measuring *Tbx20* levels in siRNA knockdown experiments of the respective TFs showed reduction of *Tbx20* mRNA levels by 20–50% (data not shown). These results demonstrate that binding of Nkx2-5 and Gata4 is indeed functional and activates *Tbx20*

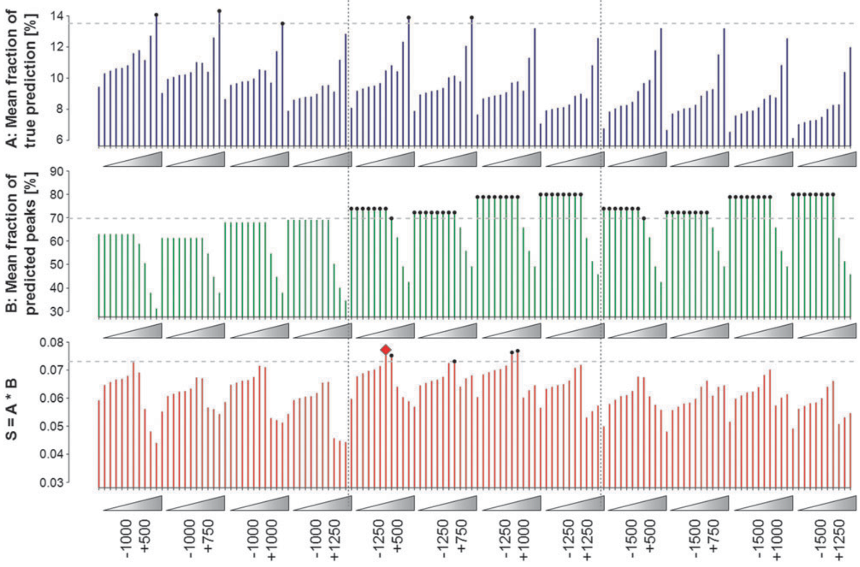


**Fig. 5** Optimization of TFBS prediction. Results are shown for the TRANSFAC Matcher and a subset of promoter settings. The upstream (−) and downstream (+) lengths used as promoter are placed below the plot. Triangles indicate the level of conservation from 0% to 100%. Dashed horizontal lines mark best 5 scores, values above this score are highlighted with black dots. The red diamond highlights the best scoring prediction setting.

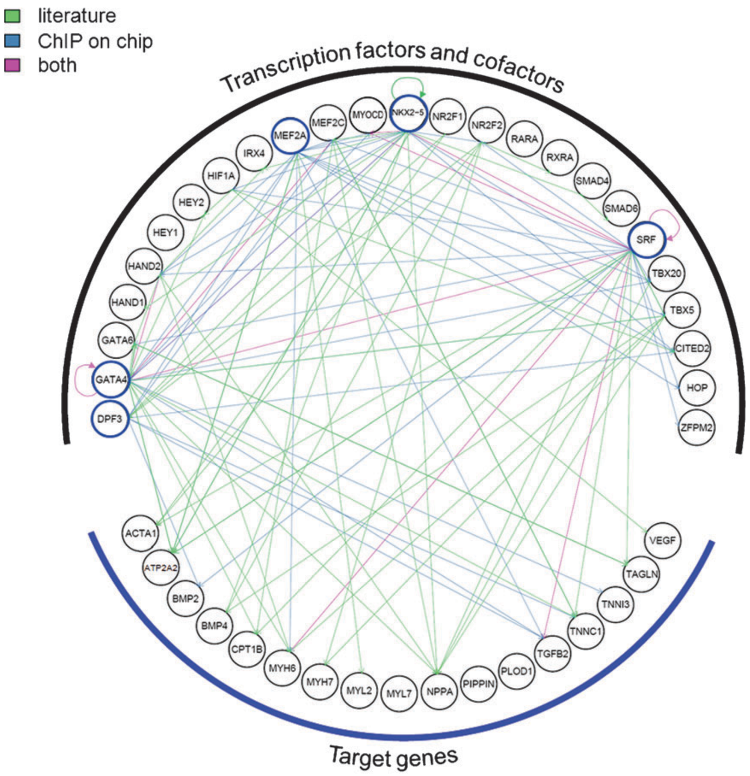


**Fig. 6** Regulatory network based on TF binding information known from literature (green) and ChIP (blue). Red arrows indicate regulatory interactions found in both. TFs encircled in blue were investigated by ChIP.

expression. Potentially, posttranscriptional modifications could explain why neither of the two TFs are part of this correlated gene group.[47] Concerning the *GATA4*, *NR2F1*, *NR2F2* and *TAGLN* correlated gene group, several TFs were found that had predicted binding sites in all promoters of the four genes. Among them are TBX5, GATA6 as well as GATA4 and the two NR2F factors. Identification of the latter three TFs is quite remarkable as all the TFs present in this correlated gene group appear to show regulatory interactions with each other that could explain the found correlation. As seen in Fig. 6, some connections have already been described in literature (Gata6 → *Gata4*[48]/*Tagln*[49]; Gata4 → *Gata4*[48]) but no binding was found in our ChIP data. However, it must be kept in mind that the ChIP experiments were performed using mouse cardiomyocytes, whereas the predictions are based on transcription patterns from human patient material.

In order to substantiate the predicted TF regulations, we finally incorporated the transcription factor affinity prediction (TRAP)[50] algorithm as a new method. TRAP is based on a physical binding model which aims to predict TF affinities to a given promoter sequence similar to ChIP experiments. The provided affinity measure is continuous and allows easy ranking of promoters with the highest affinity for each TF matrix. As an advantage over classical TFBS prediction methods, TRAP also incorporates contributions from weak binding sites and might therefore be a more sensitive measure to predict regulations. For each TF we computed its top-10 affinity table comprising the promoters with the highest affinities.

Applying TRAP to the correlated gene group comprising *HAND2*, *MEF2C*, *SMAD4* and *TBX20*, we did not find any TF which had high affinity for all four gene promoters. Remarkably, SMAD4 could not be found in any of the top-10 affinity tables computed for all TFs in our data set, although the SMAD4 promoter was predicted to be bound by a large fraction of TFs (Fig. 7). Regarding the results of the TFBS prediction, NKX2-5 was assigned by TRAP to two of the remaining three genes, namely *MEF2C* and *HAND2* (confirmed by literature and ChIP, respectively), but did not show high affinity for *TBX20*. However, binding of Nkx2-5 to *Tbx20* was observed in ChIP. Therefore, we believe NKX2-5 to be a crucial factor for the stated correlation.

In the case of the *GATA4*, *NR2F1*, *NR2F2* and *TAGLN* correlated gene group, both GATA4 and GATA6 appeared to have all four gene promoters in their TOP-10 affinity tables. This underlines the results of the TFBS prediction in which they also showed binding to all group members. Furthermore, it highlights these GATA proteins as potential auto-regulatory key factors in the given subnetwork. In addition, SMAD6 showed high affinity to three of the four correlated genes,

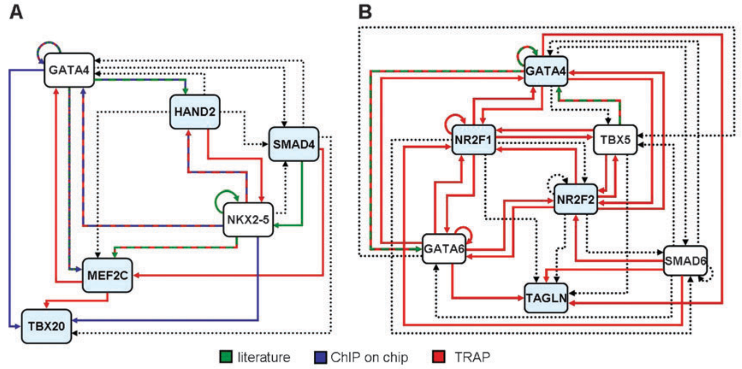


**Fig. 7** Predicted regulatory subnetworks for two correlated gene groups (A + B). Genes composing a group are marked light blue. Confirmation of predicted binding by literature, ChIP on chip and/or TRAP is depicted in colors. Unconfirmed predictions are indicated by dashed lines.

namely *NR2F1*, *NR2F2* and *TAGLN* and was predicted to be bound by GATA4 itself, which implies a functional role further downstream in the regulatory cascade.

In summary, using TRAP we could partially confirm results of the TFBS prediction and extract possible key regulators. However, as shown in case of NKX2-5, the highest affinity prediction does not always reflect biological binding known from literature or identified by ChIP. We believe that a combination of different approaches as done in this study will lead to more significant results in the light of biological authenticity.

## Conclusion

A variety of methods to predict transcription networks has been proposed recently, however, integrative approaches combining complex clinical phenotype data with advanced bioinformatic and biochemical methods are still lacking. Here, we present the first cardiac transcription networks based on predicted transcription factor binding sites and gene expression profile disturbances in samples of congenital malformed hearts. The idea to use this complex phenotype was driven by the assumption that a broad panel of cardiac phenotypes associated with a range of genomic sequence variations and different modifiers, potentially underlying the phenotype, would lead to ranges of expression patterns rather than distinct profiles. This should enable the identification of transcriptional dependencies. We combine several methods such as linear models, correlation analyses based on expression profiles as well as the prediction of *cis* regulatory elements to predict the transcription networks. Furthermore, we verify our obtained networks using data derived by literature and ChIP.

However, one has to bear in mind that expression profiling detects only transcript abundance and not the activity of the encoded proteins. Posttranslational modifications, such as phosphorylation of MEF2 proteins, NKX2-5 and GATA4,[51–54] allow fine-tuning of gene activity independent of expression levels and add an additional layer of complexity to the network of transcription factors operating in the developing heart. In addition, several evolutionary conserved micro-RNAs function as regulators of target RNAs, one of which is miR-1 that negatively regulates cardiac growth during mouse development by inhibiting translation of *Hand2*.[55]

## Material and methods

### Patient samples

All cardiac samples were obtained from the German Heart Center during cardiac surgery after short-term cardioplegia, with ethical approval by the institutional review committee and informed consent of the patients or their parents. Biopsies were taken from the right ventricle and atrium of patients with different cardiac malformations as well as from normal human hearts. All samples were directly snap-frozen in liquid nitrogen after excision and stored at −80 °C. Clinical characteristics of the study subjects are shown in Fig. 1 and Supplemental Fig. S1.‡

### RNA isolation and quantitative real-time PCR

Total RNA of all cardiac tissues were extracted using TRIzol (Invitrogen, Germany) according to manufacturer's instructions. 5 μg of total RNA were reverse transcribed using AMV-reverse transcriptase (Promega) and random hexamer primers (Amersham Pharmacia Biotech). Real-time PCR was carried out using SYBR Green PCR master mix (ABgene) on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). Intron spanning primers for 46 genes were designed using the Primer Express software (Applied Biosystems) and are available upon request. Expression levels were normalized using a normalization factor calculated from the three out of four most consistent house-keeping genes. In this setting B2M, HPRT and ABL were included according to the GeNorm software as described previously.[28] Before any further analysis, gene expression vectors were centered for comparability.

### Data analysis

If not mentioned otherwise, all bioinformatics analyses were carried out using R and Bioconductor packages[56] as well as Perl and its BioPerl modules. A total of 39 matrices associated with 15 of 22 TFs within the heart data set were retrieved from TRANSFAC[35] (version 11.3). Applying a pre-filtering step, we eliminated low quality matrices to reduce the number of false positive predictions. We used the predefined matrix similarity thresholds applied for matching by the TRANSFAC Match algorithm. By excluding matrices with a predefined matrix similarity score of less than 0.8 we reduced the number of matrices to 27 assigned to 15 TFs. In a post-filtering step, we removed again two matrices showing a very high number of average predictions per promoter. In total this led to 25 matrices associated with 15 TFs as shown in Supplemental Table 2.‡ Finally, predictions from matrices belonging to one TF were combined in order to build the basis for the construction of regulatory networks.

For the TRANSFAC Match algorithm the ''minimize the sum of both error rates'' options were used that set predefined cut-offs for matrix and core similarity.[34] TFBS prediction using the Rahmann-Matcher was carried out with a balanced type I and type II error and a *p*-value cutoff of 0.05.[33] The TRAP algorithm was used with the standard settings on all promoters of our dataset with the same settings as optimized before and overall promoter affinity was extracted for later analysis.[50] After deriving an affinity score for every individual promoter and TRANSFAC matrix in our data set, we extracted promoters with the ten highest affinities for each TRANSFAC matrix. Next, we combined matrices to TFs as in the TFBS prediction analysis thereby deriving TOP-10 affinity tables.

Based on transcription start sites in Ensembl (version 48), we used 10 kb upstream and 3 kb downstream of the 42 selected genes as promoter regions. Upstream distances gradually increasing from 200 bp to 10 kbp and downstream distances from 100 bp to 3 kbp were considered. To assess conservation of promoter sequences, the full mouse human BlastZ alignment was downloaded from Ensembl (human assembly NCBI 36; mouse assembly NCBI m37). In addition to the single nucleotide conservation masking provided by the alignment, a 100 bp window was shifted along the promoters and windows exceeding a given percentage of conservation remained unmasked. Thresholds ranging from 0% to 100% conservation were evaluated in continuous steps of 10%. Repetitive and transcribed regions were not masked. For computation of the defined score *S*, we marked a prediction as true if it was located in a range of 250 bp apart from a respective middle of a ChIP peak. Furthermore, peaks were marked as predicted if they had at least one true prediction assigned. Predictions as well as peaks were evaluated with respect to the tested promoter settings and peaks lying outside of the evaluated promoter regions were excluded.

## Abbreviations

In general mouse gene symbols are italicized, first letter upper case all the rest lower case, while human genes are indicated by all letters being in upper case. AoArch, aortic arch; ASDII, atrial septal defect of secundum type; Bpsys, systolic blood pressure; ChIP, chromatin immunoprecipitation; IAS, interatrial septal defect; Infund, infundibular; Insuff, insufficiency; IVS, Intreventricular septum; LA, RA, left/right atrium; LR, left to right; LSVC, left superior caval vein present; LV, RV, left/right ventricle; MPA, main pulmonary artery; Perim, perimembranous; PFO, patent foramen ovale; PV, pulmonary valve; RL, right to left; siRNA, small interfering RNA; Sysgrad, systolic gradient; TF, transcription factor; TFBS, transcription factor binding site; TGA, transposition of the great arteries.

## Acknowledgements

We are deeply grateful to the patients and families for their cooperation. We gratefully acknowledge Dr Siegrun Mebus for preparation of patient samples, Ilona Dunkel for technical assistance and Tammo Krueger for bioinformatic help. We thank Dr Szymon Kielbasa, Dr Thomas Manke and Helge Roider for help with the Rahmann Matcher and TRAP. This work has been supported by the European Community's Sixth Framework Programme contract ('HeartRepair') LSHM-CT-2005-018630. Martje Toenjes is supported by a fellowship of the Studienstiftung des Deutschen Volkes.

## References

1 G. Chua, M. D. Robinson, Q. Morris and T. R. Hughes, *Curr. Opin. Microbiol.*, 2004, **7**, 638–646.
2 D. J. Reiss, N. S. Baliga and R. Bonneau, *BMC Bioinf.*, 2006, **7**, 280.
3 J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink and E. D. Jarvis, *Bioinformatics*, 2004, **20**, 3594–3603.
4 M. Bansal, G. D. Gatta and D. di Bernardo, *Bioinformatics*, 2006, **22**, 815–822.
5 E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, *Nat. Genet.*, 2003, **34**, 166–176.
6 A. D. Smith, P. Sumazin, Z. Xuan and M. Q. Zhang, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 6275–6280.
7 Z. Hu, P. J. Killion and V. R. Iyer, *Nat. Genet.*, 2007, **39**, 683–687.
8 M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo, *Mol. Syst. Biol.*, 2007, **3**, 78.
9 K. L. Clark, K. E. Yutzey and D. W. Benson, *Annu. Rev. Physiol.*, 2006, **68**, 97–121.
10 J. I. Hoffman, *Pediatr. Cardiol.*, 1995, **16**, 103–113.
11 J. I. Hoffman, *Pediatr. Cardiol.*, 1995, **16**, 155–165.
12 S. Sperling, C. H. Grimm, I. Dunkel, S. Mebus, H. P. Sperling, A. Ebner, R. Galli, H. Lehrach, C. Fusch, F. Berger and S. Hammer, *Hum. Mutat.*, 2005, **26**, 575–582.
13 V. Garg, I. S. Kathiriya, R. Barnes, M. K. Schluterman, I. N. King, C. A. Butler, C. R. Rothrock, R. S. Eapen, K. Hirayama-Yamada, K. Joo, R. Matsuoka, J. C. Cohen and D. Srivastava, *Nature*, 2003, **424**, 443–447.
14 J. J. Schott, D. W. Benson, C. T. Basson, W. Pease, G. M. Silberbach, J. P. Moak, B. J. Maron, C. E. Seidman and J. G. Seidman, *Science*, 1998, **281**, 108–111.
15 S. M. Ware, J. Peng, L. Zhu, S. Fernbach, S. Colicos, B. Casey, J. Towbin and J. W. Belmont, *Am. J. Hum. Genet.*, 2004, **74**, 93–105.
16 D. W. Benson, G. M. Silberbach, A. Kavanaugh-McHugh, C. Cottrill, Y. Zhang, S. Riggs, O. Smalls, M. C. Johnson, M. S. Watson, J. G. Seidman, C. E. Seidman, J. Plowden and J. D. Kugler, *J. Clin. Invest.*, 1999, **104**, 1567–1573.
17 C. T. Basson, D. R. Bachinsky, R. C. Lin, T. Levi, J. A. Elkins, J. Soults, D. Grayzel, E. Kroumpouzou, T. A. Traill, J. Leblanc-Straceski, B. Renault, R. Kucherlapati, J. G. Seidman and C. E. Seidman, *Nat. Genet.*, 1997, **15**, 30–35.
18 B. G. Bruneau, G. Nemer, J. P. Schmitt, F. Charron, L. Robitaille, S. Caron, D. A. Conner, M. Gessler, M. Nemer, C. E. Seidman and J. G. Seidman, *Cell*, 2001, **106**, 709–721.
19 S. J. Cross, Y. H. Ching, Q. Y. Li, L. Armstrong-Buisseret, S. Spranger, S. Lyonnet, D. Bonnet, M. Penttinen, P. Jonveaux, B. Leheup, G. Mortier, C. Van Ravenswaaij and C. A. Gardiner, *J. Med. Genet.*, 2000, **37**, 785–787.
20 T. K. Ghosh, E. A. Packham, A. J. Bonser, T. E. Robinson, S. J. Cross and J. D. Brook, *Hum. Mol. Genet.*, 2001, **10**, 1983–1994.

21 Y. Hiroi, S. Kudoh, K. Monzen, Y. Ikeda, Y. Yazaki, R. Nagai and I. Komuro, *Nat. Genet.*, 2001, **28**, 276–280.

22 E. M. Small and P. A. Krieg, *Dev. Biol.*, 2003, **261**, 116–131.

23 J. K. Takeuchi, M. Ohgi, K. Koshiba-Takeuchi, H. Shiratori, I. Sakaki, K. Ogura, Y. Saijoh and T. Ogura, *Development*, 2003, **130**, 5953–5964.

24 Y. H. Ching, T. K. Ghosh, S. J. Cross, E. A. Packham, L. Honeyman, S. Loughna, T. E. Robinson, A. M. Dearlove, G. Ribas, A. J. Bonser, N. R. Thomas, A. J. Scotter, L. S. Caves, G. P. Tyrrell, R. A. Newbury-Ecob, A. Munnich, D. Bonnet and J. D. Brook, *Nat. Genet.*, 2005, **37**, 423–428.

25 J. K. Takeuchi, M. Mileikovskaia, K. Koshiba-Takeuchi, A. B. Heidt, A. D. Mori, E. P. Arruda, M. Gertsenstein, R. Georges, L. Davidson, R. Mo, C. C. Hui, R. M. Henkelman, M. Nemer, B. L. Black, A. Nagy and B. G. Bruneau, *Development*, 2005, **132**, 2463–2474.

26 B. Kaynak, A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, J. Vogel, H. P. Sperling, R. Pregla, V. Alexi-Meskishvili, R. Hetzer, P. E. Lange, M. Vingron, H. Lehrach and S. Sperling, *Circulation*, 2003, **107**, 2467–2474.

27 D. Seelow, R. Galli, S. Mebus, H. P. Sperling, H. Lehrach and S. Sperling, *BMC Bioinf.*, 2004, **5**, 168.

28 J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe and F. Speleman, *Genome Biol.*, 2002, **3**, RESEARCH0034.

29 P. Ellinghaus, R. J. Scheubel, D. Dobrev, U. Ravens, J. Holtz, J. Huetter, U. Nielsch and H. Morawietz, *J. Thorac. Cardiovasc. Surg.*, 2005, **129**, 1383–1390.

30 R. Tabibiazar, R. A. Wagner, A. Liao and T. Quertermous, *Circ. Res.*, 2003, **93**, 1193–1201.

31 K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel and M. Vingron, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10781–10786.

32 S. Hammer, M. Toenjes, M. Lange, J. J. Fischer, I. Dunkel, S. Mebus, C. H. Grimm, R. Hetzer, F. Berger and S. Sperling, *J. Cell. Biochem.*, 2008, DOI: 10.1002/jcb.21686.

33 S. Rahmann, T. Muller and M. Vingron, *Stat. Appl. Genet. Mol. Biol.*, 2003, **2**, art. 7.

34 A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis and E. Wingender, *Nucleic Acids Res.*, 2003, **31**, 3576–3579.

35 V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender, *Nucleic Acids Res.*, 2003, **31**, 374–378.

36 L. A. Goff, J. Davila, R. Jornsten, S. Keles and R. P. Hart, *J. Biomol. Tech.*, 2007, **18**, 205–212.

37 S. Y. Kim and Y. Kim, *BMC Bioinf.*, 2006, **7**, 330.

38 S. Nelander, E. Larsson, E. Kristiansson, R. Mansson, O. Nerman, M. Sigvardsson, P. Mostad and P. Lindahl, *BMC Genomics*, 2005, **6**, 68.

39 U. Gerland, J. D. Moroz and T. Hwa, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12015–12020.

40 R. R. Copley, M. Totrov, J. Linnell, S. Field, J. Ragoussis and I. A. Udalova, *Genome Res.*, 2007, **17**, 1327–1335.

41 I. S. Skerjanc, H. Petropoulos, A. G. Ridgeway and S. Wilton, *J. Biol. Chem.*, 1998, **273**, 34904–34910.

42 M. Tanaka, Z. Chen, S. Bartunkova, N. Yamasaki and S. Izumo, *Development*, 1999, **126**, 1269–1280.

43 D. G. McFadden, J. Charite, J. A. Richardson, D. Srivastava, A. B. Firulli and E. N. Olson, *Development*, 2000, **127**, 5331–5341.

44 E. Dodou, M. P. Verzi, J. P. Anderson, S. M. Xu and B. L. Black, *Development*, 2004, **131**, 3931–3942.

45 T. F. Plageman, Jr and K. E. Yutzey, *J. Biol. Chem.*, 2004, **279**, 19026–19034.

46 G. Nemer, F. Fadlalah, J. Usta, M. Nemer, G. Dbaibo, M. Obeid and F. Bitar, *Hum. Mutat.*, 2006, **27**, 293–294.

47 M. J. Herrgard, M. W. Covert and B. O. Palsson, *Genome Res.*, 2003, **13**, 2423–2434.

48 A. Rojas, S. De Val, A. B. Heidt, S. M. Xu, J. Bristow and B. L. Black, *Development*, 2005, **132**, 3405–3417.

49 W. Nishida, M. Nakamura, S. Mori, M. Takahashi, Y. Ohkawa, S. Tadokoro, K. Yoshida, K. Hiwada, K. i. Hayashi and K. Sobue, A Triad of Serum Response Factor and the GATA, NK Families Governs the Transcription of Smooth and Cardiac Muscle, *Genes*, 2002, **277**, 7308–7317.

50 H. G. Roider, A. Kanhere, T. Manke and M. Vingron, *Bioinformatics*, 2007, **23**, 134–141.

51 J. D. Molkentin, L. Li and E. N. Olson, *J. Biol. Chem.*, 1996, **271**, 17199–17204.

52 O. I. Ornatsky, D. M. Cox, P. Tangirala, J. J. Andreucci, Z. A. Quinn, J. L. Wrana, R. Prywes, Y. T. Yu and J. C. McDermott, *Nucleic Acids Res.*, 1999, **27**, 2646–2654.

53 H. Kasahara and S. Izumo, *Mol. Cell. Biol.*, 1999, **19**, 526–536.

54 F. Charron, G. Tsimiklis, M. Arcand, L. Robitaille, Q. Liang, J. D. Molkentin, S. Meloche and M. Nemer, *Genes Dev.*, 2001, **15**, 2702–2719.

55 Y. Zhao, E. Samal and D. Srivastava, *Nature*, 2005, **436**, 214–220.

56 R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang, *Genome Biol.*, 2004, **5**, R80.

This journal is © The Royal Society of Chemistry 2008

### 2.1.5 Genomic organization of the transcriptome

Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W, **Sperling S**. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics* 2007;**89**:580-587.

Vogel J, von Heydebreck A, Purmann A, **Sperling S**. Chromosomal Clustering of a human transcriptome reveals regulatory background. *BMC Bioinformatics* 2005;**6**:230.

To understand the global regulatory network underlying specific transcriptomes, also the genomic organization of the transcripts (Hershberg et al., 2005) has to be considered. Analyzing the cardiac transcriptome identified by the genome-wide expression analysis of malformed human hearts (Kaynak et al., 2003), the FANTOM data (Carninci et al., 2005) and GNF Symatlas (Su et al., 2004), a striking evidence for a relationship between gene expression and genomic localization was observed. Adjacent genes were defined as gene cluster depending on their expression profile. The study showed that highly coexpressed gene clusters are phylogenetically conserved, have a length limit and mainly consist of non-paralogous genes. **Figure 9** shows the genomic organization of cardiac expressed gene cluster. These clusters show a weaker functional but similar regulatory relationship to each other than general genomic neighbours. This points to so far unknown *cis*-acting units and reject cofunctionality as a driving force. It can be hypothesized that highly coexpressed gene clusters are essential for a higher order of transcriptional regulation, while specific TFs are likely to handle the fine-tuning of transcription on shorter time scales.

Thus, coordinated expression of genomic neighbours might potentially explain the association of 30% of congenital heart disease with additional developmental abnormalities, which are based on micro-deletions/ duplications of 1-3 genes in addition to recognized chromosomal aberrations.

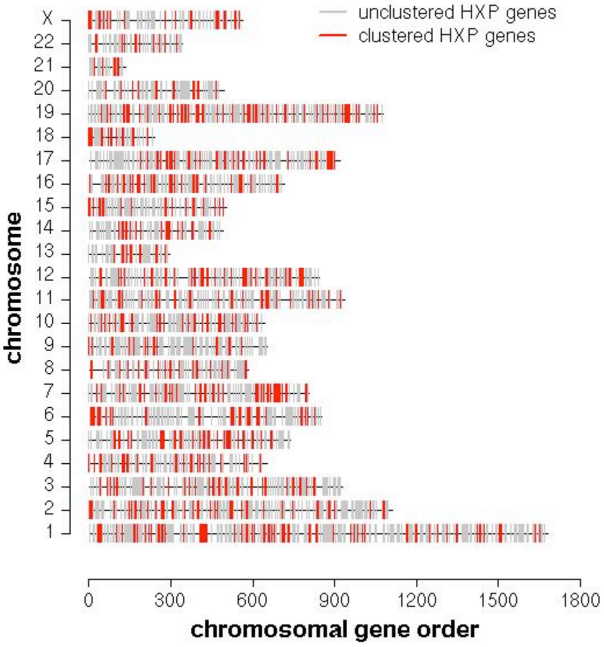


**Figure 9.** Genomic organization of the overall (gray) and clustered heart-expressed genes (HXP, red).

Available online at www.sciencedirect.com
ScienceDirect

GENOMICS

Genomics 89 (2007) 580 – 587

www.elsevier.com/locate/ygeno

# Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality

Antje Purmann [a], Joern Toedling [b], Markus Schueler [a], Piero Carninci [c,d], Hans Lehrach [a], Yoshihide Hayashizaki [c,d], Wolfgang Huber [b], Silke Sperling [a,*]

[a] *Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany*
[b] *EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
[c] *Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan*
[d] *Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan*

Received 29 August 2006; accepted 24 January 2007
Available online 21 March 2007

## Abstract

In studies of their transcriptional activity, genomes have shown a high order of organization. We assessed the question of how genomically neighboring genes are transcriptionally coupled across tissues and what could be the driving force behind their coupling. We focused our analysis on the transcriptome information for 13 tissues of *Mus musculus* and 79 tissues of *Homo sapiens*. The analysis of coexpression patterns of genomically adjacent genes across tissues revealed 2619 and 1275 clusters of highly coexpressed genes, respectively. Most of these clusters consist of pairs and triplets of genes. They span a limited genomic length and are phylogenetically conserved between human and mouse. These clusters consist mainly of nonparalogous genes and show a decreased functional and similar regulatory relationship to one another compared to general genomic neighbors. We hypothesize that these clusters trace back to large-scale, qualitative, persistent reorganizations of the transcriptome, while transcription factor regulation is likely to handle fine-tuning of transcription on shorter time scales. Our data point to so far uncharacterized *cis*-acting units and reject cofunctionality as a driving force.
© 2007 Elsevier Inc. All rights reserved.

*Keywords:* Genomics; Gene expression; Transcription regulation; Regulatory elements; Mammals

Studies of genomes and transcriptomes have shown that genes are nonrandomly located in genomes and that genes of coordinated expression appear in clusters along the genome. This raises the question of how genomes have evolved and how they function. It is evident that the location of a gene in a genome affects its expression, for example, transgene activity can depend on the chromosomal integration site, or an intact gene in a different genomic location can have a pathological phenotype. Clusters of genes that are coexpressed were first identified on a genomic scale in *Saccharomyces cerevisiae* [1,2] and *Caenorhabditis elegans* [3,4]. In the latter, clusters could be attributed to the cotranscription of these genes in operons, a process that is unusual among eukaryotes. However, there is extensive evidence for clusters of coexpressed genes across all major eukaryotes. Using a less stringent definition of a cluster that allows for intervening genes with different expression patterns led to the identification of large groups of coexpressed genes in *Drosophila melanogaster* that span 10-30 genes or, on average, 125 kb of genomic DNA [5]. In *Homo sapiens*, genes with high expression levels tend to cluster in large domains [6,7]. Other reports indicate that genes coexpressed in a given tissue or cell state are clustered along the genome [8–14]. Further, it is well known that specific gene clusters, like for β-globin or HoxD genes, are regulated by a locus and global control region, respectively [15,16]. Here, the latter includes the control of several genes unrelated in structure and function and it is currently unclear whether this is an exceptional or a common feature for higher eukaryotes. Thus, the question arises

---

[*] Corresponding author. Fax: +49 30 8413 1699.
*E-mail address:* sperling@molgen.mpg.de (S. Sperling).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.01.010

if there exist chromosomal transcriptional "hot spots" in a given tissue or if coexpressed clustered genes are mainly housekeeping genes that are expressed in many tissues [17,18]. Some reports indicate that clusters of coexpressed genes tend to be conserved through evolution, for example, coexpressed genes contain fewer breakpoints between human and mouse, indicating that they are held together by natural selection [19–21].

To date, however, we lack an understanding of how many such clusters may exist in mammals, how the transcriptional coupling of gene clusters is regulated in general, and what may be the driving force behind their formation. To address these issues, we assessed the conservation of gene clusters across tissues and species and investigated the functional and regulatory relationship between coexpressed clustered genes. We focused our analysis on two recent transcriptome datasets, namely the FANTOM3 data for 13 tissues of *Mus musculus*, obtained by cap-analysis gene expression (CAGE), and the GNF Symatlas data for 79 tissues of *H. sapiens*, obtained by microarray expression profiling [18,22].

## Results

### *Chromosomal clustering of transcriptomes*

We investigated the genomic organization of 13 *M. musculus* transcriptomes that had been extensively analyzed within the FANTOM3 project [22]. In particular, we were interested in the physical scale of coexpression of genes located adjacent to each other in the genome. We called a set of adjacent genes that are expressed in a particular tissue a cluster of coexpressed genes, independent of their expression levels. A cluster of two neighboring genes expressed in the same tissue is also called a pair, a cluster of three a triplet, and so on.

A large proportion of genes (approx 30-75%) were arranged in gene clusters along the genome without any prevalence for particular chromosomes. These clusters consisted mainly of pairs and triplets. Fig. 1A shows an example of the size distribution of the gene clusters identified in cerebellum, heart, macrophage, and muscle. To evaluate the significance of our observation, we compared the observed number of genes localized in clusters with permuted data, in which the ordering of genes on the DNA, but not their individual expression profiles across tissues, was permuted. This permutation scheme corresponds to a null hypothesis in which the coexpression of genes is independent of their genomic location, but follows the empirical correlations between tissues (Fig. 1B). Even though the number of clusters expected under the null hypotheses is high, the number of observed clusters is significantly larger.

### *Conservation of gene clusters across several tissues*

To quantify coexpression of a pair of genes in a set of $n$ tissues, we defined two coefficients. $\alpha$ is the proportion of tissues in which both genes are expressed, and $\Omega$ is the number of tissues in which either one or both genes are expressed divided by $n$. Both coefficients are numbers between 0 and 1, and $\alpha \leq \Omega$. If $\alpha = \Omega$, the two genes have an identical expression pattern across tissues, while a small ratio of $\alpha/\Omega$ indicates that their expression is not correlated. We computed these coefficients for each chromosomally neighboring pair of genes in the FANTOM3 data (Fig. 2A).

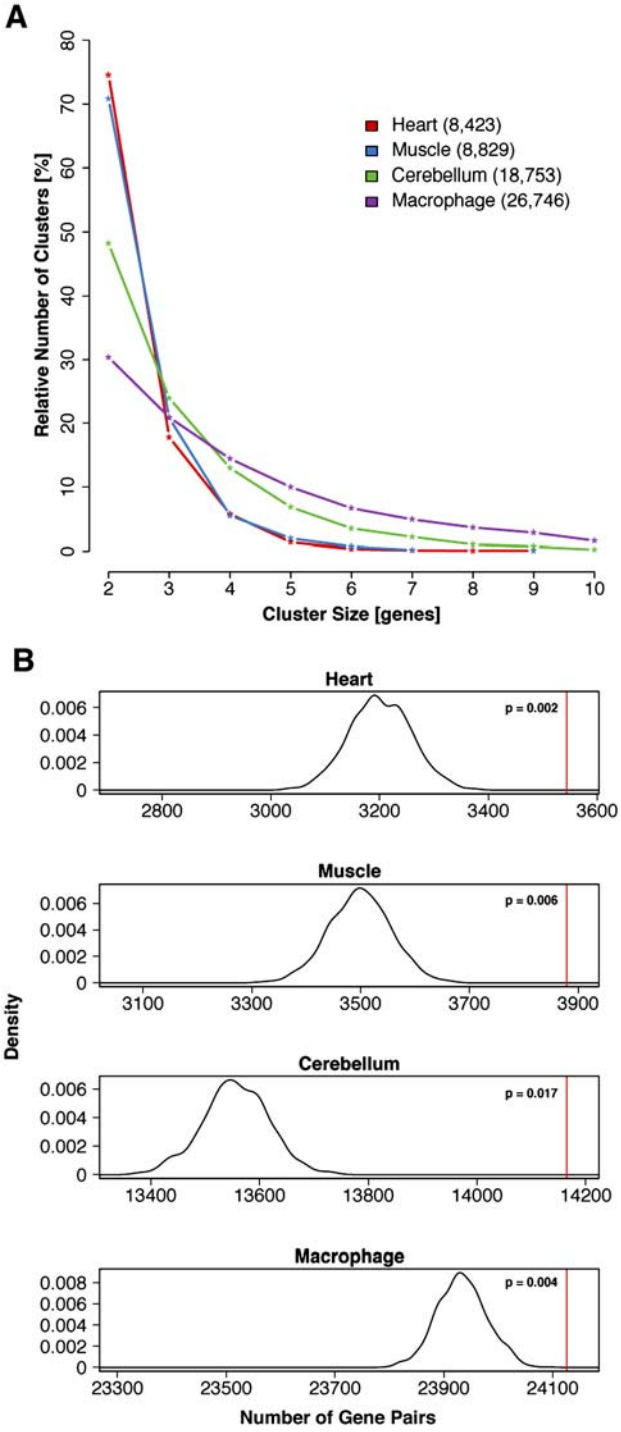


Fig. 1. Transcriptomes are organized in clusters. (A) Relative frequencies of observed gene clusters in the mouse transcriptome, for cerebellum, heart, macrophage, and muscle as examples. The total number of genes observed to be expressed in each tissue is given in parentheses. (B) To assess the statistical significance of the spatial clustering, we compared the number of observed clustered genes with the number of clusters from permuted data. Here, data are presented, as examples, for mouse cerebellum, heart, macrophage, and muscle.

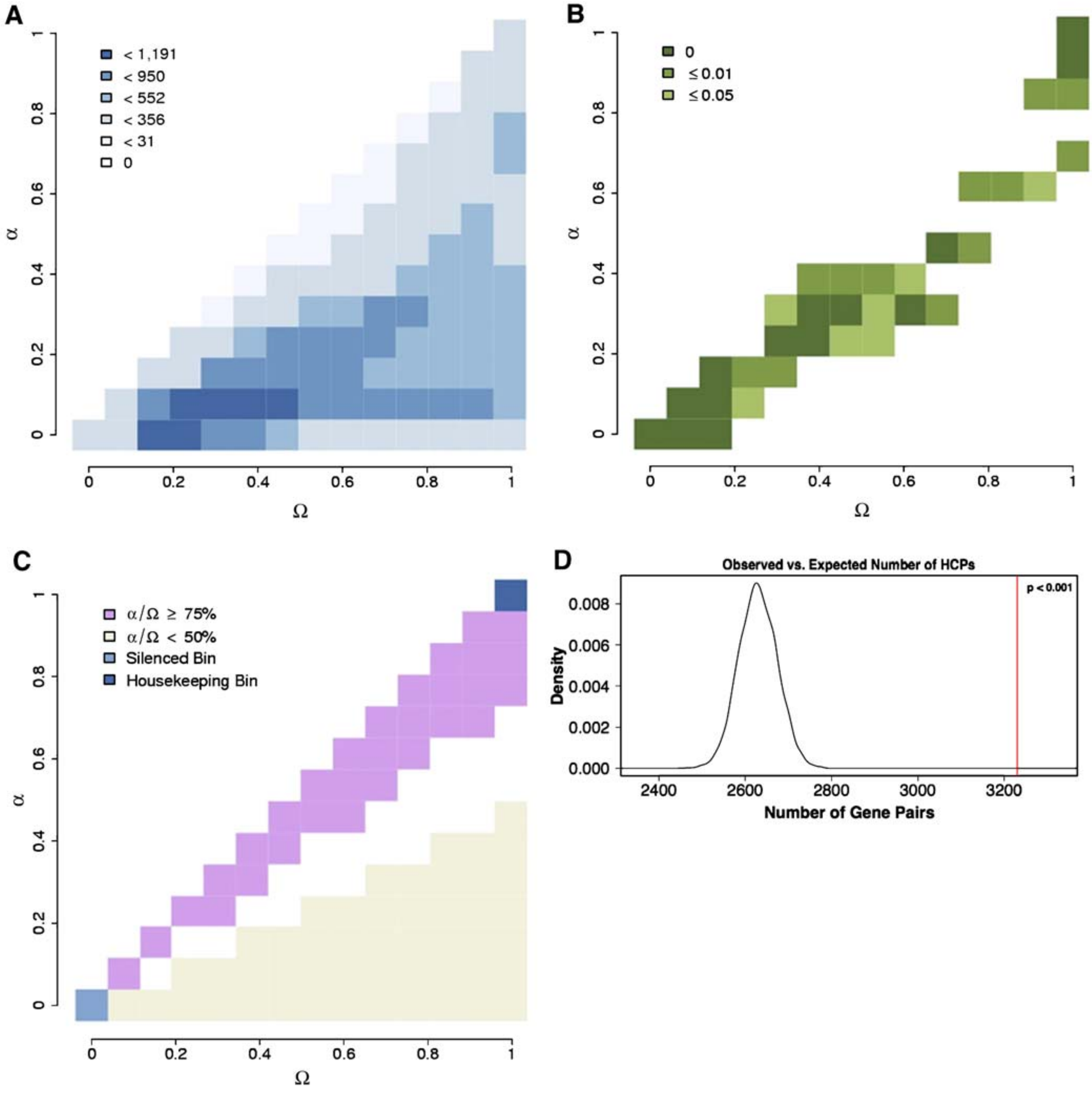


Fig. 2. Degree of coexpression of genomic neighbors defined by the coefficients α and Ω. We defined two coefficients, α and Ω, to quantify coexpression of a pair of genes in a set of *n* tissues. α is the proportion of tissues in which both genes are expressed, and Ω is the number of tissues in which either one or both genes are expressed divided by *n*. (A) Absolute bin occupancies, (B) empirical *p* values, and (C) defined coexpression categories based on the ratio between α and Ω. A high ratio between α and Ω indicates a high degree of coexpression. (D) Number of observed HCPs (red line) compared to the number of HCPs expected from permuted data.

The distribution of the bivariate coexpression measure (α, Ω) is nonrandom; certain combinations of α and Ω occur more frequently in the genome than expected if coexpression were independent of genomic location. Fig. 2B shows the *p* values for each combination of α and Ω, using the same permutation scheme as above. For each tuple (α, Ω), the empirical *p* value is given by the proportion of permutations in which equally many or more gene pairs display this coexpression pattern (α, Ω) than in the actual data.

Pairs were then assigned to one of the following coexpression categories, which depend on thresholds $\theta_{\mathrm{coex}}$ and $\theta_{\mathrm{unc}}$: (i) highly coexpressed, if $\alpha/\Omega \geq \theta_{\mathrm{coex}}$ and $\alpha < 1$; (ii) housekeeping, if $\alpha = 1$; (iii) silenced, if $\Omega = 0$; (iv) uncorrelated, if $\alpha/\Omega \leq \theta_{\mathrm{unc}}$. For the FANTOM3 data, we chose the thresholds $\theta_{\mathrm{coex}} = 0.75$ and $\theta_{\mathrm{unc}} = 0.5$ (Fig. 2C). This resulted in 3230 highly coexpressed pairs (HCPs), 154 housekeeping pairs, 36 silenced pairs, and 27,287 uncorrelated pairs (UCPs). Comparison of Fig. 1C with Fig. 1B shows that the number of

HCPs is larger than expected under the null hypothesis. Similarly, there are more housekeeping pairs, and more silenced pairs, while the frequency of UCPs is less than expected.

### Clusters decay at their flanks

We considered highly coexpressed clusters (HCCs) of genes, which consist of one or several neighboring HCPs. In each tissue, either all of the genes in the HCC or a subset of them are expressed. A feature that goes along with our definition of HCC is that there is a preference for the central genes in the HCC to be expressed, while the flanking genes are more likely to get lost, meaning that genes clustering in one tissue are expressed in shorter clusters or not as clusters at all in other tissues. Thus they show a pronounced directionality in how they decay across tissues (66% of triplets decay directed). Conversely, in the case of unrelated transcriptional regulation, proposed for uncorrelated clusters (UCCs), genes at any position in the cluster get lost at the same rate (84% of triplets decay undirected).

### Highly coexpressed clusters and housekeeping functionality

It has been suggested that housekeeping genes are often arranged in clusters along the genome [23]. However, the reverse is not true: most of our highly coexpressed pairs are expressed only in a limited fraction of tissues, hence these genes are not particularly housekeeping genes (Fig. 3).

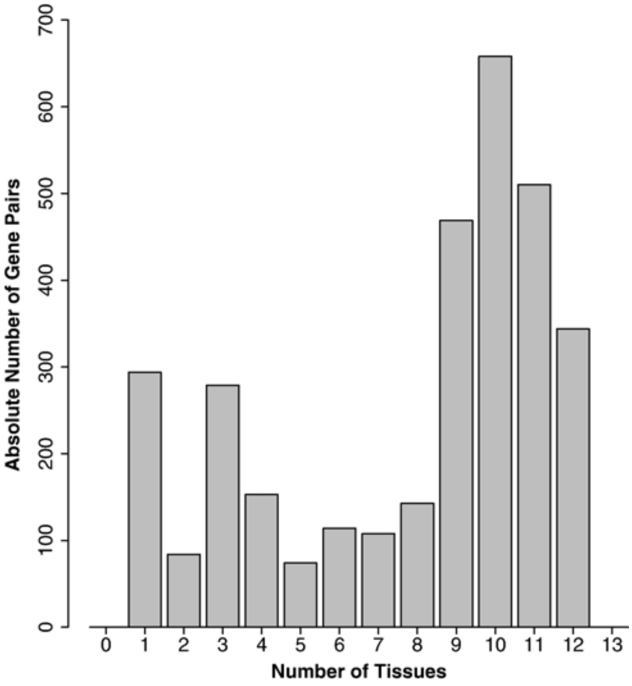


Fig. 3. Shared tissue expression of highly coexpressed gene pairs in mouse ($\alpha/\Omega \geq 0.75$ and $\alpha/\Omega < 1$). Shown is the number of gene pairs and their corresponding numbers of shared tissue expression. The number of tissues corresponds to α. Highly coexpressed gene pairs are not particularly housekeeping genes.

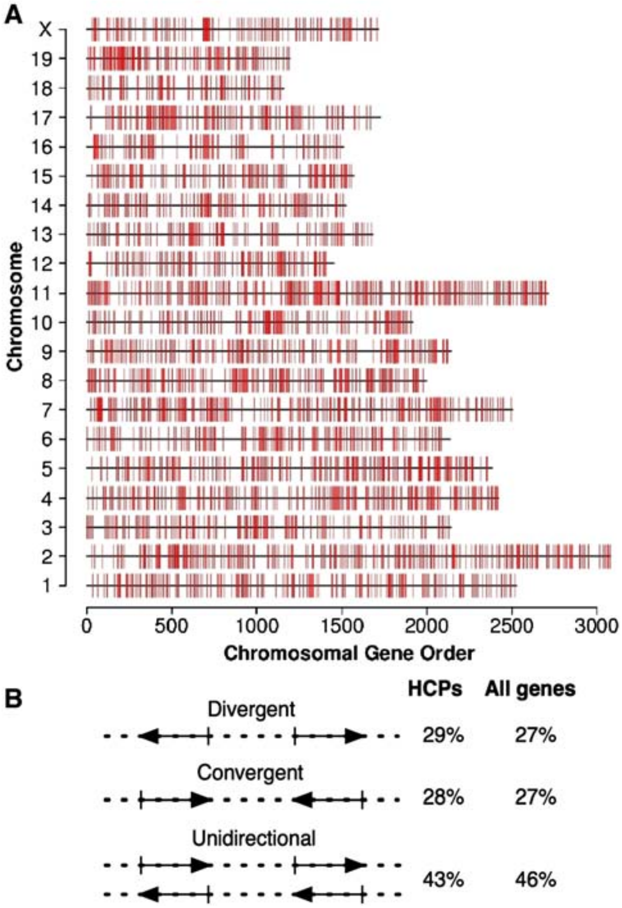


Fig. 4. Genomic distribution of highly coexpressed gene pairs. (A) Chromosomal gene order of HCPs (vertical red bars). Highly coexpressed gene pairs appear to distribute without notable prevalence among the chromosomes. (B) Distribution of strand orientations in HCPs compared to the overall distribution of orientations of adjacent genes in the mouse genome.

### Genomic location, orientation, and dimensions of HCPs

Fig. 4A provides an overview of the spatial distribution of HCCs and shows certain regions with slightly higher concentrations of genes, but an almost homogeneous distribution of these regions across all chromosomes.

We addressed whether HCPs are characterized by a particular genomic orientation that might affect their transcriptional coupling. For example, a divergent orientation would enable the sharing of regulatory sequences between adjacent genes. Therefore, we divided the HCPs into three groups based on their relative orientation (divergent, convergent, and unidirectional) and compared this grouping with all the genomic pairs of FANTOM3 regardless of their level of coexpression. As shown in Fig. 4B, we found the distribution of genomic orientations to be similar between HCPs and all genomic pairs.

We assessed the intergenic and transcriptional start site distances for HCPs and all genomic pairs. HCPs have smaller intergenic (median of 7662 bp versus 18,665 bp for all pairs, $p=3\times10^{-5}$, Wilcoxon rank sum test) and transcriptional start site distances (median of 28,781 bp versus 34,491 bp, $p=8\times10^{-8}$, Wilcoxon rank sum test). We were then interested in whether highly coexpressed clusters are characterized by limitations on their size, which could point to factors like chromatin remodeling contributing to the transcriptional coupling. We observed that the length of clusters measured in base pairs depends on the number of clustered genes. However, for HCCs the maximal observed number of adjacent genes within a cluster was limited to seven genes, and furthermore the 95% quantile of the cluster length was 320 kb (Fig. 5), which was much smaller compared to clusters of uncorrelated genes with 810 kb.

### *Functionality, paralogy, and transcriptional regulation of highly coexpressed gene clusters*

Paralogy, functional similarity, and the presence of common transcription factor binding sites (TFBSs) have been reported among genomic neighbors for a limited number of examples. We analyzed the frequency of shared Gene Ontology (GO) terms, protein domains, and TFBSs within adjacent gene pairs in FANTOM3 as a whole, as well as within HCPs. Naturally, our analysis was limited to only those genes annotated with Gene Ontology terms (36% FANTOM3 genes), protein domain information (42% FANTOM3 genes), or TFBS (33% FANTOM3 genes).

Table 1 shows that the sharing of domains and GO terms is less frequent in HCPs than in general genomic neighbors, whereas sharing of common TFBSs occurs at a similar rate. However, all of these observations are still more frequent than between nonneighboring genes, as is indicated by random permutations (as described above).

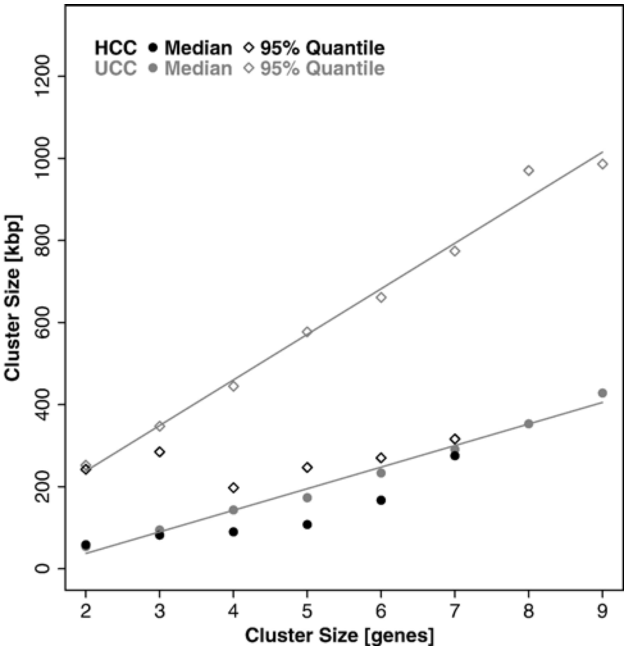


Fig. 5. Cluster sizes in base pairs and genes. Shown is the comparison of cluster sizes between HCCs and uncorrelated clusters measured in base pair length (*y* axis) as well as by the number of genes located in a particular cluster (*x* axis). The number of adjacent genes located in HCCs is limited to seven genes and their cluster length in base pairs is smaller compared to uncorrelated gene clusters.

Table 1

Functional and transcriptional properties of genomic neighbors in mouse

| | No. of annotated genomic pairs | Genomic pairs sharing similar annotations in % | No. of annotated HCPs | HCPs sharing similar annotations in % |
|---|---|---|---|---|
| GO terms | 5586 | 17.1 | 1272 | 8.8 |
| Protein domains | 7335 | 18.1 | 1567 | 10.8 |
| TF/TFBS | 4800 | 36.8/27.4 | 770 | 38.2/29.7 |

Genomic neighbors, irrespective of their coexpression, share Gene Ontology (GO) terms and protein domains to a much higher extent than highly coexpressed gene pairs (HCPs), whereas similar numbers of both groups of neighbors are potentially regulated by common transcription factors (TF) through their respective binding sites (TFBS).

To investigate the relationship between coexpression and paralogy, we analyzed gene pairs that had highly similar protein domains but showed only weak coexpression (in total 1307 gene pairs). Among them, we found members of well-known gene families that have previously been described to be clustered at certain genomic locations but to display tissue-specific expression nonetheless. For example, a family of S100 –calcium-binding proteins with its FANTOM3 tissue expression is depicted in Fig. 6 [24].

The 100 most frequently shared GO terms, protein domains, and transcription factors within general genomic neighbors are listed in the supplementary material.

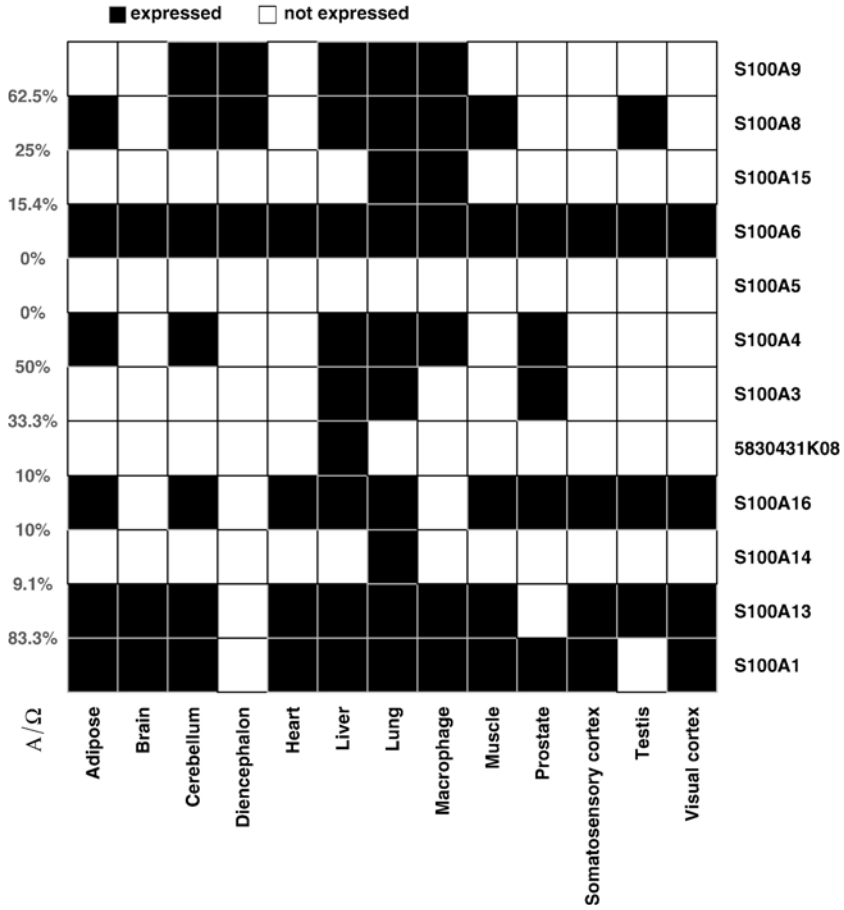
### *Comparing FANTOM3 with microarray data of human*

To verify that our observations are not limited to one single dataset and organism, we performed the same analyses as described above for the 79 tissues in the *H. sapiens* part of the GNF Symatlas dataset [18]. We used slightly relaxed thresholds for the definition of HCPs, $\theta_{coex}=0.50$ and $\theta_{unc}=0.33$, to account for the lower coverage and higher false negative rate of these data.

We observed highly similar results for all analyses performed with the FANTOM3 dataset, such as the chromosomal clustering of genes expressed in human tissues, the conservation of gene clusters across tissues, and the tissue distribution of observed HCCs, as well as for the relationship of functional similarity, paralogy, and transcriptional regulation. From these results, which are presented in the supplementary material, we conclude that our observations are not biased toward the FANTOM3 data and may hold true for mammals in general.

### *Phylogenetic conservation of highly coexpressed gene clusters*

To assess if the observed transcriptional coregulation of neighbored genes is phylogenetically conserved between *H. sapiens* and *M. musculus*, we extracted the human homologs of all FANTOM3 genes, resulting in 2245 gene pairs consisting of direct genomic neighbors in both species. We found that HCPs in mice also tend to be highly coexpressed in human, as the frequency of human HCPs among mouse



■ expressed ☐ not expressed

| A/Ω | Gene | Adipose | Brain | Cerebellum | Diencephalon | Heart | Liver | Lung | Macrophage | Muscle | Prostate | Somatosensory cortex | Testis | Visual cortex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S100A9 | ☐ | ☐ | ■ | ■ | ☐ | ■ | ■ | ■ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 62.5% | | | | | | | | | | | | | | |
| | S100A8 | ■ | ☐ | ■ | ■ | ☐ | ■ | ■ | ■ | ■ | ☐ | ☐ | ■ | ☐ |
| 25% | | | | | | | | | | | | | | |
| | S100A15 | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ■ | ■ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15.4% | | | | | | | | | | | | | | |
| | S100A6 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 0% | | | | | | | | | | | | | | |
| | S100A5 | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 0% | | | | | | | | | | | | | | |
| | S100A4 | ■ | ☐ | ■ | ☐ | ☐ | ■ | ■ | ■ | ☐ | ■ | ☐ | ☐ | ☐ |
| 50% | | | | | | | | | | | | | | |
| | S100A3 | ☐ | ☐ | ☐ | ☐ | ☐ | ■ | ■ | ☐ | ☐ | ■ | ☐ | ☐ | ☐ |
| 33.3% | | | | | | | | | | | | | | |
| | 5830431K08 | ☐ | ☐ | ☐ | ☐ | ☐ | ■ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10% | | | | | | | | | | | | | | |
| | S100A16 | ■ | ☐ | ■ | ☐ | ■ | ■ | ■ | ☐ | ■ | ■ | ■ | ■ | ■ |
| 10% | | | | | | | | | | | | | | |
| | S100A14 | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ■ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9.1% | | | | | | | | | | | | | | |
| | S100A13 | ■ | ■ | ■ | ☐ | ■ | ■ | ■ | ■ | ■ | ☐ | ■ | ■ | ■ |
| 83.3% | | | | | | | | | | | | | | |
| | S100A1 | ■ | ■ | ■ | ☐ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ☐ | ■ |

Fig. 6. Tissue expression of the S100 calcium-binding protein gene cluster. This cluster is likely to have originated by duplication of a common ancestor gene, as indicated by the genes' highly similar protein domain structures. Rows hold the genes, columns the FANTOM3 tissues. Black boxes indicate gene expression in a tissue and white boxes indicate no expression.

HCPs is 31% compared to 8% of human HCPs among mouse UCPs.

## Discussion

There is striking evidence that eukaryotic genomes show a high degree of gene organization. We focused on the large-scale, qualitative features of transcriptional regulation rather than on the fine-tuning. Thus we considered neighboring genes expressed in a particular tissue as coexpressed gene clusters regardless of expression levels of individual genes. We analyzed mouse transcriptomes of the FANTOM3 dataset derived by cloning techniques and confirmed our results on human data obtained through microarrays (GNF Symatlas).

Observing a surprisingly large number of gene clusters (sets of genomically neighboring, expressed genes), we suggested that only a subgroup of those clusters is actively transcriptionally coupled, whereas for a proportion of clusters, this observation would probably just be an effect of crowding of a given number of genes into a given genomic size. Therefore, we assessed the coexpression of gene clusters across tissues and observed a significant proportion of HCPs in addition to a small number of housekeeping gene clusters. This allowed us to extract sets of gene clusters (HCCs) that are characterized by a mainly directed decay across tissues, where the gene loss originates from only one end of the cluster and that may be indeed transcriptionally coupled. HCCs are characterized by clear-cut upper limits for physical cluster size and the number of genes making up these clusters, possibly reflecting the underlying mechanism. This finding may point to so far uncharacterized *cis*-acting units regulating the coexpression of certain sets of genes. To uncover such features further, our observation that HCCs are phylogenetically conserved between *M. musculus* and *H. sapiens* should provide the basis to extract potentially interesting conserved sequence features in the future. The coupling of highly coexpressed clusters could for example be controlled by histone modifications that are mediated by specific proteins initiating the opening or closing of chromatin and that spread along a chromosomal region until they meet a boundary element [23,25,26]. On the other hand, uncorrelated clusters probably arise as a consequence of intervening genes being transcriptionally silenced, for example, during cell differentiation. It has been shown that stem cells have a largely open chromatin formation and each step toward specialization is accompanied by down-regulation of genes in specific chromosomal regions [27]. These modifications could be stably inherited through cell division by DNA methylation, slowly reversed by silencing by histone lysine methylation or rapidly modulated by histone acetylation. For *S. cerevisiae* it has been reported that genes that are regulated by the same sequence-specific transcription factor tend to be regularly spaced across the genome [28]. Other reports suggested that the transcriptional regulation has shaped the organization of transcriptional units on the chromosome [29], and recently, genes controlled by the transcription factor *aire* were shown to be clustered along the genome [30]. These reports are in line with our finding that sharing of TFBSs is a general phenomenon among genomic neighbors, but furthermore, we saw that this does not simultaneously result in their coexpression, as TFBSs are shared at a similar rate between highly coexpressed and general genomic neighbors.

Considering a broader cluster definition of genomic genes independent of their expression, it has been shown that genes coding for proteins involved in the same metabolic pathways tend to appear in chromosomal clusters [31]. Also genes that are involved in stable protein-protein complexes tend to be more tightly linked than expected [32]. We assessed paralogy and functional similarity as potential driving forces for the arrangement of clusters. Previous reports have demonstrated the cofunctionality of coexpressed gene clusters [2,5,33–35] but did not investigate cofunctionality of genomic neighbors in general for comparison. However, highly coexpressed neighbors do not seem to have a higher degree of co-functionality than general genomic neighbors. This unexpected finding may be explained in light of models of gene duplication in which duplication leads to neofunctionalization and subfunctionalization. Neofunctionalization can result in expression of duplicate genes in tissues lacking expression of the ancestral gene, while subfunctionalization can result in division of the ancestral expression pattern onto duplicates [36–39].

We hypothesize that HCCs trace back to large-scale, persistent reorganizations of the transcriptome, while TF regulation is likely to handle the fine-tuning of transcription on shorter time scales. To date, the underlying mechanism of transcriptional coupling between genomic neighbors is a matter of speculation. Our data point to so far unknown conserved *cis*-acting units involved in this regulatory process in mammals. It is hoped that studies addressing the chromatin remodeling process, e.g., through histone modifications, modifying transcription factors, or the nuclear spacing of transcriptional events, will provide further insights.

## Methods

### *FANTOM3 dataset and chromosomal clustering*

We considered the gene set and gene ordering defined by FANTOM3, consisting of 39,593 genes mapped to build mm5of the mouse genome [22]. In FANTOM3 genes are defined as transcriptional units (TU), representing discontinuous genomic regions from which one mature mRNA is derived. We used the term “gene” synonymously for TUs throughout this report. For cluster analyses, we concentrated on genes expressed in the following tissues, from which the expression information was obtained using CAGE technology (numbers of expressed genes): adipose (19,166), brain (13,766), cerebellum (18,753), diencephalon (6567), heart (8423), liver (30,721), lung (30,560), macrophage (26,746), muscle (8829), prostate gland (10,795), somatosensory cortex (17,193), testis (13,347), visual cortex (17,216). A set of physically neighboring genes coexpressed in a particular tissue is called a cluster of coexpressed genes.

### *Transcription factor binding sites*

TFBSs conserved in human/mouse/rat alignments were considered as annotated by the UCSC Genome Browser (http://genome.ucsc.edu/). Based on the ENSEMBL gene IDs of FANTOM3, we extracted all TFBSs annotated in the 10-kb upstream region of each gene. Taking into account that the TFBS annotation of the UCSC Genome Browser is conservative, we considered two genes to have common cis-acting regulatory units if they shared one or more TFBSs. The numbers of shared TFBSs per coexpressed gene pair were compared to a randomly built dataset (1000 permutations).

### *Gene ontology*

For each gene in the FANTOM3 data, we extracted the most specific GO terms to which that gene had been annotated and disregarded their parental terms [40]. We defined that two genes have a similar GO annotation if they shared at least 50% of these most specific terms as annotation. The GO terms for the genes were obtained from the given RefSeq and LocusLink identifiers using the Bioconductor software package “biomaRt” to query the Ensembl database (build 33, May 2005) [41–43]. We compared the concordance of GO annotation for gene pairs of interest with the one expected by random permutation of gene order [44].

### *Protein domain information*

We extracted the available domain information for the encoded protein of each gene in the FANTOM3 data. We considered two genes sharing at least 50% of their domains to have a similar domain annotation. Again, the annotations were obtained from the given RefSeq and LocusLink identifiers using the Bioconductor software package biomaRt to query the Ensembl database (build 33, May 2005). To compare observed similarities in protein domain annotation for gene pairs of interest with the one expected by chance, we employed the same permutation method as for the Gene Ontology annotation (see above).

## Acknowledgments

We thank Anja von Heydebreck, Steffen Grossmann, Stefanie Hammer, and Martin Vingron for helpful discussions. We also thank Claudio Lottaz, Steffen Durinck, and Raffaello Galli for their assistance with GNF Symatlas annotation, the biomaRt package, and transcription factor binding site annotation. This work was funded by the Max Planck Society for the Advancement of Science. The preparation of the FANTOM3 dataset was funded by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science, and Technology of the Japanese Government to Y.H.; a grant from the Genome Network Project of the Ministry of Education, Culture, Sports, Science, and Technology, Japan; and The Strategic Programs for R&D of RIKEN. We also thank Chikatoshi Kai and Jun Kawai for their contribution to FANTOM3.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.01.010.

## References

[1] R.J. Cho, et al., A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.

[2] B.A. Cohen, R.D. Mitra, J.D. Hughes, G.M. Church, A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat. Genet. 26 (2000) 183–186.

[3] T. Blumenthal, et al., A global analysis of *Caenorhabditis elegans* operons, Nature 417 (2002) 851–854.

[4] P.J. Roy, J.M. Stuart, J. Lund, S.K. Kim, Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans, Nature 418 (2002) 975–979.

[5] P.T. Spellman, G.M. Rubin, Evidence for large domains of similarly expressed genes in the Drosophila genome, J. Biol. 1 (2002) 5.

[6] H. Caron, et al., The human transcriptome map: clustering of highly expressed genes in chromosomal domains, Science 291 (2001) 1289–1292.

[7] R. Versteeg, et al., The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes, Genome Res. 13 (2003) 1998–2004.

[8] J.D. Barrans, et al., Chromosomal distribution of the human cardiovascular transcriptome, Genomics 81 (2003) 519–524.

[9] S. Bortoluzzi, et al., A comprehensive, high-resolution genomic transcript map of human skeletal muscle, Genome Res. 8 (1998) 817–825.

[10] B.L. Gabrielsson, B. Carlsson, L.M. Carlsson, Partial genome scale analysis of gene expression in human adipose tissue using DNA array, Obes. Res. 8 (2000) 374–384.

[11] S. Minagawa, K. Nakabayashi, M. Fujii, S.W. Scherer, D. Ayusawa, Functional and chromosomal clustering of genes responsive to 5-bromodeoxyuridine in human cells, Exp. Gerontol. 39 (2004) 1069–1078.

[12] E. Soury, et al., Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver, Immunogenetics 53 (2001) 634–642.

[13] H. Zhang, K.H. Pan, S.N. Cohen, Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci, Proc. Natl. Acad. Sci. USA 100 (2003) 3251–3256.

[14] J.H. Vogel, A. von Heydebreck, A. Purmann, S. Sperling, Chromosomal clustering of a human transcriptome reveals regulatory background, BMC Bioinformatics 6 (2005) 230.

[15] D. Schubeler, et al., Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus, Genes Dev. 14 (2000) 940–950.

[16] F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster, Cell 113 (2003) 405–417.

[17] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, Nat. Genet. 31 (2002) 180–183.

[18] A.I. Su, et al., A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.

[19] C. Pal, L.D. Hurst, Evidence for co-evolution of gene order and recombination rate, Nat. Genet. 33 (2003) 392–395.

[20] J.W. Pepper, The evolution of evolvability in genetic linkage patterns, Biosystems 69 (2003) 115–126.

[21] G.A. Singer, A.T. Lloyd, L.B. Huminiecki, K.H. Wolfe, Clusters of co-expressed genes in mammalian genomes are conserved by natural selection, Mol. Biol. Evol. 22 (2005) 767–775.

[22] P. Carninci, et al., The transcriptional landscape of the mammalian genome, Science 309 (2005) 1559–1563.

[23] L.D. Hurst, C. Pal, M.J. Lercher, The evolutionary dynamics of eukaryotic gene order, Nat. Rev. Genet. 5 (2004) 299–310.

[24] I. Marenholz, C.W. Heizmann, G. Fritz, S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature), Biochem. Biophys. Res. Commun. 322 (2004) 1111–1122.

[25] S. Cai, H.J. Han, T. Kohwi-Shigematsu, Tissue-specific nuclear architecture and gene expression regulated by SATB1, Nat. Genet. 34 (2003) 42–51.

[26] R. van Driel, P.F. Fransz, P.J. Verschure, et al., The eukaryotic genome: a system regulated at different hierarchical levels, J. Cell Sci. 116 (2003) 4067–4075.

[27] K. Akashi, et al., Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis, Blood 101 (2003) 383–389.

[28] F. Kepes, Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites, J. Mol. Biol. 329 (2003) 859–865.

[29] R. Hershberg, E. Yeger-Lotem, H. Margalit, Chromosomal organization is shaped by the transcription regulatory network, Trends Genet. 21 (2005) 138–142.

[30] J.B. Johnnidis, et al., Chromosomal clustering of genes controlled by the aire transcription factor, Proc. Natl. Acad. Sci. USA 102 (2005) 7233–7238.

[31] J.M. Lee, E.L. Sonnhammer, Genomic gene clustering analysis of pathways in eukaryotes, Genome Res. 13 (2003) 875–882.

[32] S.A. Teichmann, R.A. Veitia, Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective, Genetics 167 (2004) 2121–2125.

[33] Y. Fukuoka, H. Inaoka, I.S. Kohane, Inter-species differences of co-expression of neighboring genes in eukaryotic genomes, BMC Genomics 5 (2004) 4.

[34] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression analysis of human genes across many microarray data sets, Genome Res. 14 (2004) 1085–1094.

[35] H.R. Ueda, et al., Genome-wide transcriptional orchestration of circadian rhythms in Drosophila, J. Biol. Chem. 277 (2002) 14048–14052.

[36] L. Huminiecki, K.H. Wolfe, Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse, Genome Res. 14 (2004) 1870–1879.

[37] M. Lynch, Gene duplication and evolution, in: A. Moya, E. Font (Eds.), Evolution: from Molecules to Ecosystems, Oxford Univ. Press, Oxford, 2004, pp. 33–47.

[38] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.

[39] S. Ohno, Evolution by Gene and Genome Duplication, Springer-Verlag, Berlin, 1970.

[40] M. Ashburner, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.

[41] S. Durinck, et al., BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis, Bioinformatics 21 (2005) 3439–3440.

[42] R.C. Gentleman, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (2004) R80.

[43] E. Birney, et al., Ensembl 2006, Nucleic Acids Res. 34 (2006) D556–D561.

[44] R.C. Gentleman, D. Scholtens, B. Ding, V.J. Carey, W. Huber, Case studies using graphs on biological data, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer-Verlag, New York, 2005.

BMC Bioinformatics

Research article

Open Access

# Chromosomal clustering of a human transcriptome reveals regulatory background

Jan H Vogel[1], Anja von Heydebreck[2], Antje Purmann[1] and Silke Sperling*[1]

Address: [1]Cardiovascular Genetics Group, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany and [2]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Email: Jan H Vogel - jan.vogel@gmail.com; Anja von Heydebreck - Anja.von.Heydebreck@merck.de; Antje Purmann - purmann@molgen.mpg.de; Silke Sperling* - sperling@molgen.mpg.de

* Corresponding author

Published: 19 September 2005

*BMC Bioinformatics* 2005, **6**:230 doi:10.1186/1471-2105-6-230

Received: 17 May 2005
Accepted: 19 September 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/230

© 2005 Vogel et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** There has been much evidence recently for a link between transcriptional regulation and chromosomal gene order, but the relationship between genomic organization, regulation and gene function in higher eukaryotes remains to be precisely defined.

**Results:** Here, we present evidence for organization of a large proportion of a human transcriptome into gene clusters throughout the genome, which are partly regulated by the same transcription factors, share biological functions and are characterized by non-housekeeping genes. This analysis was based on the cardiac transcriptome identified by our genome-wide array analysis of 55 human heart samples. We found 37% of these genes to be arranged mainly in adjacent pairs or triplets. A significant number of pairs of adjacent genes are putatively regulated by common transcription factors ($p = 0.02$). Furthermore, these gene pairs share a significant number of GO functional classification terms. We show that the human cardiac transcriptome is organized into many small clusters across the whole genome, rather than being concentrated in a few larger clusters.

**Conclusion:** Our findings suggest that genes expressed in concert are organized in a linear arrangement for coordinated regulation. Determining the relationship between gene arrangement, regulation and nuclear organization as well as gene function will have broad biological implications.

## Background

To understand the global regulatory network underlying specific transcriptomes, several distinct aspects have to be considered [1,2]; (A) the genomic organization of those transcripts [3], (B) their regulation by general and specific transcription factors, (C) the influence of epigenetic effects such as e.g. histone modifications [4], (D) the local environment [5,6], and (D) the functional role of the transcripts as well as their protein products as nodes of the network. In the present report, we show for the first time for a human transcriptome that there is a relationship between the genomic organization, transcriptional regulation and functional role.

It has long been known that transcriptional regulation is related to chromosomal gene order; prokaryotic operons are the best known example [7]. For lower eukaryotes coexpressed adjacent genes were first described in

Saccharomyces cerevisiae [8,9]. For a part of those gene pairs, a common transcriptional activation was proposed through a shared upstream activating sequence, which occurs in the promoter region of one of the two genes. Furthermore, correlated triplets, but not quadruples, were found to occur more often than expected in yeast. Reports for Caenorhabditis elegans [10], Drosophila melanogaster [11,12], Homo sapiens and Mus musculus [13-15] showed coexpression of co-localized genes in higher eukaryotes, and reports of particular gene cluster such as the human β-globin locus [16], the interleukin-13 gene locus [17] and others [18-20] indicate the association to regulated chromatin domains. However, on a global scale only few insights into the molecular mechanisms of the transcriptional regulation of clustered genes have been gathered so far, and data about small clusters of adjacent genes have only been partially analyzed. Beside the finding of housekeeping gene clusters throughout the human genome [14], no evidence for a functional correlation of clustered genes had been shown. However, coexpressed genes in general (regardless of their localization) appear to function in similar biological processes [21].

In this paper we describe the chromosomal co-localization in adjacent pairs of a large proportion of the human transcriptome in heart in the context of their expression dynamics, their transcriptional regulation and their function in shared biological processes. We examined the cardiac transcriptome using data from our previous genome-wide array analysis [22] and found profound evidence for a significant clustering of more than 37% of those genes located mainly in pairs or triplets. A significant proportion of these clustered genes have common putative transcription factor binding sites within their promoter regions and share common biological functions.

## Results

To characterize genomic organization of the cardiac transcriptome, we investigated a set of 3.172 heart-expressed genes (HXP) identified in our previous study that represent the cardiac transcriptome based on the analysis of 55 human heart samples [22]. This gene set reflects all genes continuously transcribed in the analyzed heart samples. Thus, we focused on the information whether or not a gene and herewith a particular genomic region is transcribed at all and considered expressed neighboring genes as coexpressed gene clusters regardless of the expression levels of individual genes. We assigned the position of this HXP set with regard to the whole human genome as represented by Ensembl compared to the HU2 gene set represented on the arrays. In order to reflect the actual adjacency of genes on the chromosomes, we defined gene neighbors according to their Ensembl annotation, rather than using the HU2 gene set as a basis (Fig. 1).

### *Chromosomal distribution and gene clusters*

First, we analyzed the overall chromosomal distribution of HXP and observed no overrepresentation of HXP on specific chromosomes (Chi-Square-Test, $p = 0.2$). On average each chromosome contained a proportion of 18% HXP genes out of the overall analyzed dataset. Upon closer analysis of the distinct localization of HXP genes, we observed small groups of physically adjacent genes along the chromosomes throughout the genome. In Figure 2 the chromosomal distribution of the overall HXP genes and neighboring coexpressed HXP genes is represented. We calculated the number of gene clusters made up of two to five physically adjacent HXP genes and measured the statistical significance of this local clustering by comparing the numbers of HXP gene clusters with a random distribution obtained by 100,000 permutations of HU2. We observed a significantly higher number of adjacent gene pairs (881 genes, $p = 0.01$) and gene triplets (307 genes, $p = 0.02$) in HXP than would be expected for a random distribution, whereas the number of quadruples and quintuples did not differ significantly (Table 1). In total, we found 1,179 HXP genes to be locally clustered. Further, we analyzed whether there was any prevalence regarding gene orientation within these clusters compared to a random distribution obtained by 10,000 permutations. Besides an enrichment of co-oriented gene pairs within clusters of size ≥ 2 ($p = 0.03$), we observed no bias in the numbers of anti-oriented gene pairs in clusters ≥ 2 ($p = 0.2$) as well as for co- and anti-oriented gene triplets in clusters ≥ 3 ($p = 0.3$ and $p = 0.6$, respectively).

### *Gene clusters and housekeeping functionality*

Previously, it had been suggested that housekeeping genes are arranged in clusters along the genome [14]. Therefore, we assessed the tissue expression of our HXP gene set and the subset of coexpressed adjacent genes in 79 human tissues, for which the expression information of protein-coding genes had been recorded by the GNF Symatlas [15] (Fig. 3). We observed a slightly bimodal distribution for the overall HXP gene set, with a major peak corresponding to expression in 79 tissues. This distribution differed from the one observed for coexpressed adjacent genes. Here, the majority of genes showed an expression in a distinct number of tissues, not reflecting a housekeeping-like expression profile.

### *Transcriptional regulation of gene clusters*

We extended our analysis to determine to what extent these gene clusters are regulated by common transcription factors. For this purpose, we identified the putative transcription factor binding sites (TFBS) for the HXP set using the CORG database [23]. CORG is based on TFBS in non-coding regions conserved between human and mouse (see Methods), and enabled us to identify binding sites of 276 distinct binding site models in the promoter regions

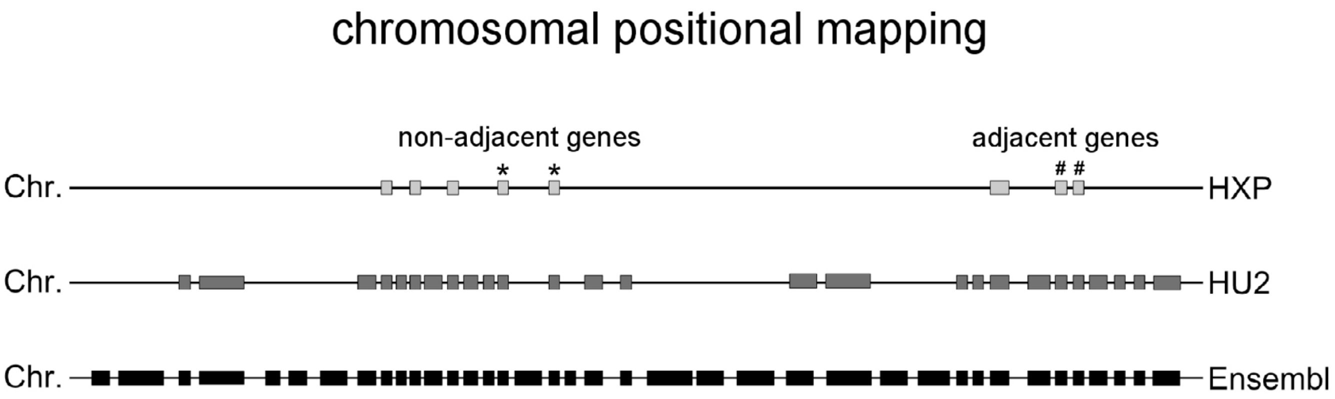


**Figure 1**
**Gene localization based on Ensembl, HU2 and HXP**. Genes marked with * illustrate an adjacent gene pair in reference to HU2 and HXP, but not referring to the Ensembl gene set (131 cases). Therefore, only genes like those marked with # are noted as clusters throughout.

of a total of 1,777 HXP genes. For the remaining HXP genes no conserved non-coding region could be identified.

Taking into account that several models can represent binding sites of one transcription factor, the identified TFBS pertained to only 216 distinct transcription factors. Within the HXP set with assigned TFBS, 501 genes belonged to chromosomally clustered gene pairs and for 171 pairs, binding sites for both genes were predicted [see Additional file 1]. We observed the largest number of common predicted transcription factors for the antisense-sense gene pair consisting of the *HOOK2* protein and transcription factor *JUNB*, which share binding sites for 64 transcription factors. The mean number of observed common transcription factors predicted to regulate adjacent genes was 4.3. There were 23 genes in the HXP set with binding sites for more than 75 different transcription factors, which we consider to be outliers.

Finally, we tested the significance of the observed number of gene pairs potentially regulated by at least one common transcription factor by comparing this number to the numbers of gene pairs with commonly assigned transcription factors seen in 10,000 random permutations of HXP genes. Regardless of whether the calculation was based on the number of distinct TFBS or distinct transcription factors, the common transcriptional regulation was significantly greater within the observed gene clusters than expected from the random distribution ($p = 0.02$ and $p = 0.03$, respectively). This significance was not influenced by genes defined as outliers with regard to their large number of common TFBS. Furthermore, the identified common transcription factors belonged to a broad panel of transcription factors classes without prevalence of particular families. Taking the gene orientation into account again, we found a homogenous distribution of strand orientation within observed gene pairs regulated by common transcription factors. An example of a gene cluster regulated by common transcription factors is given in Figure 4.

### *Functional relationship of gene clusters*

Further, we analyzed to what extent heart-expressed genes organized as adjacent pairs are involved in the same biological process using the Gene Ontology (GO) database as a source of annotations of biological functions. Naturally, our analysis was limited to those genes annotated with GO terms that were around 60% of all HXP genes (1,921 genes) including those located in 176 gene clusters. A total of 2,158 different GO terms could be assigned to the overall HXP set with 1,241 GO terms mapping to genes located in clusters. To focus on non-generic functional annotations, we calculated the number of GO terms shared within pairs of adjacent HXP genes for the 5th and 6th level of granularity of the GO hierarchy and compared these to a random distribution generated from HXP genes to account for the tissue-specificity of the analyzed dataset. We observed a significant enrichment of shared GO terms within pairs of adjacent genes for both levels of granularity ($p << 0.0001$ for each). Only eight of the 96 and 45 pairs sharing the 5th and 6th level of functional classification are conspicuous in terms of having originated by gene duplications [see Additional file 2] (for an example see Fig. 4).

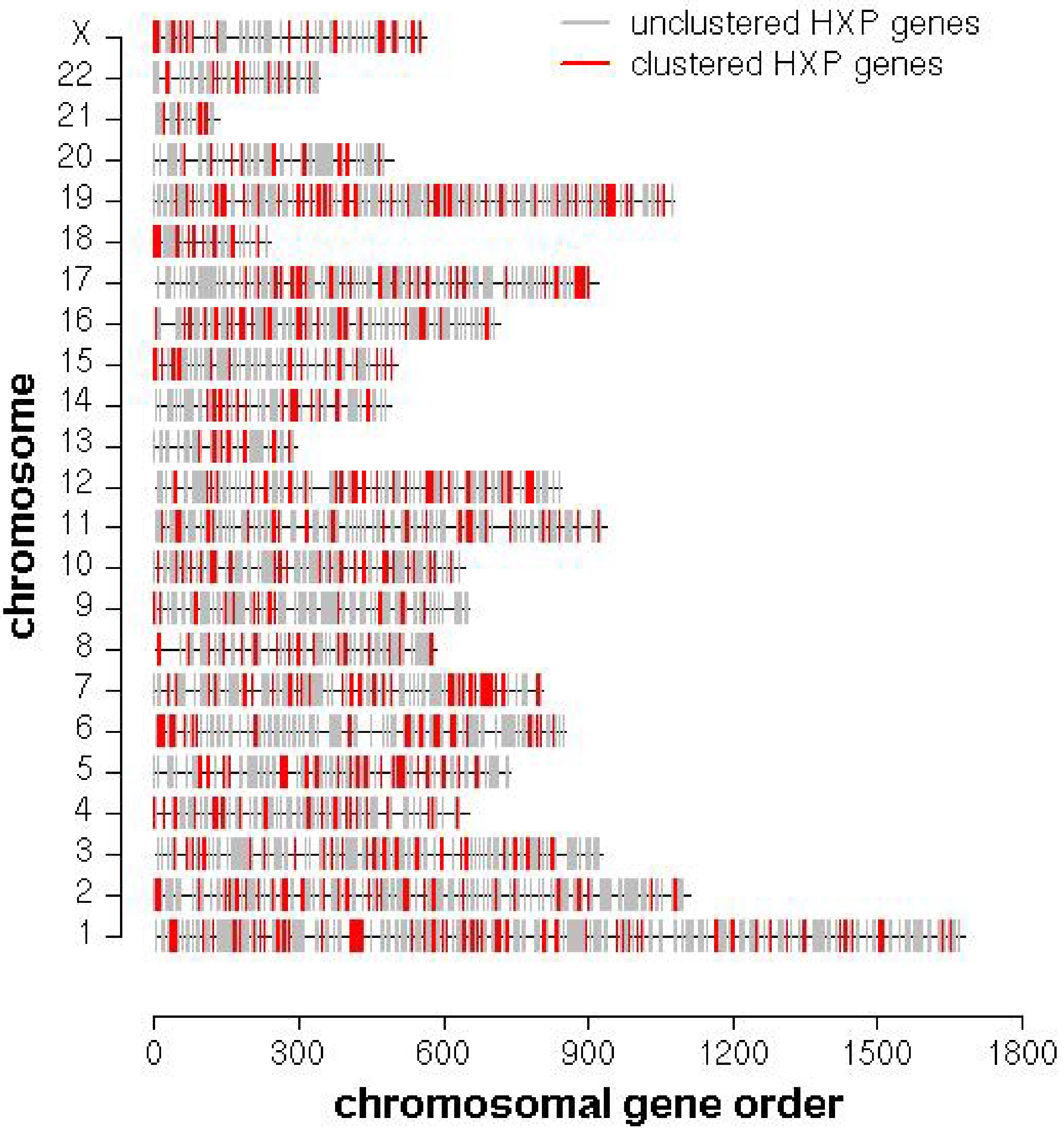


**Figure 2**

**Genomic organization of the overall and clustered heart-expressed genes**. The chromosomal gene order of clustered genes is represented in red, whereas the non-clustered genes are shown in gray, both appear to distribute without notable prevalence among the chromosomes.

***Expression pattern of adjacent genes***

Finally, we attempted to determine whether genes located close to each other show a correlation of their expression levels within our previously analyzed patient cohort, which consisted of 5 cardiac phenotypes with characteristic expression profiles [22]. First, we built correlation

**Table 1: Numbers of detected heart-expressed gene clusters of different sizes.**

| Cluster-size | Number of gene cluster | ENSG | p-value | Z-score |
|---|---|---|---|---|
| 2 | 480 | 881 | 0.01 | 3.89 |
| 3 | 79 | 207 | 0.02 | 3.24 |
| 4 | 15 | 57 | 0.05 | 2.68 |
| 5 | 1 | 5 | 0.7 | 0.07 |

ENSG represents the number of unique Ensembl genes in all clusters of the given size. The p-values and Z-scores result from a comparison with a random distribution obtained by 100,000 permutations of the HU2 dataset.

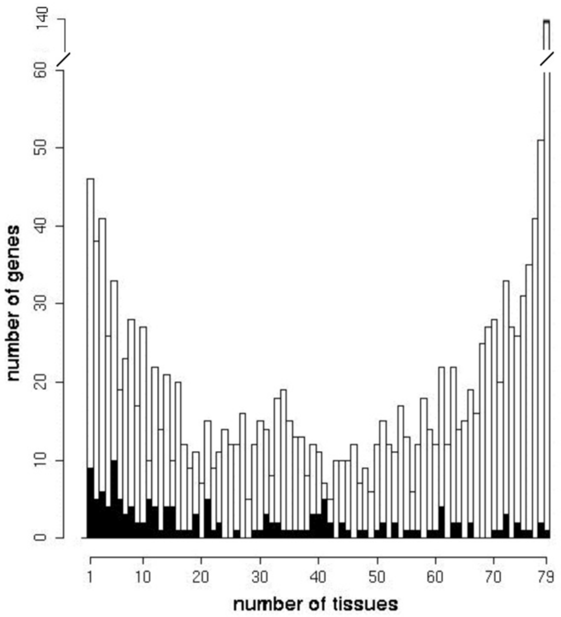


**Figure 3**
**Tissue expression of overall and clustered HXP genes**. The numbers of tissue expression of clustered heart-expressed genes are presented filled black. Of the 79 analyzed tissues, the expression profile of clustered genes shows a broad panel of tissue expression but only very few clustered genes are expressed in the majority of tissues. Thus we concluded that clustered coexpressed genes are mainly non-housekeeping genes.

maps for visual inspection of the overall HXP gene set. To construct such matrices, we sorted the HXP genes of each chromosome according to their arrangement on the chromosome and calculated the correlation of expression of each possible gene pair using Pearson's correlation coefficient. We observed groups of coexpressed genes (*cor* $\geq 0.5$) on several chromosomes. As an example, the correlation matrix of human chromosome 10 shows areas of coexpressed genes between genes 7–10 and 27–30 [see Additional file 3]. Correlation maps provide a gross overview of the coexpression of genes located nearby each other. Therefore, we further analyzed the coexpression of gene pairs with respect to their distance on two different scales: the base pair distance and the number of genes located between pairs. For neither of these scales do we observe significant coexpression of co-localized gene pairs.

## Discussion

Our data suggest that a large proportion of the cardiac transcriptome in human is linearly arranged in small groups of adjacent genes such that the genes within each group tend to be regulated by the same transcription factors and appear to share granular biological processes. Even though the numbers of shared transcription factors and shared GO terms are small, they are considerable larger than what could be expected by chance. These findings provide powerful evidence that gene clustering plays a potentially important role in concerted gene regulation through the location of regulatory elements with respect of the regulated genes.

We focused our analysis on the information whether or not a gene and herewith a particular genomic region is transcribed at all in the human heart and did not consider rates of transcription. The proportion of genes arranged in clusters exceeded the number reported previously, which could be explained by different definitions of coexpression, as other reports defined coexpression mainly based on the correlation coefficient of continuous valued expression levels or considered only highly expressed genes [13]. Creating an expression neighborhood through the localization of regulatory elements would be an efficient means to increase regional gene activity, which has been shown to be influenced by the local concentration of regulatory proteins [24-26]. In addition, such regulatory elements may influence the expression status within a chromosomal region e.g. by spreading of histone modifications. Further support for concerted regulation of clustered genes is provided by our observation of functional relatedness between clustered genes. Apart from the identification of housekeeping genes arranged in clusters such relatedness has not been reported in human [14].

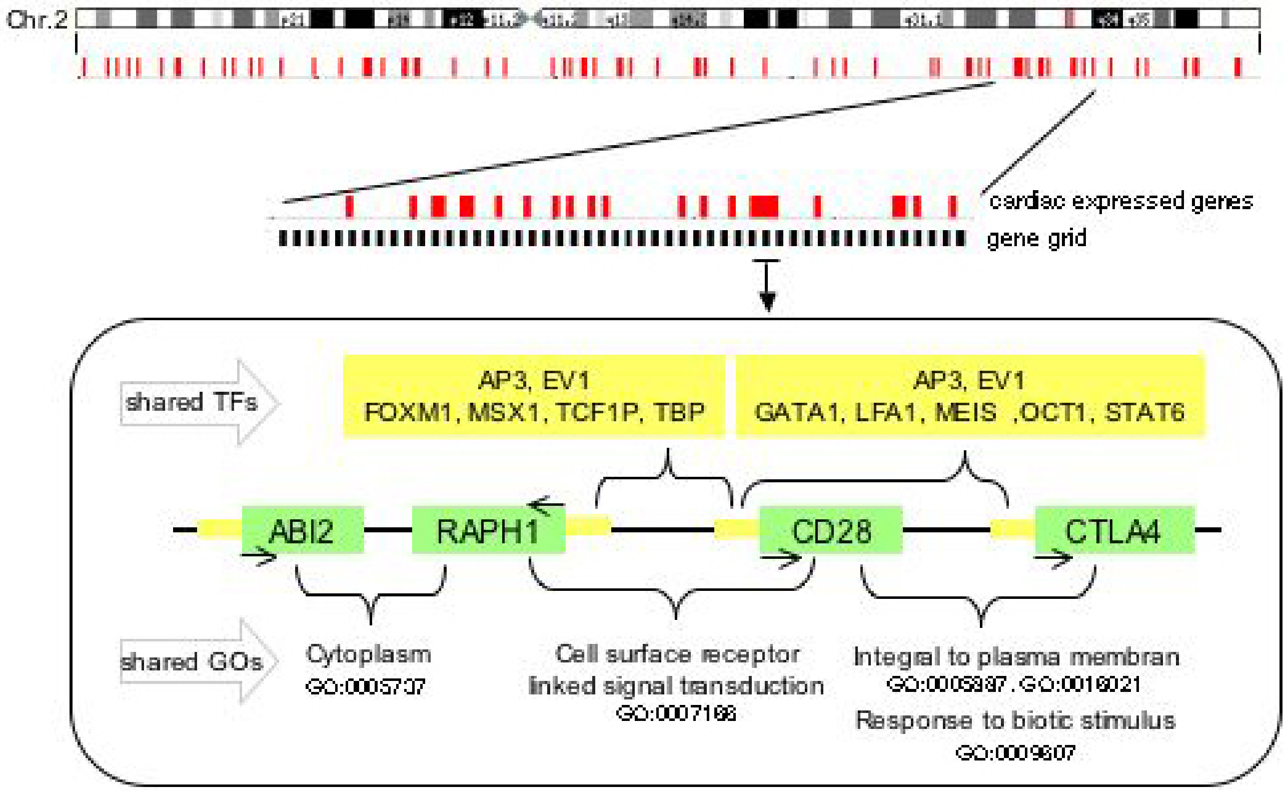


**Figure 4**

**Example of four adjacent coexpressed genes**. Shown is the gene cluster distribution on human chromosome 2 with an example of four adjacent genes regulated in part by common transcription factors and sharing gene ontology categories. Coexpressed gene clusters are shown in red. Each coexpressed gene pair ABI2 – RAPH1; RAPH1 – CD28 and CD28 – CTLA4 shares GO terms as indicated. Genes of the triplet RAPH1 – CD28 – CTLA4 have transcription factor binding sites for AP3 and EV1 in common. Furthermore, each gene pair RAPH1 – CD28 and CD28 – CTLA4 shares additional TFBS. Genes are marked in green, arrows indicate the strand orientation, promotor regions and transcription factors are colored in yellow.

However without considering gene localization, it has been shown recently that coexpressed genes are often functionally related [21]. Further studies will be required to show if there is an evolutionary constraint upon the disruption of co-localized genes. So far, sequencing projects are still in process and a comparison between close relatives such as human and for example mouse would not be sufficient due to their high synteny. From our analysis, we propose that duplication events alone are not sufficient as an explanation.

In the past some reports suggested that tissue-specific genes, e.g. genes specifically expressed in human skeletal muscle and adipose tissue, are preferentially located on certain chromosomes [27,28]. For the cardiovascular transcriptome in human, a disparate chromosomal distribution has been reported with enrichment on chromosomes 17, 19 and 22 [29,30]. With the present knowledge of genome annotation that reveals the inhomogeneous chromosomal distribution of the human genome, we cannot confirm such observations for the cardiac transcriptome.

Finally, we analyzed correlation of expression levels of genes located in clusters, as it has been suggested for the human transcriptome and other organism. By using the base pair distance as well as the gene distance between gene pairs, we failed to observe any significant correlation between co-localization and expression levels. Recently, it has been proposed that such correlation could be influenced by the probe localization on the array [31], which was corrected for in our primary array analysis. Further, we took into account that our dataset was based on different cardiac phenotypes caused by multiple factors, which could increase background noise and reduce the ability to recognize distinct coexpression of genes in our sample collection.

## Conclusion

In summary, we provide evidence that the linear arrangement of genes expressed in concert is due to coordinated regulation by common transcription factors. We suggest that determining the relationship between nuclear organization and gene arrangement will lead to a deeper understanding of how transcriptomes, dedicated to a particular cellular function or fate, are controlled. Here, meta-analysis of the large-scale transcriptome array data beginning to appear in public repositories, could build the basis for the discovery of the nodes between gene regulation and nuclear organization. Those analyses could provide insights into a regulatory constraint such that genes localized in clusters tend to be coregulated throughout several tissues. Furthermore, such a regulatory constraint may be a crucial factor in the development of human diseases caused by partial deletions or insertions of chromosomal units separating genes regulated in concert.

## Methods

### HXP dataset

The data composition and the classification of expressed genes was done using the Human Unigene II clone set containing 74.695 IMAGE clones [22]. The genomic localizations of the clones were determined via Ensembl and CrossMatch sequence comparison. In summary, 40.416 clones were assigned to 16.260 Ensembl genes, which resulted in 67% coverage of the Ensembl dataset version 11.31.1. In a previous array study, we identified the cardiac transcriptome by hybridisation experiments of 55 cardiac samples using arrays containing the above IMAGE clones [22]. The finally defined heart-expressed subset (HXP) of IMAGE-clones contained 3.172 Ensembl genes represented by 4.167 clones. This gene set refers to the 15% highest expressed genes in at least 4 heart samples whose natural log ratios had a standard deviation across the samples of at least 0.5, and were obtained out of the overall set of human genes (HU2) after normalization of expression levels for array position and averaging over duplicates. The analyzed samples belonged to four categories, (1) the normal right/left atrial and ventricular samples, (2) right atrial and ventricular samples obtained from patients with Tetralogy of Fallot, (3) right atrial samples of patients with ventricular septal defect, and (4) right atrial samples of patients with atrial septal defect. For the further analysis, we excluded the Y chromosome since the dataset was composed of samples from males and females. Functional categories of these genes were assigned using the Gene Ontology classification.

### GFN Symatlas

To analyze the distribution of tissue expression of the HXP dataset, we used the microarray gene expression information for 79 human tissues in the GNF Symatlas dataset [15]. This dataset contained almost 34,000 probe sets with 'present', 'marginal', or 'absent' calls for each probe set in each tissue. We considered a probe set expressed if it had a 'present' or 'marginal' call. If probe sets with different expression calls had the same chromosomal location, we considered the 'present' or 'marginal' call in case where one of the probe sets had an 'absent' call. Probe Sets with ambiguous location information were excluded, such that the resulting dataset consisted of 10,715 Ensembl genes with distinct chromosomal locations. Based on the Ensembl gene IDs we could map 1,600 HXP genes including 183 clustered genes to expression information represented by the GNF Symatlas.

### Physical annotation of gene distances

To measure the distance of a pair of genes, we used the number of base pairs between them as well as defining the distance of a pair of genes with respect to the amount of Ensembl genes located between them (Fig. 1). A pair of genes was defined as adjacent if there are no other genes in the Ensembl dataset that lie between the two different gene loci. By using these different scales, we were able to distinguish between gene pairs located very close to each other in a chromosomal region with high gene density, as well as describing the relationship of a gene pair independent of its physical localization on the chromosome. Groups or pairs of directly neighboring genes that we analyzed are referred to as gene clusters. The definition of 'neighborhood' refers to the genes included in the Ensembl dataset.

### Chromosomal distribution of heart-expressed genes

The number of genes on each chromosome was calculated for the HXP and HU2 dataset. The two distributions were compared using the Chi-square test. The numbers of detected gene clusters were compared to a randomly built dataset based on the genes in the HU2 dataset (100,000 permutations). Furthermore, we used the hypergeometrical distribution to assess whether genes located close to each other share similar expression levels ($cor \geq 0.65$) more often than genes located further apart.

### Identification of regulatory transcription factors

The identification of putative regulatory transcription factors binding sites was done using the TRANSFAC database. To reduce the rate of false positives among putative binding sites, we first filtered our data by searching for conserved, non-coding upstream regions of orthologous genes in human and mouse as annotated in the *Comparative Regulatory Genomics database* CORG [23]. CORG is based on the assumption that high levels of sequence conservation in non-coding upstream regions of orthologous genes are likely to reflect common regulatory elements.

### *Identification of similar gene expression levels*

To determine similarity in gene expression for pairs of genes, we calculated the Pearson correlation coefficient as well as the Euclidian distance of their expression values across all 55 previously analyzed tissues [22]. To assess the potential relationship of gene localization and similar expression, we used distance-correlation plots and correlation matrices. Correlation matrices give a rough overview of similar expression levels of genes located on the same chromosome by displaying color-coded correlation coefficients (see for example web supplement Fig. 1S).

## Authors' contributions

JHV acquired the data and performed the main analysis of data. AvH has made substantive intellectual contribution to concept and interpretation of data. AP participated in analysis of data and drafting of article. SS conceived the project, managed and participated in analysis and interpretation of data.

## Additional material

### Additional File 1

***Transcription factors shared by gene clusters****. Represented are clustered heart-expressed genes with their HUGO gene names, Ensembl gene IDs, strand orientations, transcriptional start site distances and transcription factors for which binding sites could be predicted. The shared transcription factors are indicated in italic.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-230-S1.pdf]

### Additional File 2

***Gene Ontology terms shared by gene clusters****. Shown are clustered heart-expressed genes annotated with Ensembl gene IDs, HUGO gene names and shared Gene Ontology (GO) terms of layer 4 and 5. The GO term IDs and descriptions are given.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-230-S2.pdf]

### Additional File 3

***Correlation matrix of human chromosome 10****. The matrix of Pearson correlation coefficients between the expression profiles of heart-expressed genes is shown color-coded, with genes being arranged according to their order on the chromosome. Gene pairs with similar expression levels are depicted in blue, anti-correlated pairs are shown in red.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-230-S3.pdf]

## Acknowledgements

We thank Raffaello Galli and Christoph Dieterich for technical assistance, and Sarah Teichmann for helpful discussions. This work was supported by a grant from The Max-Planck-Society for the Advancement of Science.

## References

1. van Driel R, Fransz PF, Verschure PJ: **The eukaryotic genome: a system regulated at different hierarchical levels.** *J Cell Sci* 2003, **116(Pt 20):**4067-4075.
2. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5(4):**299-310.
3. Hershberg R, Yeger-Lotem E, Margalit H: **Chromosomal organization is shaped by the transcription regulatory network.** *Trends Genet* 2005, **21(3):**138-142.
4. Grewal SI, Moazed D: **Heterochromatin and epigenetic control of gene expression.** *Science* 2003, **301(5634):**798-802.
5. Brown KE: **Chromatin folding and gene expression: new tools to reveal the spatial organization of genes.** *Chromosome Res* 2003, **11(5):**423-433.
6. Stein GS, Lian JB, Montecino M, Stein JL, van Wijnen AJ, Javed A, Pratap J, Choi J, Zaidi SK, Gutierrez S, Harrington K, Shen J, Young D, Pockwinse S: **Nuclear microenvironments support physiological control of gene expression.** *Chromosome Res* 2003, **11(5):**527-536.
7. Lawrence JG: **Shared strategies in gene organization among prokaryotes and eukaryotes.** *Cell* 2002, **110(4):**407-413.
8. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2(1):**65-73.
9. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26(2):**183-186.
10. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.** *Nature* 2002, **418(6901):**975-979.
11. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome.** *Nature* 2002, **420(6916):**666-669.
12. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol* 2002, **1(1):**5.
13. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291(5507):**1289-1292.
14. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31(2):**180-183.
15. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16):**6062-6067.
16. Schubeler D, Francastel C, Cimbora DM, Reik A, Martin DI, Groudine M: **Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus.** *Genes Dev* 2000, **14(8):**940-950.
17. Yamashita M, Ukai-Tadenuma M, Kimura M, Omori M, Inami M, Taniguchi M, Nakayama T: **Identification of a conserved GATA3 response element upstream proximal from the interleukin-13 gene locus.** *J Biol Chem* 2002, **277(44):**42399-42408.
18. Choudhary SK, Wykes SM, Kramer JA, Mohamed AN, Koppitch F, Nelson JE, Krawetz SA: **A haploid expressed gene cluster exists as a single chromatin domain in human sperm.** *J Biol Chem* 1995, **270(15):**8755-8762.
19. Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY: **Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster.** *Nucleic Acids Res* 2005, **33(5):**1435-1444.
20. Lawson GM, Knoll BJ, March CJ, Woo SL, Tsai MJ, O'Malley BW: **Definition of 5' and 3' structural boundaries of the chromatin domain containing the ovalbumin multigene family.** *J Biol Chem* 1982, **257(3):**1501-1507.

21. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14(6):**1085-1094.
22. Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S: **Genome-wide array analysis of normal and malformed human hearts.** *Circulation* 2003, **107(19):**2467-2474.
23. Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M: **CORG: a database for Comparative Regulatory Genomics.** *Nucleic Acids Res* 2003, **31(1):**55-57.
24. DeKoter RP, Singh H: **Regulation of B lymphocyte and macrophage development by graded expression of PU.1.** *Science* 2000, **288(5470):**1439-1441.
25. Lundgren M, Chow CM, Sabbattini P, Georgiou A, Minaee S, Dillon N: **Transcription factor dosage affects changes in higher order chromatin structure associated with activation of a heterochromatic gene.** *Cell* 2000, **103(5):**733-743.
26. Wallin JJ, Gackstetter ER, Koshland ME: **Dependence of BSAP represser and activator functions on BSAP concentration.** *Science* 1998, **279(5358):**1961-1964.
27. Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, d'Alessi F, Tiso N, Pallavicini A, Toppo S, Cannata N, Valle G, Lanfranchi G, Danieli GA: **A comprehensive, high-resolution genomic transcript map of human skeletal muscle.** *Genome Res* 1998, **8(8):**817-825.
28. Gabrielsson BL, Carlsson B, Carlsson LM: **Partial genome scale analysis of gene expression in human adipose tissue using DNA array.** *Obes Res* 2000, **8(5):**374-384.
29. Barrans JD, Ip J, Lam CW, Hwang IL, Dzau VJ, Liew CC: **Chromosomal distribution of the human cardiovascular transcriptome.** *Genomics* 2003, **81(5):**519-524.
30. Dempsey AA, Pabalan N, Tang HC, Liew CC: **Organization of human cardiovascular-expressed genes on chromosomes 21 and 22.** *J Mol Cell Cardiol* 2001, **33(3):**587-591.
31. Kluger Y, Yu H, Qian J, Gerstein M: **Relationship between gene co-expression and probe localization on microarray slides.** *BMC Genomics* 2003, **4(1):**49.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

## 2.2 DNA-binding and epigenetic transcription factors relevant for cardiac development and disease

The heterogeneity of congenital heart disease (CHD) associated with single-gene defects in patients as well as the broad phenotype spectrum seen in mouse models point to a complex genetic network with modifier genes, genetic polymorphisms and the influence of environmental factors (Srivastava, 2001; Solloway and Harvey, 2003; Olson, 2004). In the following, three potential key nodes of the cardiac transcription network (CITED2, TBX20 and DPF3) have been further investigated to define their impact on the development of CHD in human.

### 2.2.1 Mutations of the transcription factor CITED2 lead to congenital heart defects

> **Sperling S**, Grimm CH, Dunkel I, Mebus S, Sperling HP, Ebner A, Galli R, Lehrach H, Fusch C, Berger F, Hammer S. Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Hum Mutat* 2005;**26**:575-582.

CITED2 is a ubiquitously expressed hypoxia-inducible transcriptional cofactor and interacts with CREBBP and EP300 with high affinity . (Bhattacharya et al., 1999; Leung et al., 1999; Freedman et al., 2003). This binding competitively inhibits the interaction between EP300 and the transcription factor HIF1A blocking transcriptional activation by HIF1A. Further, CITED2 coactivates TFAP2 isoforms and both can be detected together at the *Pitx2c* promoter in embryonic mouse hearts, suggesting a role for CITED2 in left-right patterning through the Nodal-PITX2C pathway (Braganca et al., 2003; Bamforth et al., 2004; Weninger et al., 2005). CITED2 is essential for normal development of the heart, as mice lacking *Cited2* die in utero showing various cardiac malformations including atrial and ventricular septal defects, right-sided aortic arches, double-outlet right ventricle, common arterial trunk and overriding aorta (Bamforth et al., 2001; Bamforth et al., 2004; Weninger et al., 2005). Deficiencies in TFAP2 coactivation have been suggested to cause laterality defects in *Cited2-/-* mice, but also dysregulation of hypoxia-activated gene transcription may account for the cardiac malformations seen in *Cited2*-/- embryos (Yin et al., 2002).

In the presented report, 392 patients reflecting a broad range of CHD were analyzed and seven potential diseases causing mutations in eight patients discovered. These mutations that were exclusively observed in the patient cohort and not found in 192 control individuals give raise to cardiac septal defects as well as outflow tract abnormalities associated with malrotation of the great arteries (Table 1). This reflects the range of defects observed in *Cited2*-/- embryos.

| Nucleotide Variation | Amino Acid Variation | # Patients | Type of Congenital Heart Defect |
|---|---|---|---|
| c.-91G>A | | 2 | Tetralogy of Fallot<br>Perimembranous VSD and secundum ASD |
| c.-81T>C | | 1 | Situs inversus totalis, transposition of the great arteries and perimembranous VSD |
| c.456C>T | p.His52His | 1 | Perimembranous VSD and right ventricular outflow tract obstruction |
| c.508_534del27 | p.Ser170_Gly178del | 1 | Perimembranous VSD |
| c.534_535ins27 | p.Gly178_Ser179ins9 | 1 | Secundum ASD |
| c.592_597delAGCGGC | p.Ser198_Gly199del | 1 | Sinus venosus ASD, abnormal pulmonary venous return to the right atria |
| c.1268A>G | | 1 | Tetralogy of Fallot |

**Table 1. *CITED2* mutations identified among 392 patients with congenital heart disease.**

Three of these *CITED2* mutations alter the amino acid sequence (p.Ser170_Gly178del, p.Gly178_Ser179ins9, p.Ser198_Gly199del) and cluster in the serine-glycine rich junction of the protein, which therefore represents a potential hotspot for mutations in CITED2. The further analysis of these mutations using reporter-gene assays revealed their functional implications, such as that all three mutations lead to a significant loss in HIF1A transcriptional repressive capacity of CITED2. Moreover, the TFAP2C coactivation of the p.Ser170_Gly178del mutant was significantly diminished. These findings indicate a modifying role for the serine-glycine rich junction in CITED2 function to which no functionality had been assigned so far. One might speculate that variations in this region cause conformational changes, altering the ability of the EP300 binding domain to interact with CREBBP and EP300 or to recruit other cofactors.

This suggests that the detected CITED2 mutations are potential risk factors for CHD and account for ~2% (8/392) of the studied patient cohort of sporadic CHD cases. The broad phenotypical spectrum of heart defects seen in *Cited2*-/- mice as well as in the patients points to potentially further relevant, but currently unknown modifying factors. In the future it will be of interest to evaluate the functionality of the serine-glycine-rich junction of CITED2 and the non amino acid altering mutations observed in the coding region, the 5´ UTR and 3´UTR.

To gain insights in the inheritance of the mutations, it should be envisaged to analyze family triplets, which unfortunately have not been available for the present study.

Finally, comparing patients and controls no significant differences in allele frequencies of the common variants were observed, which is in accordance with a previous study (Volcik et al., 2004). Haplotype analysis showed that only 3 out of 32 possible haplotypes accounted for at least 98.6% of the investigated chromosomes, which suggests the existence of only one haplotype block. Therefore two htSNPs could be extracted that are sufficient for haplotype determination.

In summary, this is the first evidence that CITED2 is a disease causing gene for congenital heart malformations in human, in particular for septal defects and malrotations of the great arteries.

**HUMAN MUTATION 26(6), 575–582, 2005**

*RESEARCH ARTICLE*

# Identification and Functional Analysis of *CITED2* Mutations in Patients With Congenital Heart Defects

**Silke Sperling,[1]* Christina H. Grimm,[1] Ilona Dunkel,[1] Siegrun Mebus,[2] Hans-Peter Sperling,[2] Arno Ebner,[3] Raffaello Galli,[1] Hans Lehrach,[1] Christoph Fusch,[3] Felix Berger,[2] and Stefanie Hammer[1]**

*[1]Max Planck Institute for Molecular Genetics, Berlin, Germany; [2]German Heart Center Berlin, Berlin, Germany; [3]Department of Pediatrics, Ernst-Moritz-Arndt University, Greifswald, Germany*

*Communicated by Mark H. Paalman*

**Recent reports have demonstrated that mice lacking the transcription factor Cited2 die in utero showing various cardiac malformations. We present for the first time functionally relevant mutations of *CITED2* in patients with congenital heart defects (CHDs). *CITED2* encodes a CREBBP/EP300 interacting transcriptional modulator of HIF1A and TFAP2. To study the potential impact of sequence variations in *CITED2* for CHDs in humans, we screened a cohort of 392 well-characterized patients and 192 control individuals using DHPLC, sequencing, and Amplifluor$^{TM}$ genotyping techniques. We identified 15 *CITED2* nucleotide alterations. Seven of these alterations were found only in CHD patients and were not detected in controls, including three mutations leading to alterations of the amino acid sequence (p.Ser170_Gly178del, p.Gly178_Ser179ins9, and p.Ser198_Gly199del). All three of these amino acid changing mutations cluster in the serine-glycine-rich junction of the protein, to which no functionality had heretofore been assigned. Here we show that these mutations significantly reduce the capacity of CITED2 to transrepress HIF1A, and that the p.Ser170_Gly178del mutation significantly diminishes TFAP2C coactivation. This reveals a modifying role for the serine-glycine-rich region in CITED2 function. In summary, the observation of these mutations in patients with septal defects indicates that *CITED2* has a causative impact in the development of CHD in humans. Hum Mutat 26(6), 575–582, 2005. © 2005 Wiley-Liss, Inc.**

KEY WORDS: CITED2; congenital heart defects; CHD; septal defects; arterial malrotation

## INTRODUCTION

Congenital heart defects (CHDs) account for the largest number of birth defects in humans, with an incidence of about eight per 1,000 live births and are the leading noninfectious cause of mortality in newborns. Although major insights into the cardiac developmental process have been gained in studies of animal models, such as mice, chicken, and zebrafish, little is known about the genetic basis in humans. The overwhelming majority of congenital heart malformations do not segregate in Mendelian ratios, although they show familial aggregation, which suggests that genetic factors play a role in their development. Almost 30% of major cardiac malformations are associated with additional developmental abnormalities and result from a recognized chromosomal anomaly or occur as part of a syndrome. Primarily based on knowledge gained from model organisms, disease genes have been identified in a few syndromes, familial nonsyndromic conditions, and very few sporadic cases. The identified disease genes point to a key role of transcription factors in the process of cardiac maldevelopment. It has been demonstrated that TBX5 (MIM# 601620) mutations are frequent causes of Holt-Oram syndrome (MIM# 142900) [Basson et al., 1997; Li et al., 1997] and mutations in the transcriptional coactivators CREBBP (MIM# 600140) and EP300 (MIM# 602700) are associated with cardiac malformations in Rubinstein-Taybi syndrome (MIM# 180849) [Petrij et al., 1995; Roelfsema et al., 2005]. Moreover, causative gene defects have been described for nonsyndromic congenital heart malformations, e.g., mutations in the cardiac transcription factors NKX2.5 (MIM# 600584) and GATA4 (MIM# 600576) [Schott et al., 1998; Garg et al., 2003; Pizzuti et al., 2003; Reamon-Buettner and Borlak, 2004; Ware et al., 2004; Ching et al., 2005]. The heterogeneity of CHDs associated with single-gene defects in patients, and the broad phenotype spectrum seen in mouse models point to a complex genetic network with modifier genes, genetic polymorphisms, and the influence of environmental factors [Bamford et al., 2000; Srivastava, 2001; Solloway and Harvey, 2003; Olson, 2004]. For example, human mutations in the cardiac homeobox protein NKX2.5 cause a diverse set of congenital heart malformations that include septal defects, cardiomyopathy, outflow tract defects, hypoplastic left heart, and associated arrhythmias. Here we present an analysis of the *CITED2* (MIM# 602937) gene in a patient cohort representing

Received 13 April 2005; accepted revised manuscript 16 August 2005.

*Correspondence to: Silke Sperling, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany.
E-mail: sperling@molgen.mpg.de

Grant sponsor: Bioprofile; Grant numbers: BMBF; Bioprofile 0313040B; Grant sponsor: Max Planck Society (Tandem Project).
DOI 10.1002/humu.20262

Published online in Wiley InterScience (www.interscience.wiley.com).

© 2005 WILEY-LISS, INC.

a broad phenotype spectrum, and show for the first time its impact as a disease gene for CHDs in humans.

CITED2 is an ubiquitously expressed hypoxia-inducible transcriptional cofactor and interacts with high affinity with CREBBP and EP300 [Bhattacharya et al., 1999; Leung et al., 1999; Freedman et al., 2003]. This binding competitively inhibits the interaction between EP300 and the transcription factor HIF1A (MIM# 603348), blocking transcriptional activation by HIF1A. Further, CITED2 coactivates TFAP2 isoforms (MIM#s 107580, 601601, 601602) and both can be detected together at the *Pitx2c* (MIM# 601542) promoter in embryonic mouse hearts. This suggests that CITED2 plays a role in left–right patterning through the Nodal-PITX2C pathway [Braganca et al., 2003; Bamforth et al., 2004; Weninger et al., 2005]. CITED2 is essential for normal development of the heart, as evidenced by the fact that mice lacking *Cited2* die in utero showing various cardiac malformations including atrial and ventricular septal defects, right-sided aortic arches, double-outlet right ventricle, common arterial trunk and overriding aorta [Bamforth et al., 2001, 2004; Weninger et al., 2005]. Deficiencies in TFAP2 coactivation have been suggested to cause laterality defects in *Cited2*−/− mice, but also dysregulation of hypoxia-activated gene transcription may account for the cardiac malformations seen in *Cited2*−/− embryos [Yin et al., 2002].

In the present study we performed a mutation screen of the *CITED2* gene in a cohort of patients with well-characterized phenotypes of sporadic nonsyndromic CHD [Kaynak et al., 2003]. Novel *CITED2* mutations were identified and their functional significance was investigated by transactivation and subcellular-localization assays. The results indicate that *CITED2* has a causative impact on the development of CHD in humans.

## MATERIALS AND METHODS

### Patient Samples

Patient blood samples were obtained from the German Heart Center after ethics approval was granted by the institutional review committee and informed consent was obtained from the patients or their parents. Phenotypic information was documented in detail in a cardiovascular genetic database established at the Max-Planck-Institute for Molecular Genetics [Seelow et al., 2004]. Genomic DNA was prepared from the blood samples by standard procedures, and DNA was purified when necessary using DNA Cleanup (Qiagen, Hilden, Germany; www.qiagen.com). Control DNA samples were obtained from the Community and Molecular Medicine Newborn Survey (University Hospital Greifswald, Greifswald, Germany).

### Mutation Detection

Both exons of the human *CITED2* gene, including the entire 5′ and 3′ untranslated regions (GenBank: NM_006079.3), were amplified by PCR using the primers described in Table 1A. All PCR reactions were performed using 50 ng of genomic DNA, 200 μM of dNTPs and 500 nM of primer. PCR products were denatured for 10 min at 95°C and subjected to denaturing HPLC (DHPLC) analysis on the automated WAVE$^{TM}$ nucleic acid fragment analysis system (Transgenomic, San Jose, CA; www.transgenomic.com) as described previously [Eng et al., 2001]. The fragments were eluted with temperatures calculated by the DHPLC melt program for the successful resolution of heteroduplexes (http://insertion.stanford.edu/melt.html) [Jones et al., 1999]. Samples with double- or triple-peaked DHPLC chromatograms were purified using Qiagen PCR purification and sequenced by the Services in Molecular Biology Company (Berlin, Germany).

### Amplifluor$^{TM}$ Allele-Specific PCR

The Amplifluor$^{TM}$ genotyping assay based on PCR amplification in the presence of tailed allele-specific primers, a common reverse primer, and universal fluorescence labeled Amplifluor$^{TM}$ primers (Serologicals, Temecula, CA; www.serologicals.com) was performed as described previously [Myakishev et al., 2001; Rickert et al., 2004]. The primers designed by the Amplifluor$^{TM}$ assay architect software (www.assayarchitect.com) are listed in Table 1B. The 5-μl amplification reactions contained 25 nM of FAM- and JOE-labeled Amplifluor$^{TM}$ primers, 25 nM of tailed allele-specific primers, 375 nM of reverse primer, 1 × reaction buffer (Serologicals), 0.2 mM of dNTPs, 0.25 U of HotStar Taq Polymerase (Qiagen), and 20 ng of genomic DNA. The amplification profiles were as follows: 96°C, 10 min; (95°C, 30 sec; 56°C, 30 sec; 72°C, 40 sec) × 45 cycles; 72°C, 3 min. The amplification signals were analyzed via endpoint measurement using the ABI Prism 7900HT system (Applied Biosystems, Darmstadt, Germany; www.appliedbiosystems.com).

### Haplotype Analysis

Haplotype structure was determined using HAPLOVIEW [Barrett et al., 2005]. Linkage disequilibrium (LD) was calculated as D′ values using an expectation-maximization (EM) algorithm, and haplotype frequencies comparing the patient and the control cohort were analyzed by $\chi^2$ tests.

### Plasmids

The open reading frames of wild-type (wt) and mutant *CITED2* were amplified by PCR from genomic DNA and cloned into pcDNA3.1(+) (Invitrogen, Karlsruhe, Germany; www.invitrogen.com) to obtain expression vectors for CITED2-wt and the CITED2 mutants. The resulting clones were verified by sequencing. To create N- and C-terminal GFP-CITED2 fusion proteins, the open reading frames were amplified by PCR from the plasmid DNA and cloned in frame into pEGFP-N1 and pEGFP-C1 (BD Biosciences, Palo Alto, CA). The HIF1A reporter system pGal4-HIF1A and pGal4-Luc was described previously and kindly provided by L.E. Huang (NCI, Bethesda, MD) [Huang et al., 1998]. The pGal4-HIF1A plasmid contains the Gal4-DNA binding domain fused to the C-terminal transactivation domain of HIF1A, which is stable under nonhypoxic conditions. In pGal4-Luc the Luciferase reporter gene is under the control of Gal4-DNA binding sites. The TFAP2 responsive reporter plasmid pAP2-Blue harboring a Luciferase reporter gene under the control of three TFAP2 response elements, and the human TFAP2C expression plasmid pRSV-TFAP2C were a kind gift from Helen Hurst (Hammersmith Hospital, London, UK) [Bosher et al., 1996; Bamforth et al., 2001].

### Transcriptional Assay

HepG2 cells were maintained in Dulbecco's modified Eagle's medium with 10% fetal bovine serum. Cells were seeded into 96-well plates and on the next day at 60% confluency were transfected using Fugene6 (Roche, Mannheim, Germany; www.roche-applied-science.com) according to the manufacturer's instructions. Wt and mutant *CITED2* constructs or empty vector were cotransfected together with pGal4-HIF1A and pGal4-Luc or pRSV-TFAP2C and pAP2-Blue, respectively. In all wells a pRL-TK Luciferase vector (Promega, Mannheim, Germany; www.prome

TABLE 1. Primers Used for CITED2 Amplification and Allele-Specific PCR

| Name | Sequence |
|---|---|
| A: Primers for amplification of *CITED2* exons | |
| exon1_F1 | GCTCATTGTTGGCAGCTGC |
| exon1_R1 | TTCGCCTCACGCTCTTCCTC |
| exon2_F1 | ATCTGCCCTTTTCACTTCCAG |
| exon2_R1 | GGAGTTGTTAAACCTGGCCG |
| exon2_F2 | TGTGAACGGAGGGCACCCC |
| exon2_R2 | CGAGCTGCTGCCAGAGCCG |
| exon2_F3 | ACCAGATGAACGGGACAAAC |
| exon2_R3 | CGGTCCAAACCCATTTCTAT |
| exon2_F4 | GCCCAATGTCATAGACACTG |
| exon2_R4 | ATTCACGCCGAAGAAGTTG |
| exon2_F5 | GGCGAAAGAAATCAAACCC |
| exon2_R5 | AATGTCAAGGCTACAAAAACGA |
| exon2_F6 | CTGCCACTTTTTTTCCTGTTT |
| exon2_R6 | AAAATGAAGCGAGATGGCAGT |
| exon2_F7 | TAGTTGGTTGCATGAACTTC |
| exon2_R7 | AACTATTAGCACAGTGTCAAA |
| exon2_F8 | GTCAGTGGCAAACATTTCACAGA |
| exon2_R8 | TGTTCAACTCAAAGACGGGG |
| B: Primers for Amplifluor$^{\text{TM}}$ allele-specific PCR | |
| c.–91_A_green_r | GAAGGTGACCAAGTTCATGCTTTCAGCAGCACATAGAGGGGAT |
| c.–91_G_red_r | GAAGGTCGGAGTCAACGGATTAGCAGCACATAGAGGGGAC |
| c.–91_com_f | CGCTTTGCACGCCAGGAA |
| c.–81_T_green_r | GAAGGTGACCAAGTTCATGCTTGACCGGCTCAGCAGCACA |
| c.–81_C_red_r | GAAGGTCGGAGTCAACGGATTACCGGCTCAGCAGCACG |
| c.–81_com_f | CGCTTTGCACGCCAGGAA |
| c.115_117_CAC_green_r | GAAGGTGACCAAGTTCATGCTTCTGCTGCTGCTGGTGGT |
| c.115_117_delCAC_red_r | GAAGGTCGGAGTCAACGGATTGCTGCTGCTGCTGGTGAT |
| c.115_117_com_f | ATGGGCATGGGGCAGTT |
| c.1040_T_green_f | GAAGGTGACCAAGTTCATGCTTCCTTGACATTCACCCACCTCT |
| c.1040_C_red_f | GAAGGTCGGAGTCAACGGATTCTTGACATTCACCCACCTCC |
| c.1040_com_r | CAACGAAAAAGACCAAGTTAGCTA |
| c.1268_G_green_r | GAAGGTGACCAAGTTCATGCTAAGCGAGATGGCAGTTTGC |
| c.1268_A_red_r | GAAGGTCGGAGTCAACGGATTTGAAGCGAGATGGCAGTTTGT |
| c.1268_com_f | GGAAAAATTGCATTAGTTGGTTGCAT |
| rs1131400_C_green_r | GAAGGTGACCAAGTTCATGCTAAGCGCCCGTGGTTCATG |
| rs1131400_A_red_r | GAAGGTCGGAGTCAACGGATTAAGCGCCCGTGGTTCATT |
| rs1131400_com_f | GACTGGAAATGGCAGACCATAT |
| rs1131431_C_green_r | GAAGGTGACCAAGTTCATGCTGTGCAGTAATATCTGCCCTTCG |
| rs1131431_T_red_r | GAAGGTCGGAGTCAACGGATTGTGCAGTAATATCTGCCCTTCAA |
| rs1131431_com_f | GGAAAAATTGCATTAGTTGGTTGCAT |
| rs1804687_C_green_f | GAAGGTGACCAAGTTCATGCTCCGGTCCTGGACGCGACCA |
| rs1804687_G_red_f | GAAGGTCGGAGTCAACGGATTCCGGTCCTGGACGCGACGA |
| rs1804687_com_r | CTCGGAGGACTGGGCTGGCAA |
| rs2001409_T_green_f | GAAGGTGACCAAGTTCATGCTTCCTCGGTCTTCGGAGCAGAAT |
| rs2001409_A_red_f | GAAGGTCGGAGTCAACGGATTCCTCGGTCTTCGGAGCAGAAA |
| rs2001409_com_r | AAGAGCCCCAGCCAGCTT |
| rs4177_TTT_green_f | GAAGGTGACCAAGTTCATGCTGTCAGTGGCAAACATTTCACAGATTT |
| rs4177_delTTT_red_f | GAAGGTCGGAGTCAACGGATTGTCAGTGGCAAACATTTCACAGATTA |
| rs4177_com_r | ACAGTGTCAAAAATGTTGAAGACAGA |

ga.com) was cotransfected to control for transfection efficiency. Subsequently, 40 hr after transfection the cells were washed with phosphate-buffered saline (PBS) and lysed in 50 µl of passive lysis buffer (Promega). Firefly and *Renilla* Luciferase activities were measured using the Dual-Luciferase-Reporter Assay System (Promega) in a Centro LB960 luminometer (Berthold, Bad Wildbad, Germany; www.berthold.com). Firefly Luciferase activities were normalized to *Renilla* Luciferase activity and the results for the samples transfected with CITED2-wt construct were set to 100%. The results shown represent a minimum of three independent experiments performed in at least triplicates.

### Immunofluorescence and Subcellular Localization

HepG2 and HEK293 cells were seeded onto glass coverslips 6 hr prior to transfection at about 60% confluency. GFP-CITED2 expression constructs containing wt and mutant *CITED2* were transfected using Fugene6 (Roche) according to the manufacturer's instructions. Then 48 hr after transfection, the cells were fixed in 4% paraformaldehyde/PBS at RT for 15 min, washed in PBS, and mounted with Vectashield containing DAPI (Vector Laboratories, Inc., Burlingame, CA; www.vectorlabs.com). The cells were analyzed by fluorescence microscopy.

## RESULTS

### Phenotypes of the Analyzed CHD Patient Cohort

*CITED2* mutation analysis was performed using DHPLC analysis in a cohort of 392 unrelated nonsyndromic patients who showed a broad spectrum of CHDs. The cardiac phenotypes of the analyzed patients are described in Table 2. To allow a more detailed visualization of the panel of analyzed CHD phenotypes in the overall patient cohort, as well as their association with the observed genotypes, we have set up a freely accessible interactive Web supplement using the database front-end d-matrix applied to our cardiovascular genetics database (http://dmatrix.molgen.mpg.de/SV) [Seelow et al., 2004]. A further 192 individuals from the Greifswald Newborn Survey served as controls.

### Identification of *CITED2* Mutations

From a total of 392 patient samples, we identified five sequence variations that were already listed in dbSNP (NCBI), as well as 10 novel *CITED2* nucleotide alterations (three amino acid deletions, one amino acid insertion, one amino acid substitution, one silent nucleotide alteration, two nucleotide substitutions in the 5′UTR, and two alterations in the 3′UTR). The localization of these sequence variations and the predicted effects on the CITED2 amino acid sequence, as well as the frequencies detected in CHD patients and controls, are shown in Table 3. Seven of the novel sequence variations were detected only in CHD patients and not in controls, namely c.−91G>A, and c.−81T>C in the untranslated 5′ region; c.456C>T, c.508_534del27 (p.Ser170_Gly178del), c.534_535ins27 (p.Gly178_Ser179ins9), and c.592_597delAGCGGC (p.Ser198_Gly199del) in the coding region; and c.1268A>G in the untranslated 3′ region (nomenclature based on GenBank NM_006079.3, *CITED2* cDNA, with +1 corresponding to A of the initiation codon; www.hgvs.org/mutnomen/). All patient-exclusive mutations that alter the amino acid sequence of CITED2 (p.Ser170_Gly178del, p.Gly178_Ser179ins9, and p.Ser198_Gly199del) cluster in the serine-glycine-rich junction [Leung et al., 1999] of the protein (Fig. 1).

TABLE 2. Congenital Heart Defects of Analyzed Patients

| Cardiovascular anomalies | Patients genotyped (N = 392) |
|---|---|
| Situs inversus totalis | 5 (1.3%) |
| Dextrocardia | 2 (0.5%) |
| D-transposition of the great arteries | 21 (5.4%) |
| Right aortic arch | 30 (7.7%) |
| Secundum atrial septal defect | 129 (32.9%) |
| Sinus venosus atrial septal defect | 26 (6.6%) |
| Partial anomalous pulmonary venous return | 11 (2.8%) |
| Perimembranous ventricular septal defect | 180 (45.9%) |
| Incomplete atrioventricular septal defect | 3 (0.8%) |
| Complete atrioventricular septal defect | 16 (4.1%) |
| Tetralogy of Fallot | 46 (11.7%) |
| Hypoplastic left heart syndrome | 2 (0.5%) |
| Pulmonary atresia | 14 (3.6%) |
| Pulmonary stenosis | 56 (14.3%) |
| Double inlet left ventricle | 1 (0.3%) |
| Double outlet right ventricle | 13 (3.3%) |
| Left superior vena cava | 29 (7.4%) |
| Aortic isthmus stenosis | 19 (4.8%) |

The phenotype characteristics of patients with potential disease-causing mutations are shown in Table 4. We observed amino acid altering mutations in one patient with a sinus venosus atrial septal defect and abnormal pulmonary venous return to the right atria, one patient with an atrial septal defect of the secundum type (ASDII), and one patient with a perimembranous ventricular septal defect (VSD). Furthermore, we found non-amino acid altering mutations in two patients with tetralogy of Fallot; one patient with situs inversus totalis, transposition of the great arteries, and perimembranous VSD; one patient with perimembranous VSD and ASDII; and one patient with perimembranous VSD and right ventricular outflow tract stenosis.

### Influence of CITED2 Mutations on TFAP2 Transactivation and HIF1A Transrepression

To assess the functional significance of the amino acid altering CITED2 mutations found in CHD patients, we tested their influence on reporter gene transactivation and repression. To date, no functionality has been assigned to the serine-glycine-rich junction of the protein harboring the observed mutations. Previous reports showed that CITED2 acts as a binding partner and transcriptional coactivator of TFAP2 [Bamforth et al., 2001], and interacts with the histone acetylases CREBBP/EP300 via its C-terminus. The latter leads to a transcriptional repression of HIF1A due to overlapping binding sites [Bhattacharya et al., 1999]. In our experiments CITED2-wt coactivated the TFAP2C-mediated stimulation of a TFAP2-reporter construct, as described previously [Bamforth et al., 2001]. However, the p.Ser170_Gly178del mutant

TABLE 3. Localization and Frequencies of *CITED2* Variants

| | | | | Patients | | | Controls | | |
|---|---|---|---|---|---|---|---|---|---|
| dbSNP | Position | Nucleotide variation$^{a}$ | Amino acid variation | Mut chr | Total chr | Mut allele freq | Mut chr | Total chr | Mut allele freq |
| | **Exon1 (5′UTR)** | **c.−91G>A**$^{b}$ | | **2** | **332** | **0.0060** | **0** | **368** | **0.0000** |
| | **Exon1 (5′UTR)** | **c.−81T>C**$^{b}$ | | **1** | **362** | **0.0028** | **0** | **368** | **0.0000** |
| rs1804687 | Exon1 (5′UTR) | c.−52G>C | | 15 | 352 | 0.0426 | 16 | 332 | 0.0482 |
| rs2001409 | Exon1 (5′UTR) | c.−24A>T | | 65 | 358 | 0.1816 | 65 | 356 | 0.1826 |
| rs1131400 | Exon2 | c.21C>A | p.Ala7Ala | 75 | 372 | 0.2016 | 67 | 382 | 0.1754 |
| | **Exon2** | **c.115_117delCAC** | **p.His39del** | **2** | **728** | **0.0027** | **2** | **388** | **0.0052** |
| | **Exon2** | **c.456C>T**$^{b}$ | **p.His52His** | **1** | **686** | **0.0015** | **0** | **382** | **0.0000** |
| | **Exon2** | **c.479A>T** | **p.His160Leu** | **1** | **702** | **0.0014** | **1** | **382** | **0.0026** |
| | **Exon2** | **c.508_534del27**$^{b}$ | **p.Ser170_Gly178del**$^{c}$ | **1** | **702** | **0.0014** | **0** | **382** | **0.0000** |
| | **Exon2** | **c.534_535ins27**$^{b}$ | **p.Gly178_Ser179ins9**$^{c}$ | **1** | **702** | **0.0014** | **0** | **382** | **0.0000** |
| | **Exon2** | **c.592_597delAGCGGC**$^{b}$ | **p.Ser198_Gly199del**$^{c}$ | **1** | **648** | **0.0015** | **0** | **378** | **0.0000** |
| | **Exon2 (3′UTR)** | **c.1040C>T** | | **13** | **372** | **0.0349** | **10** | **374** | **0.0267** |
| rs1131431 | Exon2 (3′UTR) | c.1248C>T | | 55 | 374 | 0.1471 | 61 | 382 | 0.1597 |
| | **Exon2 (3′UTR)** | **c.1268A>G**$^{b}$ | | **1** | **372** | **0.0027** | **0** | **382** | **0.0000** |
| rs4177 | Exon2 (3′UTR) | c.1497_1499delTTT | | 55 | 350 | 0.1571 | 59 | 378 | 0.1561 |

[a] Systematic nomenclature for SNPs (www.hgvs.org) based on GenBank NM_006079.3 (*CITED2* cDNA) and counting +1 as A of the initiation codon. Novel sequence variations are in bold.

[b] Sequence variations not found in the control cohort.

[c] Patient-exclusive CITED2 protein mutations altering transcriptional properties. Mut, mutant; chr, chromosome; freq, frequency.

showed significantly reduced costimulation capacity compared to wt, and reached only half-maximal coactivation (Fig. 2A). All other CITED2 mutants coactivated TFAP2C to the same extent as the CITED2-wt. Next, we confirmed with a HIF1A responsive reporter system [Huang et al., 1998] that CITED2-wt is an efficient repressor of HIF1A transcriptional activation independently of hypoxia [Bhattacharya et al., 1999]. Again, the p.Ser170_Gly178del mutant displayed a significant loss of activity, as it was only able to repress HIF1A with about 60% efficiency compared to wt. Moreover, in the HIF1A reporter system, the mutations p.Ser198_Gly199del and p.Gly178_Ser179ins9 also significantly affected the activity of CITED2, revealing only about 75% repressive activity compared to wt. In contrast, the p.His39del mutation and the p.His160Leu amino acid substitution, which had also been found in controls, did not alter CITED2 mediated HIF1A repression significantly (Fig. 2B). These results show that deletions or insertions within the serine-glycine-rich junction modulate CITED2 signal transduction and point to a causative impact of CITED2 on the development of congenital heart diseases in human.

### Influence of CITED2 Mutations on Subcellular Localization

To further evaluate whether the functional changes are due to altered subcellular localization of the protein, transient transfections were carried out using N- and C-terminal GFP fusion constructs of mutant and wt CITED2, followed by fluorescence microscopy. CITED2-wt was detected mainly in the nucleus and to a lesser extent in the cytoplasm of HEK and HepG2 cells. However, none of the CITED2 mutations altered cellular localization or expression of the protein (Fig. 3 and data not shown). Thus, the diminished TFAP2 coactivation and HIF1A repression of the CITED2 mutants are not caused by an incorrect localization of the protein.

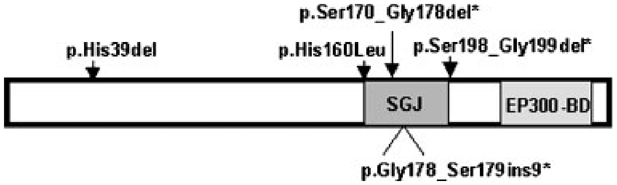


FIGURE 1. **Position of mutations in the CITED2 protein observed in CHD patients. The discovered mutations cluster mainly in the serine-glycine rich junction (SGJ; p.Ser161_Gly199), which therefore represents a hotspot for mutations in the protein. The position of the EP300 binding domain (EP300-BD; p.Asp224_Phe255) is indicated. Mutations found only in CHD patients and not in the control cohort are marked with*.**

### Haplotype Analysis of the *CITED2* Gene

Finally, to evaluate the polymorphisms observed in *CITED2*, we calculated the Hardy-Weinberg equilibrium using multiple tests and a two-sided significance level of 5%. None of the genotyped polymorphisms showed a significant deviation from the Hardy-Weinberg equilibrium. For the haplotype analysis there were 359 individuals available with successful genotyping for each of the six *CITED2* polymorphisms and a minor allele frequency of $>1\%$. The EM algorithm showed that three out of 32 possible haplotypes exceeded a frequency of 1%, and these accounted for 98.6% of the chromosomes within our samples, with the main haplotype G-A-C-C-C-TTT showing a frequency of 65.4% (Table 5). An allelic association between the different loci (represented as $D'$ values) is indicated in Figure 4. This analysis suggests the existence of one single haplotype block [Gabriel et al., 2002], and we identified rs1131400 and rs1131431 as haplotype tag (ht) SNPs, which are sufficient for determining the corresponding haplotype. However, none of the common haplotypes showed an association with CHD when patients and control individuals were compared (data not shown).

## DISCUSSION

Previous reports of mice lacking *Cited2* suggested that it plays a direct role in the development of the AV canal and cardiac septa, and that it is required for the normal establishment of the left–right axis. *Cited2*−/− embryos show a variety of cardiac malformations, including atrial and ventricular septal defects, abnormal heart looping with overriding aorta, and outflow tract abnormalities [Bamforth et al., 2001, 2004; Weninger et al., 2005]. In the present study we analyzed 392 patients with a broad range of CHDs, and discovered seven potential disease-causing mutations in eight patients. These mutations, which were exclusively observed in the patient cohort and not found in 192 control individuals, give rise to cardiac septal defects as well as outflow tract abnormalities associated with malrotation of the great arteries. This reflects the range of defects observed in *Cited2*−/− embryos.

Three of these *CITED2* mutations (p.Ser170_Gly178del, p.Gly178_Ser179ins9, and p.Ser198_Gly199del) alter the amino acid sequence and cluster in the serine-glycine rich junction of the protein, which therefore represents a potential hotspot for mutations in CITED2. Our further analysis of these mutations using reporter-gene assays revealed their functional implications

TABLE 4. ***CITED2* Mutations Identified among 392 Patients with Congenital Heart Disease**

| Nucleotide variation | Amino acid variation | # Patients | Type of congenital heart defect |
|---|---|---|---|
| c.–91G>A | | 2 | Tetralogy of Fallot<br>Perimembranous ventricular septal defect and secundum atrial septal defect |
| c.–81T>C | | 1 | Situs inversus totalis, transposition of the great arteries and perimembranous ventricular septal defect |
| c.456C>T | p.His52His | 1 | Perimembranous ventricular septal defect and right ventricular outflow tract obstruction |
| c.508_534del27 | p.Ser170_Gly178del | 1 | Perimembranous ventricular septal defect |
| c.534_535ins27 | p.Gly178_Ser179ins9 | 1 | Secundum atrial septal defect |
| c.592_597delAGCGGC | p.Ser198_Gly199del | 1 | Sinus venosus atrial septal defect, abnormal pulmonary venous return to the right atria |
| c.1268A>G | | 1 | Tetralogy of Fallot |

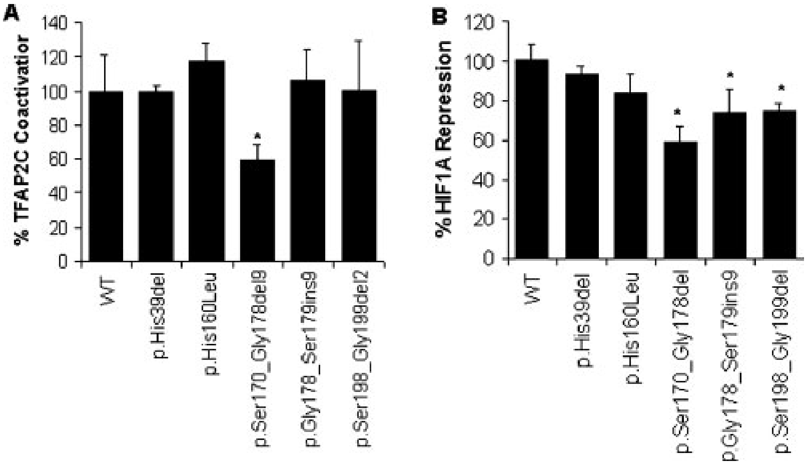


FIGURE 2. Transcriptional modulation of CITED2 variants. Luciferase activity of a TFAP2C-stimulated TFAP2-reporter construct (**A**) and a Gal4-HIF1A stimulated Gal4-reporter construct (**B**) cotransfected with CITED2-wt or mutant constructs as indicated. The specific CITED2 construct used is shown below each bar. Luciferase activities were measured and the mean fold-coactivation/repression as compared to wt is expressed as a percentage. Each bar represents a minimum of three independent experiments performed in at least triplicates (* significantly different from wt, $P < 0.05$).

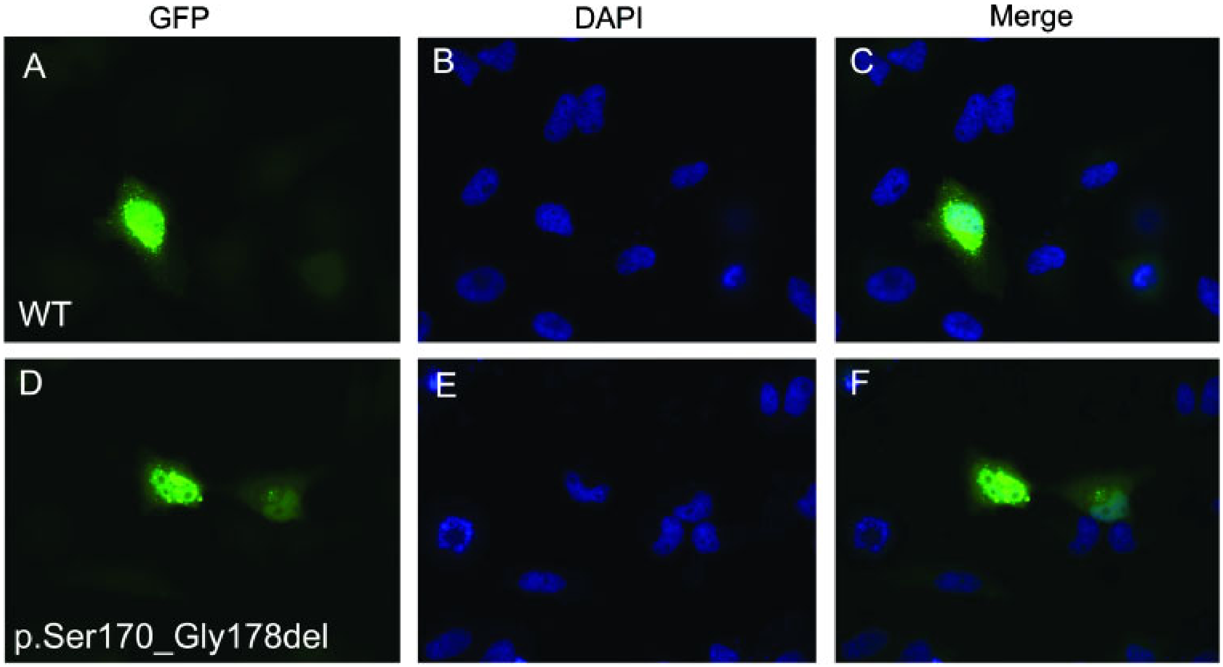


FIGURE 3. Subcellular localization of CITED2. For CITED2-wt and the p.Ser170_Gly178del mutant fluorescence of GFP-fusion proteins (**A** and **D**) and DAPI staining (**B** and **E**) are shown individually and merged (**C** and **F**).

TABLE 5. *CITED2* Haplotypes With Estimated Frequencies >1%

| No. | Haplotype[a] | Frequency (%) |
|---|---|---|
| 1 | G-A-**C**-C-**C**-TTT | 65.4 |
| 2 | G-T-**A**-C-**C**-TTT | 18.3 |
| 3 | G-A-**C**-C-**T**-delTTT | 14.9 |

[a]Haplotypes are designated with the SNPs in the following order: rs1804687-rs2001409-rs1131400-c.1040C>T-rs1131431-rs4177. htSNPs are in bold.

(e.g., all three mutations lead to a significant loss in HIF1A transcriptional repressive capacity of CITED2). Moreover, we observed a significantly diminished TFAP2C coactivation of the p.Ser170_Gly178del mutant. These findings indicate a modifying role for the serine-glycine rich junction in CITED2 function, to which no functionality had been assigned to date. One might speculate that variations in this region cause conformational changes, altering the ability of the EP300 binding domain to interact with CREBBP and EP300 or to recruit other cofactors.

This suggests that the detected CITED2 mutations are potential risk factors for CHD and account for ~2% (8/392) of our patient cohort of sporadic CHD cases. The broad phenotypical spectrum of heart defects seen in *Cited2*−/− mice, as well as in our patients, points to other, potentially relevant but currently unknown modifying factors. In the future it will be of interest to evaluate the functionality of the serine-glycine-rich junction of CITED2 and the non-amino acid altering mutations observed in the coding region, the 5′UTR and 3′UTR. To gain insights into the

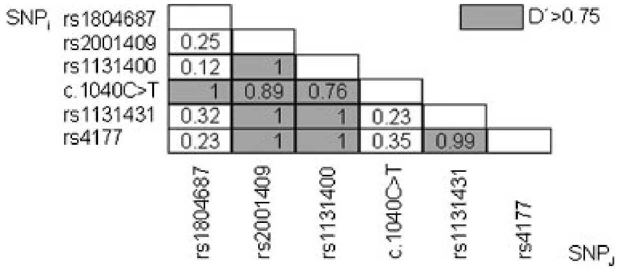


FIGURE 4. **Pairwise allelic association of SNPs in *CITED2* as measured by D′ (numbers). Only SNPs with minor allele frequencies >1% are included.**

inheritance of the mutations, it would be useful to analyze family triplets, which unfortunately were not available for the present study.

Finally, in a comparison of patients and controls we did not observe any significant differences in allele frequencies of the common variants, in accordance with a previous study [Volcik et al., 2004]. Haplotype analysis showed that only three out of 32 possible haplotypes accounted for at least 98.6% of the investigated chromosomes, which suggests the existence of only one haplotype block (Table 5; Fig. 4). Therefore, two htSNPs were extracted that are sufficient for haplotype determination.

In summary, we present the first evidence that CITED2 is a disease-causing gene for congenital heart malformations (particularly septal defects and malrotations of the great arteries) in humans.

## ACKNOWLEDGMENTS

We thank Bernd Timmermann for the haplotype analysis, Bettina Moser and Richard Reinhardt for the DHPLC analysis, and Stefan Wiemann for helpful discussions. We are deeply grateful to the patients and families for their cooperation.

## REFERENCES

Bamford RN, Roessler E, Burdine RD, Saplakoglu U, dela Cruz J, Splitt M, Goodship JA, Towbin J, Bowers P, Ferrero GB, Marino B, Schier AF, Shen, Muenke M, Casey B. 2000. Loss-of-function mutations in the EGF-CFC gene CFC1 are associated with human left-right laterality defects. Nat Genet 26:365–369.

Bamforth SD, Braganca J, Eloranta JJ, Murdoch JN, Marques FI, Kranc KR, Farza H, Henderson DJ, Hurst HC, Bhattacharya S. 2001. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nat Genet 29:469–474.

Bamforth SD, Braganca J, Farthing CR, Schneider JE, Broadbent C, Michell AC, Clarke K, Neubauer S, Norris D, Brown NA, Anderson RH, Bhattacharya S. 2004. Cited2 controls left–right patterning and heart development through a Nodal-Pitx2c pathway. Nat Genet 36:1189–1196.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J, Renaul B, Kucherlapati R, Seidman JG, Seidman CE. 1997. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. Nat Genet 15:30–35.

Bhattacharya S, Michels CL, Leung MK, Arany ZP, Kung AL, Livingston DM. 1999. Functional role of p35srj, a novel p300/CBP binding protein, during transactivation by HIF-1. Genes Dev 13:64–75.

Bosher JM, Totty NF, Hsuan JJ, Williams T, Hurst HC. 1996. A family of AP-2 proteins regulates c-erbB-2 expression in mammary carcinoma. Oncogene 13:1701–1707.

Braganca J, Eloranta JJ, Bamforth SD, Ibbitt JC, Hurst HC, Bhattacharya S. 2003. Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2. J Biol Chem 278:16021–16029.

Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, Thomas NR, Scotter AJ, Caves LS, Tyrrell GP, Newbury-Ecob RA, Munnich A, Bonnet D, Brook JD. 2005. Mutation in myosin heavy chain 6 causes atrial septal defect. Nat Genet 37: 423–428.

Eng C, Brody LC, Wagner TM, Devilee P, Vijg J, Szabo C, Tavtigian SV, Nathanson KL, Ostrander E, Frank TS. 2001. Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. J Med Genet 38:824–833.

Freedman SJ, Sun ZY, Kung AL, France DS, Wagner G, Eck MJ. 2003. Structural basis for negative regulation of hypoxia-inducible factor-1alpha by CITED2. Nat Struct Biol 10: 504–512.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, Rothrock CR, Eapen RS, Hirayama-Yamada K, Joo K, Mastsuoka R, Cohen JC, Srivastava D. 2003. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. Nature 424:443–447.

Huang LE, Gu J, Schau M, Bunn HF. 1998. Regulation of hypoxia-inducible factor 1alpha is mediated by an O2-dependent degradation domain via the ubiquitin-proteasome pathway. Proc Natl Acad Sci USA 95:7987–7992.

Jones AC, Austin J, Hansen N, Hoogendoorn B, Oefner PJ, Cheadle JP, O'Donovan MC. 1999. Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. Clin Chem 45(8 Pt 1):1133–1140.

Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S. 2003. Genome-wide array analysis of normal and malformed human hearts. Circulation 107:2467–2474.

Leung MK, Jones T, Michels CL, Livingston DM, Bhattacharya S. 1999. Molecular cloning and chromosomal localization of the human CITED2 gene encoding p35srj/Mrg1. Genomics 61: 307–313.

Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T, Bonnet D, Lyonnet S, Young ID, Raeburn JA, Buckler AJ, Law DJ, Brook JD. 1997. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. Nat Genet 15:21–29.

Myakishev MV, Khripin Y, Hu S, Hamer DH. 2001. High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. Genome Res 11:163–169.

Olson EN. 2004. A decade of discoveries in cardiac biology. Nat Med 10:467–474.

Petrij F, Giles RH, Dauwerse HG, Saris JJ, Hennekam RC, Masuno M, Tommerup N, van Ommen GJ, Goodman RH, Peters DJ, Breuning MH. 1995. Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. Nature 376:348–351.

Pizzuti A, Sarkozy A, Newton AL, Conti E, Flex E, Digilio MC, Amati F, Gianni D, Tandoi C, Marino B, Crossley M, Dallapiccola B. 2003. Mutations of ZFPM2/FOG2 gene in sporadic cases of tetralogy of Fallot. Hum Mutat 22:372–377.

Reamon-Buettner SM, Borlak J. 2004. TBX5 mutations in non-Holt-Oram syndrome (HOS) malformed hearts. Hum Mutat 24:104.

Rickert AM, Borodina TA, Kuhn EJ, Lehrach H, Sperling S. 2004. Refinement of single-nucleotide polymorphism genotyping methods on human genomic DNA: amplifluor allele-specific polymerase chain reaction versus ligation detection reaction-TaqMan. Anal Biochem 330:288–297.

Roelfsema JH, White SJ, Ariyurek Y, Bartholdi D, Niedrist D, Papadia F, Bacino CA, den Dunnen JT, van Ommen GJ, Breuning MH, Hennekam RC, Peters DJ. 2005. Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. Am J Hum Genet 76: 572–580.

Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, Maron BJ, Seidman CE, Seidman JG. 1998. Congenital heart disease caused by mutations in the transcription factor NKX2-5. Science 281:108–111.

Seelow D, Galli R, Mebus S, Sperling HP, Lehrach H, Sperling S. 2004. d-matrix–database exploration, visualization and analysis. BMC Bioinformatics 5:168.

Solloway MJ, Harvey RP. 2003. Molecular pathways in myocardial development: a stem cell perspective. Cardiovasc Res 58: 264–277.

Srivastava D. 2001. Genetic assembly of the heart: implications for congenital heart disease. Annu Rev Physiol 63:451–469.

Volcik KA, Zhu H, Finnell RH, Shaw GM, Canfield M, Lammer EJ. 2004. Evaluation of the Cited2 gene and risk for spina bifida and congenital heart defects. Am J Med Genet A 126:324–325.

Ware SM, Peng J, Zhu L, Fernbach S, Colicos S, Casey B, Towbin J, Belmont JW. 2004. Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. Am J Hum Genet 74:93–105.

Weninger WJ, Floro KL, Bennett MB, Withington SL, Preis JI, Barbera JP, Mohun TJ, Dunwoodie SL. 2005. Cited2 is required both for heart morphogenesis and establishment of the left–right axis in mouse development. Development 132:1337–1348.

Yin Z, Haynie J, Yang X, Han B, Kiatchoosakun S, Restivo J, Yuan S, Prabhakar NR, Herrup K, Conlon RA, Hoit BD, Watanabe M, Yang YC. 2002. The essential role of Cited2, a negative regulator for HIF-1alpha, in heart development and neurulation. Proc Natl Acad Sci USA 99:10488–10493.

### 2.2.2 Characterization of T-box transcription factor TBX20 in human hearts

Hammer S, Toenjes M, Lange M, Fischer JJ, Dunkel I, Mebus S, Grimm CH, Hetzer R, Berger F, **Sperling S**. Characterization of TBX20 in human hearts and its regulation by TFAP2. *J Cell Biochem* 2008;**104**:1022-1033.

During heart development *Tbx20* expression becomes gradually enriched in the atrioventricular channel, the outflow tract and the developing right ventricle and valves (Iio et al., 2001; Kraus et al., 2001; Stennard et al., 2003; Plageman and Yutzey, 2004; Takeuchi et al., 2005). It is essential for the correct formation of these structures as reduced *Tbx20* expression results in abnormal heart morphogenesis in zebrafish and mouse models (Szeto et al., 2002; Brown et al., 2005; Cai et al., 2005; Singh et al., 2005; Stennard et al., 2005; Takeuchi et al., 2005). Mechanistically, Tbx20 interacts with major players in the regulation of cardiac development such as Tbx5, Gata4, Gata5, Isl1 and Nkx2-5, acting as a transcriptional repressor of *Tbx2* or activator of *Mef2C* and *Nkx2-5* (Stennard et al., 2003; Brown et al., 2005; Cai et al., 2005; Singh et al., 2005; Takeuchi et al., 2005; Shelton and Yutzey, 2007). Thus Tbx20 has been recognized as a key component of the genetic network controlling regional identity, proliferation and differentiation within the developing heart in a dose-sensitive manner (Toenjes et al., 2008). Phenotypes of mouse embryos with a mild reduction of TBX20 levels show its role in right ventricular growth and outflow tract development (Takeuchi et al., 2005). Recently, mutations in the T-box DNA binding domain of *TBX20* have been detected in two families with cardiac pathologies including septation defects and cardiomyopathy (Kirk et al., 2007). The regulation of *TBX20* and its impact as disease gene for TOF in humans, however, had not been investigated so far.

#### *Cloning of human TBX20 splice variants and their expression in TOF patients*

To characterize the human *TBX20* gene in more detail alignments of known murine *Tbx20* transcripts with the human genome were made. This analysis suggested the potential presence of further *TBX20* splice variants in addition to the annotated human transcript harboring 6 exons (NM_020417). Consequently the full-length human TBX20A transcript was cloned, which is homologous to the mouse *Tbx20a* splice variant. This novel human isoform contains a region of 150 amino acids C-terminal to the T-box, which is predicted to carry strong transactivation and transrepression domains in mice (Stennard et al., 2003). The corresponding murine *Tbx20a* transcript has been shown to be the most abundant splice variant of *Tbx20* in mouse. In accordance with this, quantitative real-time PCR analysis of cDNA derived from normal human heart samples showed a much stronger expression of the *TBX20A* isoform compared to the previously described splice variant in human.

In addition to mutations causing deficient transcription factor activity, the regulatory network during cardiac development has been shown to be dependent on the amount of transcription factors present (Cai et al., 2005; Singh et al., 2005; Takeuchi et al., 2005). Therefore the expression of the T-box genes *TBX5* and *TBX20* was analyzed in biopsies of patients with Tetralogy of Fallot and normal human hearts as well as matched biopsies from patients with isolated ventricular septal defect. Quantitative real-time PCR displayed a significant upregulation of both *TBX20* isoforms in TOF samples but no change of expression levels of *TBX5* (**Figure 10**).

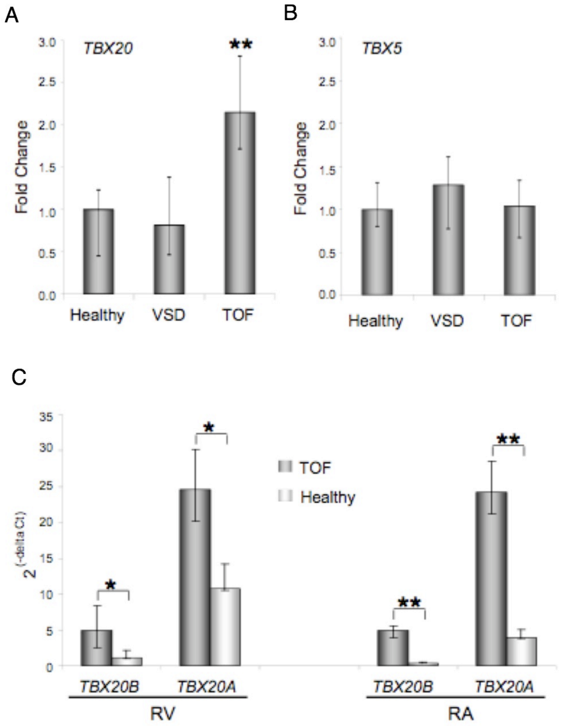


**Figure 10. Overexpression of *TBX20* variants in cardiac samples of patients with TOF.** (A) *TBX20* (both spliceforms) and (B) *TBX5* mRNA expression levels in right ventricular biopsies of patients with Tetralogy of Fallot (TOF; n=13), isolated ventricular septal defects (VSD; n=12) and normal human hearts (healthy; n=6) were quantified by real-time PCR. (C) Expression of *TBX20* splice variants in right ventricular (RV) and right atrial (RA) samples of patients with TOF (n=4) compared to normal human hearts (n=4) as determined by real-time PCR. Results represent median expression levels with 25% and 75% quantile. (*) indicates statistical significance according to Wilcoxon testing. (*) $p<0.05$; (**) $p<0.005$.

### *Regulation of TBX20 by TFAP2*

So far the only described signalling molecule upstream of *Tbx20* was Bmp2, as cultured chicken embryo explants display overexpression of *Tbx20* in its presence (Plageman and Yutzey, 2004). The core promoter of *TBX20* was identified to be located between -629bp and -527bp upstream of the translation start site of *TBX20*. This sequence stretch is responsible for 95% of the transcriptional activity resulting from the *TBX20* locus. In accordance to this, an extended 5´UTR for the *TBX20* transcripts of 527bp was discovered (**Figure 11**). Therefore the mapped transcriptionally active region is about 100bp upstream of the TSS and represents the *TBX20* core promoter. Its sequence is highly conserved between mice and human and contains a GC rich region, harboring potential binding sites for the transcription factors TFAP2 and SP1 as well as E2F. In vivo, all three isoforms of TFAP2, namely TFAP2A, TFAP2B and TFAP2C repress the *TBX20* promoter by 2-3fold, whereas SP1 and E2F do not alter *TBX20* promoter activity. In addition, TFAP2 transcription factors are able to bind to the *TBX20* promoter in vitro and in vivo.

Members of the TFAP2 family share a homologous C-terminal helix-span-helix domain responsible for dimerization and DNA-binding and a proline-glutamine rich transactivation domain at the N-terminus (Eckert et al., 2005). Interestingly, the three TFAP2 family members shown to regulate *TBX20* are expressed in the neural crest during development (Chazaud et al., 1996; Moser et al., 1997). This region contributes to cardiogenesis as progenitor cells from the cardiac neural crest migrate into the developing heart and participate in septation and outflow tract morphogenesis (Harvey, 2002). Moreover, TFAP2A and TFAP2B have been associated with congenital heart defects (Satoda et al., 2000; Brewer et al., 2002). The TFAP2C family member so far had not been implicated in CHD, however, recent studies in zebrafish embryos showed redundant activities of Tfap2a and Tfap2c in neural crest development (Li and Cornell, 2007).

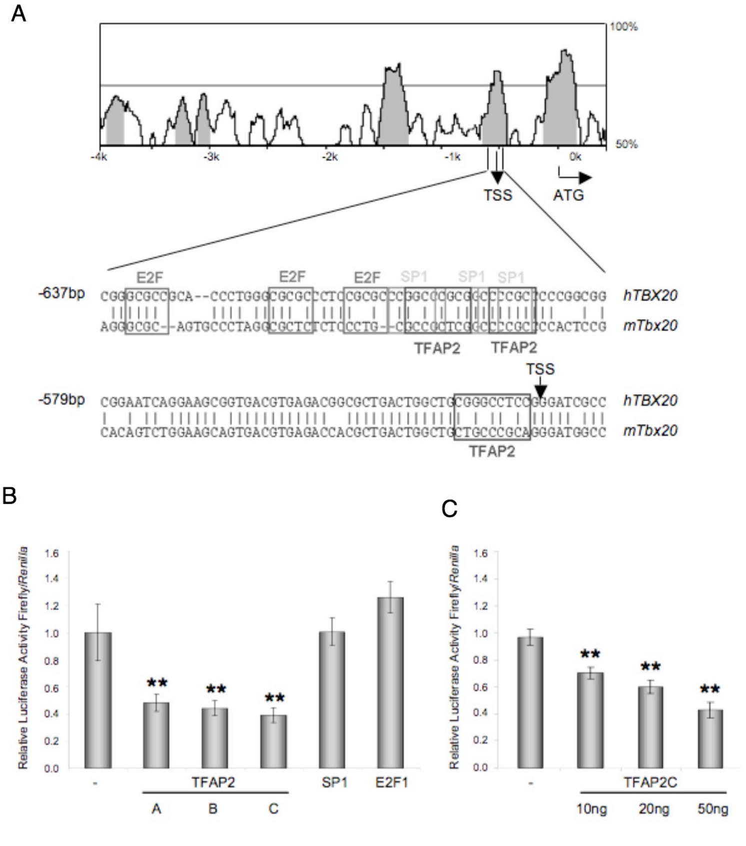


**Figure 11. Transcription factor binding sites and regulation of the *TBX20* core promoter.** (A) Alignment of the human and mouse *TBX20* 5´flanking sequence and potential transcription factor binding sites identified by TRANSFAC (Matys et al., 2003). Predicted transcription factor binding sites for E2F, SP1 and TFAP2 in the putative core promoter are boxed. (B and C) Regulation of the *TBX20* promoter by various transcription factors. HEK293 cells were transfected with expression vectors for the transcription factors as indicated or corresponding empty vectors. Normalized mean luciferase activity is shown compared to unstimulated activity of the -667bp to -7bp construct set as one.

Taken together, overexpression of *TBX20* in TOF patients may result from lack of repression by TFAP2C. Whereas mutational analysis did not show any structural alterations of the TFAP2C DNA binding domain or its cofactor CITED2, a known causative factor for CHD (Schott et al., 1998; Garg et al., 2003; Ware et al., 2004; Sperling et al., 2005), gene expression analysis demonstrated downregulation of *TFAP2C* mRNA in cardiac biopsies from TOF patients. The expression profiling and functional analysis support a role of TFAP2C as a direct transcriptional regulator of *TBX20,* which adds another piece to the transcriptional network important for cardiac development. Animal studies, however, have not yet addressed the consequences of *TBX20* gain of function. These experiments will demonstrate whether elevated levels of *TBX20* alone can mirror the cardiac malformations seen in TOF patients and explain how the cardiac transcriptional network is influenced by *TBX20* overexpression.

**Journal of Cellular Biochemistry 104:1022–1033 (2008)**

# Characterization of TBX20 in Human Hearts and Its Regulation by TFAP2

**Stefanie Hammer,[1] Martje Toenjes,[1] Martin Lange,[1] Jenny J. Fischer,[1] Ilona Dunkel,[1] Siegrun Mebus,[2] Christina H. Grimm,[1] Roland Hetzer,[2] Felix Berger,[3] and Silke Sperling[1]***

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany
[2]Department of Cardiac Surgery, German Heart Center, Berlin, Germany
[3]Department of Pediatric Cardiology, German Heart Center, Berlin, Germany

**Abstract** The T-box family of transcription factors has been shown to have major impact on human development and disease. In animal studies Tbx20 is essential for the development of the atrioventricular channel, the outflow tract and valves, suggesting its potential causative role for the development of Tetralogy of Fallot (TOF) in humans. In the presented study, we analyzed *TBX20* in cardiac biopsies derived from patients with TOF, ventricular septal defects (VSDs) and normal hearts. Mutation analysis did not reveal any disease causing sequence variation, however, *TBX20* is significantly upregulated in tissue samples of patients with TOF, but not VSD. In depth analysis of *TBX20* transcripts lead to the identification of two new exons 3′ to the known *TBX20* message resembling the mouse variant *Tbx20a*, as well as an extended 5′UTR. Functional analysis of the human *TBX20* promoter revealed a 100 bp region that contains strong activating elements. Within this core promoter region we recognized functional binding sites for TFAP2 transcription factors and identified TFAP2 as repressors of the *TBX20* gene in vitro and in vivo. Moreover, decreased *TFAP2C* levels in cardiac biopsies of TOF patients underline the biological significance of the pathway described. In summary, we provide first insights into the regulation of TBX20 and show its potential for human congenital heart diseases. J. Cell. Biochem. 104: 1022–1033, 2008. © 2008 Wiley-Liss, Inc.

**Key words:** T-box; TBX20; TFAP2; congenital heart disease; gene expression

Congenital heart defects (CHD) account for the largest number of birth defects in human, with an incidence of about eight per 1,000 live births. Nearly 30% of major cardiac malformations are associated with additional developmental abnormalities and result from a recognized chromosomal anomaly or occur as part of a syndrome. Major insights into cardiac development and disease have been gained in studies of animal models, such as mice, chicken, and zebrafish, showing that a complex molecular regulatory network is required to initiate and complete the formation of a functional heart [Cripps and Olson, 2002; Brown et al., 2005]. The transcriptional regulation process seems to play one key role in this process (e.g., *Pitx2*, *Isl1*, *Myocardin*, *Hand2*) [Bruneau, 2002], supported also by knowledge gained from mutation reports of patients (e.g., *NKX2-5*, *ZIC3*, *GATA4*, and *CITED2*) [Schott et al., 1998; Garg et al., 2003; Ware et al., 2004; Sperling et al., 2005]. Tetralogy of Fallot (TOF) is a combination of anatomic abnormalities arising mainly from the maldevelopment of the right ventricular outflow tract. Clinically, TOF is characterized by a subaortic ventricular septal defect (VSD), right ventricular infundibular stenosis, aortic valve overriding the right ventricle and right ventricular hypertrophy. As for the overwhelming majority of CHD, the molecular pathology of TOF is so far still poorly understood and major efforts to identify associated molecular factors are currently undertaken.

---

S. Hammer and M. Toenjes contributed equally to this study.

Grant sponsor: BMBF; Grant number: # 0313349; Grant sponsor: EU FP6 IP Moltools; Grant number: # 503155; Grant sponsor: EU FP6 IP HeartRepair; Grant number: # 018630.

*Correspondence to: Silke Sperling, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany. E-mail: sperling@molgen.mpg.de

Received 16 October 2007; Accepted 7 December 2007

DOI 10.1002/jcb.21686

© **2008 Wiley-Liss, Inc.**

T-box genes represent a family of transcription factors that share a highly conserved DNA-binding region (called T-box) and are suggested to play a crucial role in the development of CHD in human. Several family members show cardiac expression during early embryogenesis, such as *Tbx1*, *Tbx2*, *Tbx3*, *Tbx5*, *Tbx18*, and *Tbx20* [Plageman and Yutzey, 2005; Stennard and Harvey, 2005]. Deletions of *TBX1* have been shown in individuals with DiGeorge syndrome [Yagi et al., 2003] and mutations or haploinsufficiency of *TBX5* are frequent causes of Holt–Oram syndrome associated with atrial septal defects and first or second degree atrioventricular block [Basson et al., 1997; Li et al., 1997]. Together with *Tbx5*, the T-box transcription factor *Tbx20* is one of the first genes expressed in the vertebrate cardiac lineage showing a conserved expression pattern in cardiac structures from *drosophila* to mammals [Meins et al., 2000; Kraus et al., 2001; Plageman and Yutzey, 2005]. During development *Tbx20* expression becomes gradually enriched in the atrioventricular channel, the outflow tract and the developing right ventricle and valves [Iio et al., 2001; Kraus et al., 2001; Stennard et al., 2003; Plageman and Yutzey, 2004; Takeuchi et al., 2005]. It is essential for the correct formation of these structures as reduced *Tbx20* expression results in abnormal heart morphogenesis in zebrafish and mouse models [Szeto et al., 2002; Brown et al., 2005; Cai et al., 2005; Singh et al., 2005; Stennard et al., 2005; Takeuchi et al., 2005]. Mechanistically, Tbx20 interacts with major players in the regulation of cardiac development such as Tbx5, Gata4, Gata5, Isl1, and Nkx2-5, acting as a transcriptional repressor of *Tbx2* or activator of *Mef2C* and *Nkx2-5* [Stennard et al., 2003; Brown et al., 2005; Cai et al., 2005; Singh et al., 2005; Shelton and Yutzey, 2007]. Thus Tbx20 has been recognized as a key component of the genetic network controlling regional identity, proliferation and differentiation within the developing heart in a dose-sensitive manner. Phenotypes of mouse embryos with a mild reduction of TBX20 levels show its role in right ventricular growth and outflow tract development [Takeuchi et al., 2005]. Recently, mutations in the T-box DNA binding domain of *TBX20* have been detected in two families with cardiac pathologies including septation defects and cardiomyopathy [Kirk et al., 2007]. The regulation of *TBX20* and its impact as disease gene for TOF in humans, however, has not been investigated to date.

In the study presented, we analyzed the *TBX20* gene in human and show increased *TBX20* expression levels in atrial and ventricular biopsies from TOF patients compared to patients with isolated VSD and normal human heart samples. Further, we characterized the core promoter of *TBX20* and show that TFAP2 transcription factors are direct repressors of *TBX20* in vitro and in vivo. This might represent a regulatory pathway for *TBX20* upregulation in TOF patients as *TFAP2C* expression levels are decreased in respective samples. No sequence mutations could be observed for *TBX20* or the DNA binding domain of *TFAP2C* in analyzed patients.

## MATERIALS AND METHODS

### Patient Samples

All cardiac samples were obtained from the German Heart Center during cardiac surgery with ethical approval by the Institutional Review Committee and informed consent of the patients or parents. Biopsies were taken from the right ventricle and atrium of patients with TOF as well as age and sex matched samples from individuals with VSD from the same tissue region. Samples of all four heart chambers were obtained from normal human hearts.

### RNA and DNA Isolation and Quantitative Real-Time PCR

Total RNA and genomic DNA of all cardiac tissues were extracted using TRIzol (Invitrogen, Karlsruhe, Germany) according to manufacturer's instructions. Five micrograms of total RNA was reverse transcribed and real-time PCR carried out using SYBR Green PCR master mix (ABgene, Epsorn, UK) on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA) with primers for *TBX5*, *TBX20* isoforms as well as *TFAP2* genes. The housekeeping genes *ABL*, *B2M*, and *HPRT* were used for normalization as described [Vandesompele et al., 2002].

### Mutation Analysis

Genomic DNA extracted from patient heart biopsies was amplified using the GenomiPhi-Kit (Amersham Biosciences, Piscataway, NJ).

All exons and the 700 bp promoter region of *TBX20* as well as exons 4 and 5 of *TFAP2C* were amplified by PCR using Hotstar *Taq* polymerase (Qiagen, Hilden, Germany). Sequences of the primers utilized in this study are available upon request. PCR fragments were sequenced by the Services in Molecular Biology Company (Berlin, Germany).

### Plasmid Constructs

Human *TBX20* promoter-luciferase plasmids were generated by cloning the 1,540 bp fragment of the human *TBX20* 5′flanking region between −1,546 and −7 bp relative to the initiation codon into *Kpn*I/*Nhe*I sites of the luciferase reporter gene plasmid *pGL3basic* (Promega, Mannheim, Germany). The resulting full-length promoter–reporter plasmid was denoted as *−1,546-TBX20-Luc*. Sequential deletion constructs were created as indicated in Figure 3. Expression vectors for SP1, TFAP2A, TFAP2B, TFAP2C, and E2F1 were described previously and generously donated by Guntram Suske, Helen Hurst, Ronald J. Weigel, and Joseph R. Nevins [Hagen et al., 1994; Schwarz et al., 1995; Bosher et al., 1996; Bamforth et al., 2001].

### Cell Culture, Transfection, and Luciferase Assay

The human cell lines HEK293 and HepG2 as well as C2C12 mouse myoblasts were maintained in DMEM +10% FBS. HL1 mouse cardiomyocytes were obtained from William C. Claycomb and cultured as described [Claycomb et al., 1998]. Cells were transfected using Transfast (Promega) or Dreamfect (Oz Biosciences, Marseille, France) according to manufacturers' instructions. Reporter gene assays for luciferase activity were performed as described previously [Sperling et al., 2005].

### 5′UTR Mapping

The investigation of the *TBX20* 5′UTR was carried out by PCR using cDNA derived from HEK293 cells. The reverse primer was located in exon 2 (+168 to +188 bp relative to the A of the ATG initiation codon) and a panel of forward primers upstream of the translation start site as indicated in Figure 3b.

### Electromobility Shift Assay

Nuclear extracts were prepared from HEK293 cells after transfection with TFAP2C expression plasmid or empty vector. Double-stranded oligonucleotides containing the putative TFAP2 binding sites within the *TBX20* promoter were generated by annealing complementary single-stranded oligonucleotides (cgcccggcccgcggccccgcccccggcggcggaatca) and subsequently end-labeled with digoxygenin-11-ddUTP using the DIG Gel Shift Kit 2nd Generation (Roche Diagnostics, Mannheim, Germany). For binding reactions, 3 μg of nuclear extract and 0.8 ng labeled oligonucleotides were incubated, for competition experiments a 100-fold excess of unlabeled competitor DNAs was added to the mixture. After the binding reaction, samples were subjected to electrophoresis on a 6% TBE DNA Retardation Gel (Novex, Invitrogen) and visualized by autoradiography.

### Chromatin-Immunoprecipitation (ChIP)

ChIP experiments were performed on duplicate sets of HL1 cells essentially as described previously [Horak et al., 2002]. Modifications of the assay protocol were as follows: cells were cross-linked for 10 min at 37°C and samples sonified using a Branson 250 Sonifier with 12 pulses at power-setting of 6% and 100% duty-cycle for 30 s and 2 min on ice between pulses. Immunoprecipitation was carried out with magnetic protein A/G beads (Invitrogen) and TFAP2 antibody (#sc-8977, Santa Cruz Biotechnology, Inc., CA) at 5 μg/ml concentration. Enrichment of TFAP2 target sequences over input was quantified by real-time PCR as described above.

## RESULTS

### Human *TBX20* Splice Variants and Their Expression in Normal Human Hearts

To characterize the human *TBX20* gene in more detail we generated alignments of known murine *Tbx20* transcripts with the human genome. This analysis suggested the potential presence of further *TBX20* splice variants in addition to the annotated human transcript harboring six exons (NM_020417). RT-PCR performed on cDNA from HEK293 cells as well as human myocardium showed expression of exons 7 and 8, homologous to the mouse *Tbx20a* splice variant. We cloned the full-length human *TBX20A* transcript, submitted to Genbank (accession number NM_001077653; Fig. 1a,b).

This novel human isoform contains a region of 150 amino acids C-terminal to the T-box, which

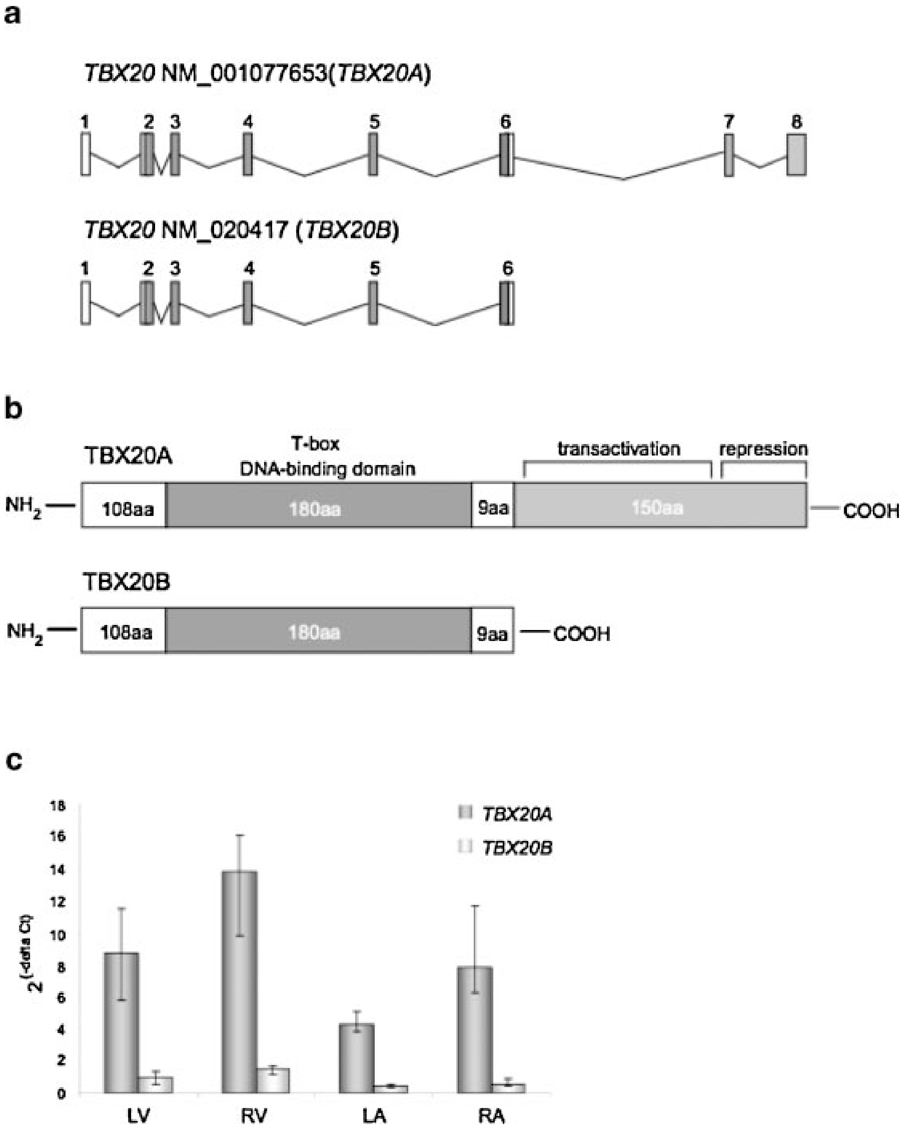


**Fig. 1.** Structure and expression of human TBX20 isoforms. **a**: Intron/exon structure of human *TBX20* transcript variant A isolated from HEK293 total cDNA compared to known *TBX20B*. Exons are represented as boxes and the position of the 180aa T-box domain is shown in dark gray. Novel exons are depicted in light gray. **b**: Schematic representation (not to scale) of the TBX20 isoforms. Note that variant TBX20A contains an extension harboring transactivation and transrepression domains. **c**: Real-time PCR analysis of *TBX20A* and *TBX20B* splice variants in cDNA derived from normal human heart tissues ($n = 4$) of left atrium (LA), right atrium (RA), left ventricle (LV), and right ventricle (RV). Results represent median expression levels with 25% and 75% quantile; assays were performed in triplicates.

is predicted to carry strong transactivation and transrepression domains in mice [Stennard et al., 2003]. The corresponding murine *Tbx20a* transcript has been shown to be the most abundant splice variant of *Tbx20* in mouse. In accordance with this quantitative real-time PCR analysis of cDNA derived from normal human heart samples showed a much stronger expression of the *TBX20A* isoform compared to the previously described splice variant in human, which is designated *TBX20B* in the paper presented. Expression profiles of the *TBX20A* and *TBX20B* transcripts were similar in cDNAs from all four chambers of the human heart (Fig. 1c).

## Mutational Analysis of *TBX20* in Patients With TOF

To analyze genomic alterations of *TBX20* potentially causative for CHD in human, we screened 23 patients with TOF by sequencing

TABLE I. Mutation Analysis of the *TBX20* Gene in Patients With TOF

| dbSNP | Position | Nucleotide variation | Amino acid variation | Mut chr | Total chr | Mut allele freq | Mut allele freq dbSNP |
|---|---|---|---|---|---|---|---|
| | 5′UTR | c.−186T > C | | 13 | 46 | 0.283 | |
| rs336283 | Exon 1 | c.39T > C | p.Ser13Ser | 33 | 46 | 0.717 | 0.735 |
| rs17675148 | Intron 3 | c.545 + 13A > G | | 9 | 38 | 0.237 | 0.263 |

Systematic nomenclature for SNPs (www.hgvs.org) based on GenBank NM_001077653 (*TBX20A* cDNA) and counting +1 as A of the initiation codon. Mut, mutant; chr, chromosome; freq, frequency.

all *TBX20* exons including their flanking intronic regions and 700 bp 5′ of the translation start site, a region potentially containing regulatory elements for *TBX20*. The results from this mutation screen are presented in Table I. We detected two previously known sequence variations showing the same distribution as in the normal population (NCBI dbSNP) and one additional nucleotide variation 5′ to the start codon. Further sequence variations, which are also currently associated with *TBX20* in dbSNP, resulted from amplification of the *TBX20* pseudogene on chromosome 12 that comprises exons 5–8 of *TBX20* on chromosome 7. However, analysis of cDNA demonstrated that the pseudogene is not transcribed suggesting its functional silence and cDNA analysis of *TBX20* showed the absence of the proposed alterations. Homology studies revealed that the mouse genome lacks a *Tbx20* pseudogene.

### Increased Cardiac *TBX20* Expression Levels in Patients With TOF

In addition to mutations potentially causing deficient transcription factor activity the regulatory network during cardiac development has been shown to be dependent on the amount of transcription factors present in the corresponding tissue [Cai et al., 2005; Singh et al., 2005; Takeuchi et al., 2005]. Therefore we questioned whether the T-box genes *TBX5* and *TBX20* would be deregulated in biopsies of 13 patients with TOF whose genomic DNA was included in the mutation analysis. A group of 8 samples of normal human hearts served as control and 12 age matched biopsies from patients with isolated VSD. Quantitative real-time PCR displayed a significant upregulation of *TBX20* in TOF samples compared to normal human hearts and VSD samples ($P < 0.005$; Fig. 2a). In contrast, expression levels of *TBX5* were not significantly altered in either group of individuals (Fig. 2b). Next, we analyzed the expression of the different *TBX20* splice variants in representative atrial and ventricular samples of TOF patients compared to normal human hearts. In these samples both *TBX20* isoforms were found to be upregulated compared to normal human hearts ($P < 0.05$ and $P < 0.005$; Fig. 2c). Again, *TBX5* levels did not differ between the groups (data not shown).

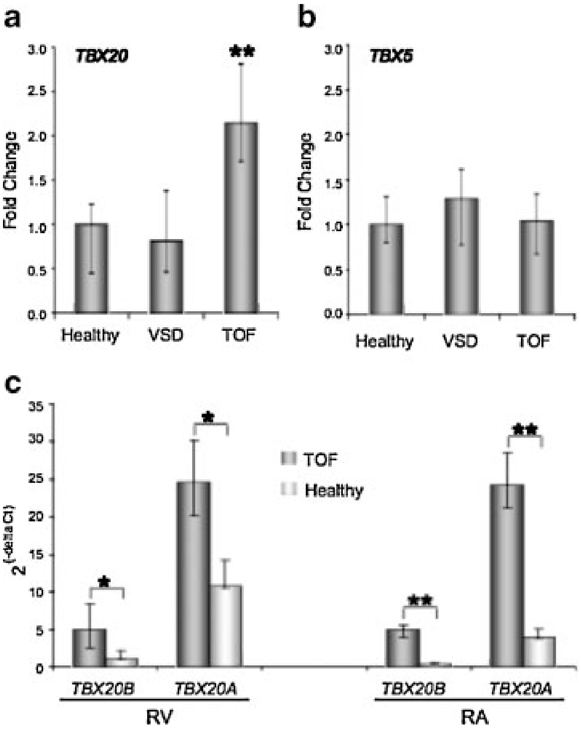


**Fig. 2.** Overexpression of *TBX20* variants in cardiac samples of patients with TOF. **a**: *TBX20* (both splice forms) and (**b**) *TBX5* mRNA expression levels in right ventricular biopsies of patients with Tetralogy of Fallot (TOF; n = 13), isolated ventricular septal defects (VSDs; n = 12) and normal human hearts (healthy; n = 6) were quantified by real-time PCR. **c**: Expression of *TBX20* splice variants in right ventricular (RV) and right atrial (RA) samples of patients with TOF (n = 4) compared to normal human hearts (n = 4) as determined by real-time PCR. Results represent median expression levels with 25% and 75% quantile. * Indicates statistical significance according to Wilcoxon testing. (*) $P < 0.05$; (**) $P < 0.005$.

### Identification of the *TBX20* Core Promoter and 5′UTR

To elucidate the regulatory region of the human *TBX20* gene we cloned a fragment comprising nucleotides −1,546 and −7 relative to the translation start site counting the A of the initiation codon as +1. This region was able to drive expression of a luciferase gene when cloned in a corresponding vector about 30-fold higher compared to the activity of the empty vector after transfection in HEK293 cells (Fig. 3a). To define the minimal promoter region of *TBX20* we generated a series of truncated constructs and characterized the basal activity in HEK293 cells. As shown in Figure 3a, a region between −629 and −527 bp relative to the translational start site is responsible for the major increase in promoter activity, as between −629 and −527 bp the transcriptional activity of the construct decreased sequentially by about five- to sixfold. Similar results were obtained in HepG2 cells as well as C2C12 mouse myoblasts (data not shown), suggesting that major regulatory elements of the *TBX20* gene are located in a region between −629 and −527 bp 5′ of the ATG initiation codon. We therefore suggest that this region represents the *TBX20* core promoter serving as recognition site for the basal transcription apparatus which is typically a 100 bp region flanking the transcriptional start site (TSS). Moreover, our data show that all constructs with inserts containing less than −527 bp exhibit only minor transcriptional activity. This 527 bp region is homologous to the murine *Tbx20* 5′UTR and using primer walking analysis we could also annotate it as the 527 bp long 5′UTR in the human *TBX20* transcripts (Fig. 3b). This TSS maps well with the one proposed by prediction programs (Dragon GSF1.0, Eponine, Mc Promoter, NNPP2.1, Promoter Scan, TSSG and TSSW).

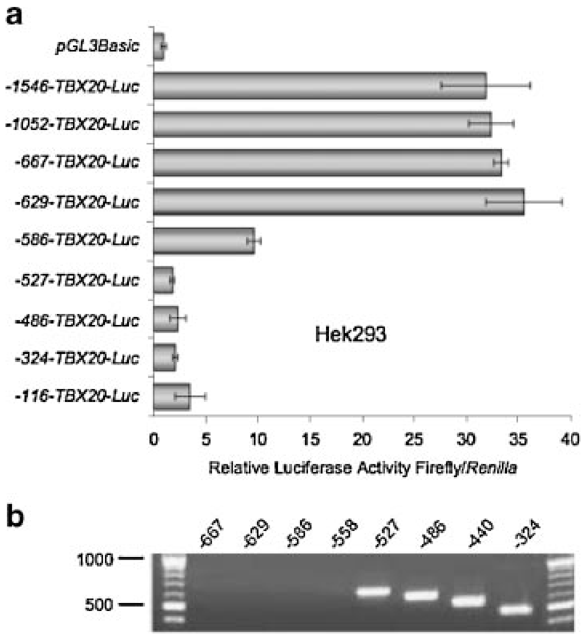


**Fig. 3.** Identification of the *TBX20* core promoter and 5′UTR. **a**: Luciferase activity assays in HEK293 cells transfected with different *TBX20* promoter constructs. The fragments between −1,546, −1,052, −667, −629, −586, −527, −486, −324, and −116 to −7 relative to the A of the initiation codon of the human *TBX20* gene (NM_001077653) were PCR amplified and cloned into *pGL3basic*. Firefly luciferase activity of the resulting plasmids was normalized to *Renilla* luciferase activity to account for differences in transfection efficiency. The mean luciferase activity of transient transformants is presented as fold change compared to basal activity of the *pGL3basic* vector from one representative experiment performed in triplicates, error bars represent standard deviations. The assays were repeated at least three times independently. **b**: Mapping of the transcriptional start site of human *TBX20* by RT-PCR analysis of cDNA from HEK293 cells with forward primers upstream of the translation start site as indicated.

### TFAP2 Isoforms Dose Dependently Downregulate the *TBX20* Promoter

Promoter analysis using TRANSFAC [Matys et al., 2003] revealed that the region identified as the *TBX20* core promoter harbors several GC-boxes that represent potential binding sites for the transcription factors SP1, E2F, and the TFAP2 family (Fig. 4a). Cotransfection of corresponding expression constructs in HEK293 cells with the −667 to −7 bp *TBX20* promoter construct in the presence of empty vector or TFAP2 expression plasmids revealed that TFAP2A, TFAP2B, and TFAP2C significantly downregulate *TBX20* promoter activity by about threefold. In contrast cotransfection with expression constructs for transcription factors SP1 and E2F had no effect on the luciferase level (Fig. 4b). The repressive effects of all three TFAP2 isoforms showed dose-dependency (Fig. 4c and data not shown).

### *TFAP2*-Response Elements Drive Promoter Activity In Vitro and In Vivo

To investigate the impact of putative TFAP2 binding sites on promoter regulation we transfected HEK293 cells with different *TBX20* promoter constructs in the presence or absence

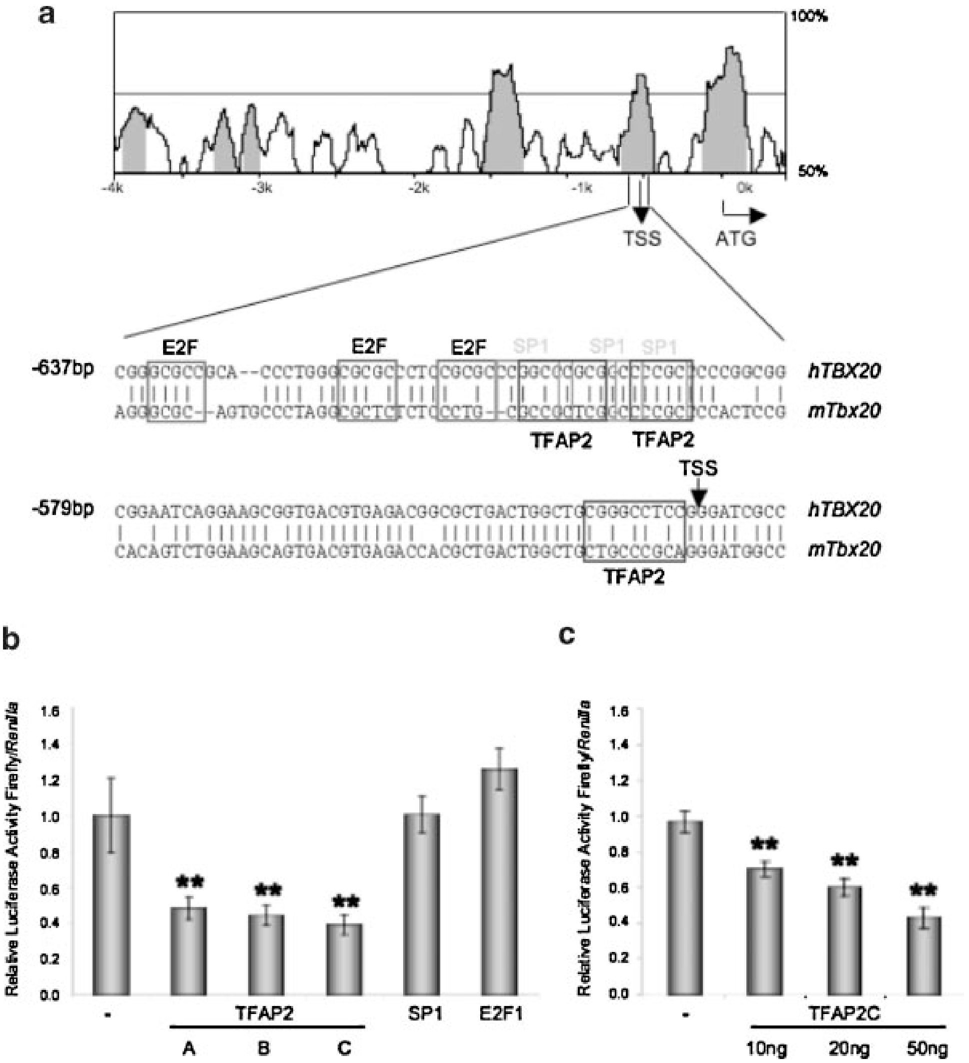


**Fig. 4.** Transcription factor binding sites and regulation of the *TBX20* core promoter. **a**: Alignment of the human and mouse *TBX20* 5′flanking sequence generated by mVISTA (http://genome.lbl.gov/vista/index.shtml) and potential transcription factor binding sites identified by TRANSFAC [Matys et al., 2003]. Predicted transcription factor binding sites for E2F, SP1, and TFAP2 in the putative core promoter are boxed. **b**,**c**: Regulation of the *TBX20* promoter by various transcription factors. HEK293 cells were transfected with expression vectors for the transcription factors as indicated or corresponding empty vectors. Normalized mean luciferase activity is shown compared to unstimulated activity of the −667 to −7 bp construct set as one.

of TFAP2 expression plasmids. TFAP2 isoforms repressed transcriptional activity of the −667 to −7 bp and −629 to −7 bp promoter constructs by two- to threefold, in contrast no effects could be observed when cotransfecting TFAP2A or TFAP2C to the −586 to −7 bp and −527 to −7 bp promoter construct (Fig. 5a and data not shown). These results suggest the functionality of a repressive TFAP2 binding site between 629 and 586 bp upstream of the *TBX20* initiation codon. To test whether TFAP2 binds to those sites in vitro, we performed gel shift assays of nuclear extracts from HEK293 cells transfected with TFAP2 expression plasmids and oligonucleotides representing the potential binding sites between −629 and −586 bp of the *TBX20* promoter. Figure 5b shows binding with nuclear extracts from TFAP2 transfected cells, whereas there is no signal in the non-transfected cells. In the presence of a 100-fold molar excess of competitor oligonucleotides, complexes of the labeled DNA fragments with TFAP2 were abolished. ChIP experiments in cultured HL1 cells showed about 20-fold enrichment of the corresponding *TBX20* core promoter in samples after precipitation of cross-linked chromatin with TFAP2 antibodies compared to unrelated promoter regions (Fig. 5c). Thus, TFAP2 binds

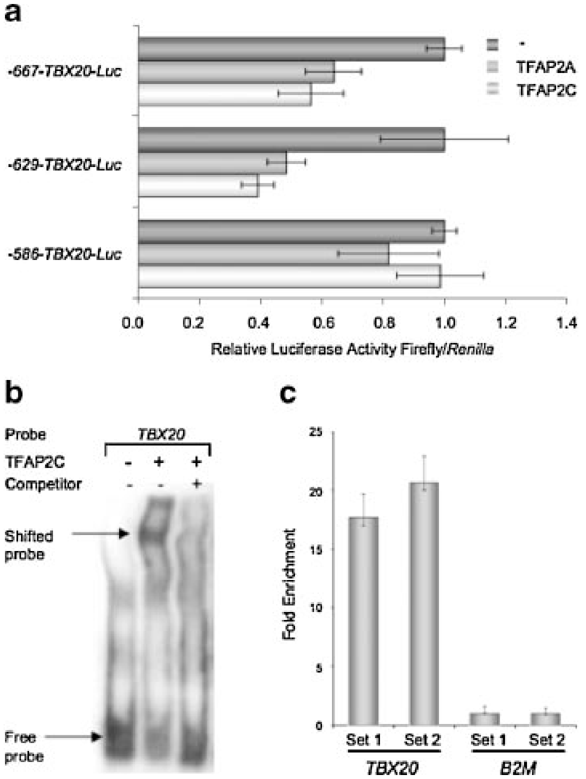


**Fig. 5.** Identification of functional TFAP2 binding sites in the *TBX20* core promoter. **a**: Effects of serial deletions of the region harboring TFAP2 binding sites on promoter regulation by TFAP2C. HEK293 cells were transfected with different *TBX20* promoter constructs as indicated in the presence or absence of TFAP2A and TFAP2C expression vectors. Normalized mean luciferase activities are shown with the luciferase activity of the corresponding unstimulated promoter constructs set to one. **b**: Electrophoretic mobility shift assay with nuclear extracts from HEK293 cells transfected with TFAP2 expression constructs or empty vector and end-labeled oligonucleotide probes containing potential TFAP2 binding sites in the presence or absence of a 100-fold excess of unlabeled oligonucleotides. **c**: Chromatin-immunoprecipitation analysis of HL1 cell extracts immunoprecipitated with TFAP2 antibody in replicates (set 1, set 2). Bound DNA was detected using real-time PCR analysis targeting the *TBX20* core promoter primers and an unrelated negative control (*B2M*).

to these regulatory elements in the *TBX20* promoter in cardiac cells in vivo.

### Decreased Expression Levels of *TFAP2C* in Biopsies of Patients With TOF

To strengthen the biological relevance of TFAP2 regulation of *TBX20* we assessed mRNA levels of *TFAP2* genes in human heart samples. *TFAP2A* and *C* mRNA was present in atrial and ventricular samples, while *TFAP2B* mRNA was not detectable by real-time PCR. Interestingly, we found that *TFAP2C* was significantly downregulated in tissue samples of patients with TOF compared to normal human hearts ($P < 0.005$; Fig. 6a) and patients with VSD ($P < 0.05$), providing a possible explanation for the overexpression of *TBX20*. In contrast, expression levels of *TFAP2A* were unchanged (Fig. 6b). Mutation analysis did not show any structural alterations of the TFAP2C DNA binding domain (data not shown) suggesting that again deregulation rather than mutation is more likely to be responsible for *TBX20* overexpression in TOF patients.

## DISCUSSION

### Splice Variants and Sequence Variations of *TBX20* in Human

The T-box genes *TBX5* and *TBX1* have long been known as disease genes for human CHD. In addition to these two family members, *TBX20* represents a key regulator of embryogenesis and particularly early cardiac development. Lack of Tbx20 leads to various cardiac malformations in animal models such as outflow tract defects and malformed valves and a disturbed expression pattern of a number of other key cardiac transcription factors [Brown et al., 2005; Cai et al., 2005; Singh et al., 2005; Takeuchi et al., 2005]. Moreover, in a recent study mutations in the T-box DNA binding domain of *TBX20* were linked to cardiomyopathy and cardiac septation defects in human

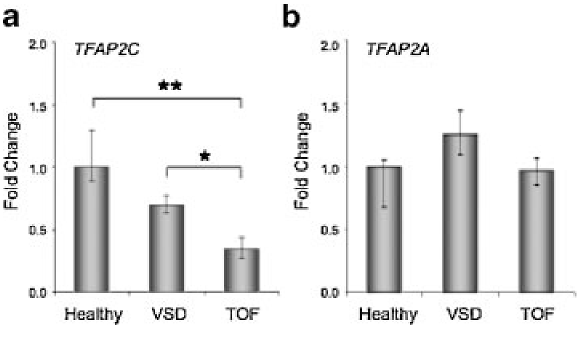


**Fig. 6.** Decreased expression levels of *TFAP2C* in cardiac samples of patients with TOF. **a**: *TFAP2C* and (**b**) *TFAP2A* mRNA expression levels in right ventricular biopsies of patients with Tetralogy of Fallot (TOF; $n = 13$), isolated ventricular septal defects (VSDs; $n = 12$) and normal human hearts (healthy; $n = 6$) were quantified by real-time PCR. Results represent median expression levels with 25% and 75% quantile. Results are shown in relation to the expression levels of the healthy human heart samples. * Indicates statistical significance according to Wilcoxon testing. (*) $P < 0.05$; (**) $P < 0.005$.

[Kirk et al., 2007]. However, the presence of different *TBX20* transcripts in human as well as sequence variations and expression levels of *TBX20* in patients with TOF have not been investigated before. Here, we discovered the presence of a human *TBX20A* splice variant homologous to the murine *Tbx20a* transcript representing the major transcript in both species. The newly discovered human TBX20A comprises C-terminal to the T-box the transactivation and transrepression domains, which are potentially of major impact for the transcriptional activity of TBX20. In contrast to the preferential expression of *Tbx20* transcripts in distinct cardiac regions during cardiac development in mouse, the human *TBX20* splice variants are equally expressed in human left and right atrial and ventricular samples of normal adult hearts.

The cardiac malformations observed in mouse models lacking Tbx20 proposed a potential primarily causative impact of TBX20 on the development of TOF in human. However, in 23 patients studied we could not identify any amino acid changing mutation. This suggests that mutations of TBX20 are not common in humans live births or they may be associated with other CHD not studied [Kirk et al., 2007]. Two sequence variations present in the dbSNP database (www.hgvs.org) could be confirmed at equal frequencies compared to the normal population. One novel nucleotide exchange was discovered within the 5′UTR. Interestingly in contrast to mouse, the human genome harbors a *TBX20* pseudogene on chromosome 12 including exons 5–8 of the *TBX20* transcript. This has to be considered when genotyping *TBX20* DNA as many *TBX20* sequence variations listed in dbSNP arise from the non-transcribed pseudogene.

## Expression of *TBX20* in Human Right Ventricular Samples

Various results from mouse studies have revealed the impact of Tbx20 as a key regulator of transcriptional networks in cardiac development. Thereby, the level of transcription factors plays an important role and is tightly regulated. However, the expression levels of transcription factors in human heart development and malformed hearts are still largely unknown. A previous study on gene expression in malformed human hearts [Kaynak et al., 2003] demonstrated disease specific molecular portraits, with a higher number of genes being upregulated in TOF patients compared to individuals with VSD. This analysis, however, did not include all transcription factors known to play a role in cardiac development. Here we determined the expression levels of the T-box transcription factors *TBX20* and *TBX5* using quantitative real-time PCR in cDNAs derived from human heart tissue samples showing elevated expression of *TBX20* in patients with TOF. In contrast, levels of *TBX5* were not altered in either of the groups.

In depth analysis of *TBX20* transcripts in human revealed that both human isoforms, namely *TBX20A* and *TBX20B* are overexpressed in patients with TOF. This upregulation could be detected in atrial and ventricular samples pointing to a general deregulation in TOF rather than adaptation processes related to cardiac pressure overload and altered hemodynamic features in the ventricle. Thus, the altered expression level of *TBX20* may have a potential impact on the development of TOF in human and we further investigated the upstream regulatory cascade of *TBX20*.

## Regulation of the *TBX20* Gene

So far biochemical and animal studies have investigated the regulation of potential target genes of Tbx20 and its interactions with other cardiac transcription factors. The regulation of the *Tbx20* gene itself, however, is largely unknown to date. The only described signaling molecule upstream of *Tbx20* is Bmp2, as cultured chicken embryo explants display overexpression of *Tbx20* in its presence [Plageman and Yutzey, 2004]. Here, we were able to identify a fragment between −629 and −527 bp upstream of the translation start site of *TBX20* that is responsible for 95% of the transcriptional activity resulting from the *TBX20* locus. In accordance to this, we discovered an extended 5′UTR for the *TBX20* transcripts of 527 bp. Therefore the mapped transcriptionally active region is about 100 bp upstream of the TSS and represents the *TBX20* core promoter. Its sequence is highly conserved between mice and human and contains a GC rich region, harboring potential binding sites for the transcription factors TFAP2 and SP1 as well as E2F. We show that all three isoforms of TFAP2, namely TFAP2A, TFAP2B, and TFAP2C repress the *TBX20* promoter by two- to threefold, whereas SP1 and E2F do not alter *TBX20* promoter activity. In addition, TFAP2 transcription factors are able to bind to the *TBX20* promoter in vitro and in vivo.

Members of the TFAP2 family share a homologous C-terminal helix-span-helix domain responsible for dimerization and DNA-binding and a proline-glutamine rich transactivation domain at the N-terminus [Eckert et al., 2005]. Interestingly, the three TFAP2 family members shown to regulate *TBX20* are expressed in the neural crest during development [Chazaud et al., 1996; Moser et al., 1997]. This region contributes to cardiogenesis as progenitor cells from the cardiac neural crest migrate into the developing heart and participate in septation and outflow tract morphogenesis [Harvey, 2002]. Moreover, TFAP2A and TFAP2B have been associated with CHD. Knock-in mice with functionally deficient *Tfap2a* display cardiac malformations in addition to failing neural tube closure and craniofacial defects [Brewer et al., 2002]. The observed cardiac malformations include a panel of defects associated with perturbed outflow tract formation such as double outlet right ventricle, persistant truncus arteriosus, TOF and severe pulmonary stenosis. In contrast, mutations of *TFAP2B* leading to haploinsufficiency or a dominant negative form of the TFAP2B protein have been associated with Char syndrome in humans, characterized by persistant ductus arteriosus, facial dysmorphism and skeletal abnormalities of the hand [Satoda et al., 2000]. These findings illustrate the role of the *TFAP2* gene family in cardiac morphogenesis, mainly outflow tract formation and cardiac septation, by controlling cell proliferation and terminal differentiation [Eckert et al., 2005; Hutson and Kirby, 2007].

The TFAP2C family member so far has not been implicated in CHD, however, recent studies in zebrafish embryos showed redundant activities of Tfap2a and Tfap2c in neural crest development [Li and Cornell, 2007]. Results presented in this study suggest that overexpression of *TBX20* in TOF patients may result from lack of repression by TFAP2C. Whereas mutational analysis did not show any structural alterations of the TFAP2C DNA binding domain or its cofactor CITED2, a known causative factor for CHD [Schott et al., 1998; Garg et al., 2003; Ware et al., 2004; Sperling et al., 2005], gene expression analysis demonstrated down-regulation of *TFAP2C* mRNA in cardiac biopsies from TOF patients.

To summarize, the present study reveals that mutations in TBX20 and the DNA binding domain of TFAP2C are unlikely to be a major cause of TOF or VSD in human. In contrast, we show that TBX20, a key transcription factor for chamber specific cell differentiation, is overexpressed in TOF patients. Our expression profiling and functional analysis support a role of TFAP2C as a direct transcriptional regulator of *TBX20* which adds another piece to the transcriptional network important for cardiac development. Animal studies, however, have not yet addressed the consequences of *TBX20* gain of function. These experiments will demonstrate whether elevated levels of *TBX20* alone can mirror the cardiac malformations seen in TOF patients and explain how the cardiac transcriptional network is influenced by *TBX20* overexpression.

## ACKNOWLEDGMENTS

We are deeply grateful to the patients and families for their cooperation. This work has been supported by the European Community's Sixth Framework Programme contract ('Heart-Repair') LSHM-CT-2005-018630.

## REFERENCES

Bamforth SD, Braganca J, Eloranta JJ, Murdoch JN, Marques FI, Kranc KR, Farza H, Henderson DJ, Hurst HC, Bhattacharya S. 2001. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nat Genet 29:469–474.

Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J, Renault B, Kucherlapati R, Seidman JG, Seidman CE. 1997. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. Nat Genet 15:30–35.

Bosher JM, Totty NF, Hsuan JJ, Williams T, Hurst HC. 1996. A family of AP-2 proteins regulates c-erbB-2 expression in mammary carcinoma. Oncogene 13:1701–1707.

Brewer S, Jiang X, Donaldson S, Williams T, Sucov HM. 2002. Requirement for AP-2alpha in cardiac outflow tract morphogenesis. Mech Dev 110:139–149.

Brown DD, Martz SN, Binder O, Goetz SC, Price BM, Smith JC, Conlon FL. 2005. Tbx5 and Tbx20 act synergistically to control vertebrate heart morphogenesis. Development 132:553–563.

Bruneau BG. 2002. Transcriptional regulation of vertebrate cardiac morphogenesis. Circ Res 90:509–519.

Cai CL, Zhou W, Yang L, Bu L, Qyang Y, Zhang X, Li X, Rosenfeld MG, Chen J, Evans S. 2005. T-box genes coordinate regional rates of proliferation and regional specification during cardiogenesis. Development 132: 2475–2487.

Chazaud C, Oulad-Abdelghani M, Bouillet P, Decimo D, Chambon P, Dolle P. 1996. AP-2.2, a novel gene related to AP-2, is expressed in the forebrain, limbs and face during mouse embryogenesis. Mech Dev 54:83–94.

Claycomb WC, Lanson NA, Jr., Stallworth BS, Egeland DB, Delcarpio JB, Bahinski A, Izzo NJ, Jr. 1998. HL-1 cells: A cardiac muscle cell line that contracts and retains phenotypic characteristics of the adult cardiomyocyte. Proc Natl Acad Sci USA 95:2979–2984.

Cripps RM, Olson EN. 2002. Control of cardiac development by an evolutionarily conserved transcriptional network. Dev Biol 246:14–28.

Eckert D, Buhl S, Weber S, Jager R, Schorle H. 2005. The AP-2 family of transcription factors. Genome Biol 6:246.

Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, Rothrock CR, Eapen RS, Hirayama-Yamada K, Joo K, Matsuoka R, Cohen JC, Srivastava D. 2003. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. Nature 424:443–447.

Hagen G, Muller S, Beato M, Suske G. 1994. Sp1-mediated transcriptional activation is repressed by Sp3. EMBO J 13:3843–3851.

Harvey RP. 2002. Patterning the vertebrate heart. Nat Rev Genet 3:544–556.

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. Proc Natl Acad Sci USA 99:2924–2929.

Hutson MR, Kirby ML. 2007. Model systems for the study of heart development and disease. Cardiac neural crest and conotruncal malformations. Semin Cell Dev Biol 18: 101–110.

Iio A, Koide M, Hidaka K, Morisaki T. 2001. Expression pattern of novel chick T-box gene, Tbx20. Dev Genes Evol 211:559–562.

Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S. 2003. Genome-wide array analysis of normal and malformed human hearts. Circulation 107:2467–2474.

Kirk EP, Sunde M, Costa MW, Rankin SA, Wolstein O, Castro ML, Butler TL, Hyun C, Guo G, Otway R, Mackay JP, Waddell LB, Cole AD, Hayward C, Keogh A, Macdonald P, Griffiths L, Fatkin D, Sholler GF, Zorn AM, Feneley MP, Winlaw DS, Harvey RP. 2007. Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. Am J Hum Genet 81:280–291.

Kraus F, Haenig B, Kispert A. 2001. Cloning and expression analysis of the mouse T-box gene tbx20. Mech Dev 100:87–91.

Li W, Cornell RA. 2007. Redundant activities of Tfap2a and Tfap2c are required for neural crest induction and development of other non-neural ectoderm derivatives in zebrafish embryos. Dev Biol 304:338–354.

Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T, Bonnet D, Lyonnet S, Young ID, Raeburn JA, Buckler AJ, Law DJ, Brook JD. 1997. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. Nat Genet 15:21–29.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.

Meins M, Henderson DJ, Bhattacharya SS, Sowden JC. 2000. Characterization of the human TBX20 gene, a new member of the T-Box gene family closely related to the Drosophila H15 gene. Genomics 67:317–332.

Moser M, Ruschoff J, Buettner R. 1997. Comparative analysis of AP-2 alpha and AP-2 beta gene expression during murine embryogenesis. Dev Dyn 208:115–124.

Plageman TF, Jr., Yutzey KE. 2004. Differential expression and function of Tbx5 and Tbx20 in cardiac development. J Biol Chem 279:19026–19034.

Plageman TF, Jr., Yutzey KE. 2005. T-box genes and heart development: Putting the “T” in heart. Dev Dyn 232: 11–20.

Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR, Pierpont ME, Gelb BD. 2000. Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. Nat Genet 25:42–46.

Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, Maron BJ, Seidman CE, Seidman JG. 1998. Congenital heart disease caused by mutations in the transcription factor NK X2-5. Science 281:108–111.

Schwarz JK, Bassing CH, Kovesdi I, Datto MB, Blazing M, George S, Wang XF, Nevins JR. 1995. Expression of the E2F1 transcription factor overcomes type beta transforming growth factor-mediated growth suppression. Proc Natl Acad Sci USA 92:483–487.

Shelton EL, Yutzey KE. 2007. Tbx20 regulation of endocardial cushion cell proliferation and extracellular matrix gene expression. Dev Biol 302:376–388.

Singh MK, Christoffels VM, Dias JM, Trowe MO, Petry M, Schuster-Gossler K, Burger A, Ericson J, Kispert A. 2005. Tbx20 is essential for cardiac chamber differentiation and repression of Tbx2. Development 132: 2697–2707.

Sperling S, Grimm CH, Dunkel I, Mebus S, Sperling HP, Ebner A, Galli R, Lehrach H, Fusch C, Berger F, Hammer S. 2005. Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. Hum Mutat 26:575–582.

Stennard FA, Harvey RP. 2005. T-box transcription factors and their roles in regulatory hierarchies in the developing heart. Development 132:4897–4910.

Stennard FA, Costa MW, Elliott DA, Rankin S, Haast SJ, Lai D, McDonald LP, Niederreither K, Dolle P, Bruneau BG, Zorn AM, Harvey RP. 2003. Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. Dev Biol 262:206–224.

Stennard FA, Costa MW, Lai D, Biben C, Furtado MB, Solloway MJ, McCulley DJ, Leimena C, Preis JI, Dunwoodie SL, Elliott DE, Prall OW, Black BL, Fatkin D, Harvey RP. 2005. Murine T-box transcription factor Tbx20 acts as a repressor during heart development, and is essential for adult heart integrity, function and adaptation. Development 132:2451–2462.

Szeto DP, Griffin KJ, Kimelman D. 2002. HrT is required for cardiovascular development in zebrafish. Development 129:5093–5101.

Takeuchi JK, Mileikovskaia M, Koshiba-Takeuchi K, Heidt AB, Mori AD, Arruda EP, Gertsenstein M, Georges R, Davidson L, Mo R, Hui CC, Henkelman RM, Nemer M, Black BL, Nagy A, Bruneau BG. 2005. Tbx20 dose-dependently regulates transcription factor networks required for mouse heart and motoneuron development. Development 132:2463–2474.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034.

Ware SM, Peng J, Zhu L, Fernbach S, Colicos S, Casey B, Towbin J, Belmont JW. 2004. Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. Am J Hum Genet 74: 93–105.

Yagi H, Furutani Y, Hamada H, Sasaki T, Asakawa S, Minoshima S, Ichida F, Joo K, Kimura M, Imamura S, Kamatani N, Momma K, Takao A, Nakazawa M, Shimizu N, Matsuoka R. 2003. Role of TBX1 in human del22q11.2 syndrome. Lancet 362:1366–1373.

### 2.2.3 Regulation of cardiac and skeletal muscle development by DPF3 – a novel epigenetic transcription factor

> Lange M, Kaynak B, Forster UB, Tonjes M, Fischer JJ, Grimm C, Schlesinger J, Just S, Dunkel I, Krueger T, Mebus S, Lehrach H, Lurz R, Gobom J, Rottbauer W, Abdelilah-Seyfried S, **Sperling S**. Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. *Genes Dev* 2008;**22**:2370-2384.

In the genome-wide gene expression study of congenital malformed human hearts, DPF3 was identified as characteristically upregulated in right ventricular myocardium of patients with Tetralogy of Fallot (Kaynak et al., 2003).

DPF3 is an evolutionary highly conserved member of the d4-protein family characterized by an N-terminal 2/3-domain unique to this protein family, a C2H2-type zinc finger and a C-terminal plant-homeodomain (PHD) zinc finger (Natalia et al., 2001). DPF3 gives rise to two splice variants (*DPF3a* and *DPF3b*) in human and mouse, four in chicken and one in zebrafish, with human and mouse DPF3 differing only by one amino acid. The human *DPF3b* variant and the *DPF3* full-length ortholog in zebrafish had not been identified previously and were cloned from human heart and zebrafish cDNA (AY803021, NM_001111169). DPF3 variants differ at the C-terminus such that DPF3a encodes a 357 amino acid protein containing a single truncated PHD finger, while DPF3b consists of 378 amino acids and a double PHD finger (**Figure 12**A).

The other members of the d4-family are DPF1 and DPF2. In the mouse, Dpf1 (Neud4) is expressed predominantly in brain and may have an important role in developing neurons through regulation of cell survival as a neurospecific transcription factor (Lessard et al., 2007). Dpf2 (ubi-d4/requiem) is ubiquitously expressed (Mertsalov et al., 2000) and implicated to be required for cell death after deprivation of trophic factors (Gabig et al., 1994).

Both splice variants of *DPF3* were found to be significantly upregulated in human right ventricular myocardial tissue of Tetralogy of Fallot (TOF) hearts compared to age and gender matched samples obtained from hearts with single ventricular septal defects as well as healthy donors (**Figure 12**B and C). TOF represents a defect in heart looping and outflow tract formation characterized by a ventricular septal defect, an overriding aorta, right ventricular outflow tract stenosis and right ventricular hypertrophy secondary to hemodynamic stress, mainly due to increased right ventricular systolic pressure. Using a multiple human tissue northern blot, *DPF3* was specifically expressed in cardiac and skeletal muscle (**Figure 12**D).

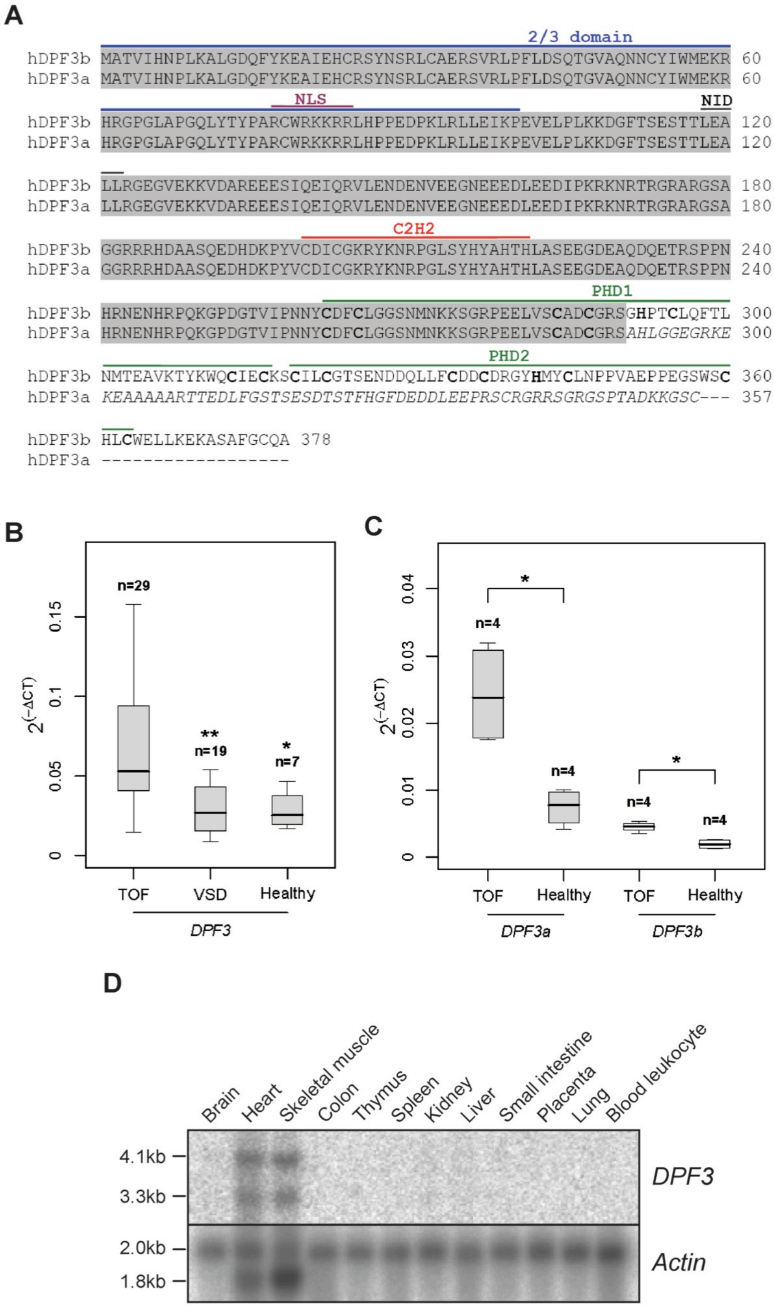


**Figure 12. DPF3 - a zinc and double PHD finger protein**. (A) Sequence conservation and divergence of human DPF3 isoforms. DPF3a (AAX20019.1) and DPF3b (NP_036206) contain an N-terminal 2/3 domain, a putative nuclear localization signal (NLS), a nuclear receptor interaction domain (NID) and a C2H2-Krüppel-like zinc finger. Note that the C-terminal double plant-homeodomain (PHD) is truncated in DPF3a. Cysteine and histidine residues of the plant-homeodomains are marked bold. (B, C) Expression of *DPF3* mRNA in malformed and normal human hearts. Real-time PCR analysis of *DPF3*mRNA levels in myocardial, right ventricular tissue from patients with Tetralogy of Fallot (TOF), ventricular septal defect (VSD) and healthy controls. Analysis of splice variant specific expression of *DPF3a* and *DPF3b* in TOF patients and healthy controls. Statistically significant differences analyzed by two-sided Wilcoxon test are indicated with asterisks (* p<0.01; ** p<0.01). Scale-bars represent ± SEM. (D) Tissue specific expression of *DPF3* mRNA in humans analyzed by Northern blot.

### *DPF3 associate with BAF chromatin remodelling complex and interacts with methylated and acetylated histones*

DPF3 contains two PHD fingers, domains frequently found in nuclear proteins whose substrate tend to be nucleosomes (Bienz, 2006). Using tandem affinity purification technique (TAP) and mass spectrometry the potential nuclear binding partners of DPF3a and DPF3b were identified. A very high percentage of proteins purified with DPF3 correspond to the BAF complex (91,2% BAF components with DPF3a and 86,8% with DPF3b as bait), such that nearly all core components could be isolated. Among the interactors of DPF3a and DPF3b was SMARCD3, the heart and somite specific subunit of the complex.

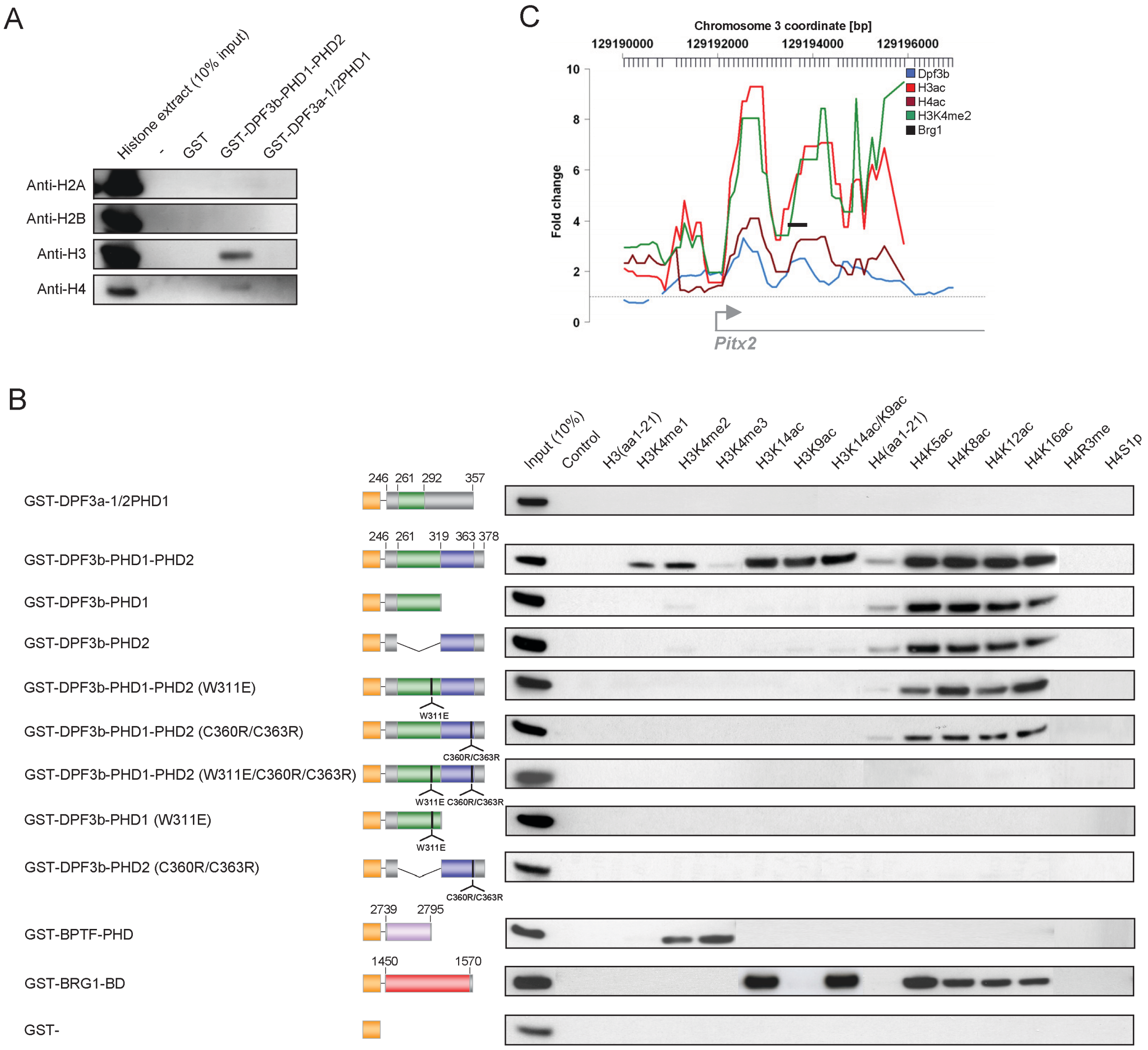


**Figure 13. The PHD fingers of DPF3b interact with modified histone tails on histone 3 and histone 4.** (A) Pulldown assays followed by western blotting and immunodetection of indicated histones using GST-DPF3 fusion proteins and calf thymus histone extracts. (B) Western blot analysis of histone peptide pulldowns with indicated GST-DPF3 fusion proteins and biotinylated peptides. GST-BPTF and GST-BRG1 fusion proteins are shown as positive / negative controls. Orange - Glutathione-S-Transferase tag (GST); green – DPF3 plant-homeodomain 1 (PHD1); blue – DPF3 plant-homeodomain 2 (PHD2); purple – BPTF plant homeodomain (PHD); red - BRG1 bromodomain**.** (C) Co-occurrence of Dpf3, BRG1, H3K4me2, and H3ac/H4ac modifications on the murine *Pitx2* locus. Normalized and smoothed relative ChIP-chip intensities and position of real-time PCR primer are shown.

It has recently become evident that proteins involved in chromatin remodelling recognize specific modifications on histone tails. The recognition of the methylation state of lysine residues on histone 3 and 4 has been shown to be mediated - among others - by the plant-homeodomain, whereas lysine acetylations are recognized by the bromodomain (Kouzarides, 2007). To address whether DPF3 generally binds to histones, a GST-pulldown system was used to test for the ability of recombinant full-length GST-DPF3 to pull down histones from calf thymus extracts followed by western analyses using histone specific antibodies against H2A, H2B, H3 and H4. DPF3b was able to pull down histones H3 and H4 but not histones H2A and H2B, whereas DPF3a did not bind any histones (**Figure 13**A). To further analyze if DPF3b binds specific histone modifications through its PHD fingers, a broad panel of histone 3 and 4 peptides harboring specific modifications was investigated such as methylations, acetylations or phosphorylations on different residues with pulldown assays (**Figure 13**B). Surprisingly, we observed specific binding of DPF3b to acetylated lysines on histone 3 and 4 (H3K14ac, H3K9ac, H4K5ac, H4K8ac, H4K12ac, H4K16ac) besides binding to mono and di-methylated lysine 4 on histone 3 (H3K4me1/2). Single DPF3-PHD fingers were sufficient for the interaction with lysine acetylations on histone 4, whereas histone 3 acetylations and methylations were only recognized by the double PHD finger (**Figure 13**B). Furthermore, DPF3a, which only contains a truncated PHD finger, did not bind any of the studied peptides. Point mutations of residues essential for the structural integrity of the aromatic cage formed by the PHD finger (W358E) as well as residues that contribute to zinc-complexing (C360R/C363R) lead to the abolishment of the binding respectively.

### *DPF3 expression pattern during embryonic development and its impact on heart and skeletal muscle development*

Whole mount in-situ hybridization in mouse embryos revealed cardiac and somite expression of *Dpf3* starting in the first differentiating cardiomyocytes of the cardiac crescent at E7.5 and in the first somites at E8.0 (**Figure 14**). Section in-situ hybridization revealed that *Dpf3* expression was restricted to the myocardial compartment of the heart. In order to analyze expression profiles of *Dpf3a* and *Dpf3b* during later stages of heart development, real-time PCR analysis was performed using cDNA obtained from embryonic hearts extracted between E9.5 and E16.5 as well as P0 and adult hearts. Expression of *Dpf3a and Dpf3b* was detectable from E9.5 onwards, although *Dpf3a* showed substantial higher expression till E11.5, where both splice variants subsequently reached a similar level of expression that remained stable until birth and adulthood. In-situ hybridization experiments in chicken and zebrafish embryos demonstrate an evolutionarily conserved expression pattern of *DPF3* orthologs.

The expression patterns of *Dpf1* and *Dpf2* were also analyzed by in-situ hybridization in mouse embryos. *Dpf1* was predominantly expressed in the developing brain, whereas *Dpf2* was ubiquitously expressed. This suggests that *Dpf3* is likely the only muscle-specific expressed d4 family member.

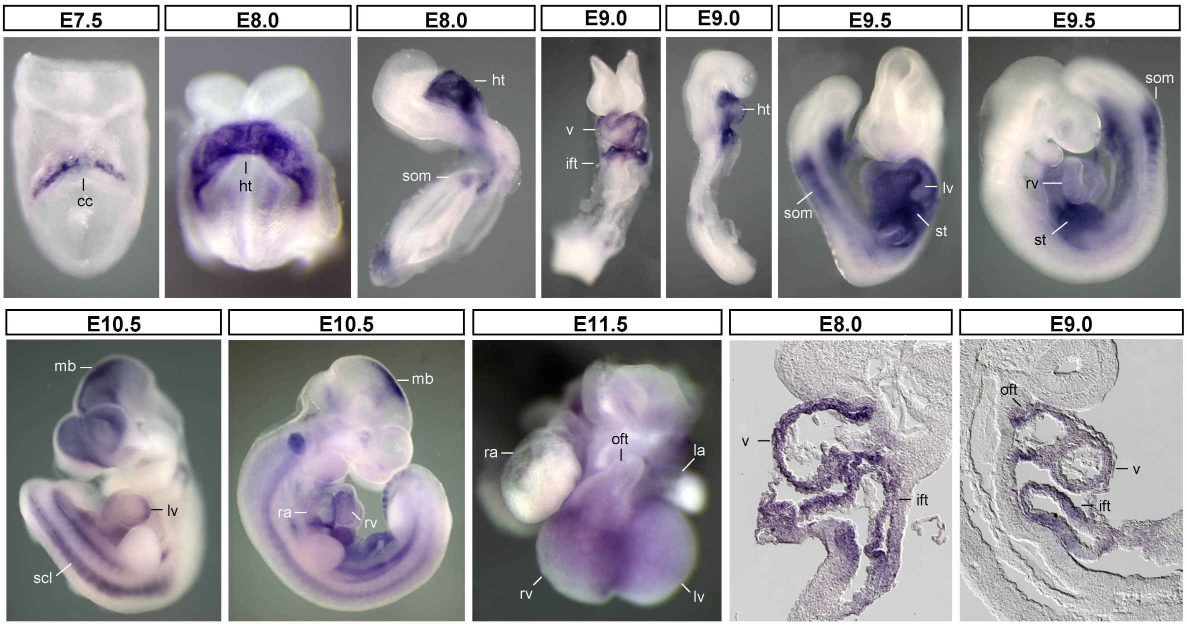


**Figure 14. Expression pattern of *Dpf3* mRNA during mouse development analyzed by in situ hybridization.** cc, cardiac crescent; ht, heart tube; v, ventricle; ift, inflow tract; oft, outflow tract; lv, left ventricle; rv, right ventricle; la, left atrium; ra, right atrium; som, somites; st, septum transversum; mb, midbrain; scl, sclerotome; a, atrium; nt, neural tube; v, ventricle; a, atrium.

Loss-of-function experiments of Dpf3 in zebrafish embryos and mouse muscle cells revealed an essential role of Dpf3 in muscle cell differentiation. In zebrafish *dpf3* Morpholino mediated knockdown lead to severely reduced cardiac contractility, incomplete cardiac looping and defective organization of cardiac and skeletal muscle fibers. Gene expression array analysis showed a marked transcriptional deregulation of structural and regulatory proteins. The set of upregulated genes contained many genes essential for transcriptional regulation, nucleosome assembly and metabolic processes, whereas genes involved in ion and electron transport were overrepresented among downregulated genes. A significantly increased expression of *cardiomyopathy associated 1* (*cmya1*, fold change 2.9) and of *actin binding protein 280-like* (*flncb*, fold change 2.5) was observed. Furthermore, the *heart and neural crest derivatives expressed 2* (*hand2*, fold change 0.5), *thymosin beta* (fold change 0.3), and a novel protein (zgc:101755) similar to mouse *actin filament capping protein of muscle Z-lines* (fold change 0.5) showed decreased expression. In morphant embryos, myofibrillay disarray, transversion of somite boundary by actin filaments, and disruption of somite boundary formation were frequently observed. In particular, the z-disc of sarcomeres representing the lateral boundaries where titin, nebulin, and the thin filaments are anchored, appeared to be affected. Transmission electron microscopy analysis revealed conservation of this phenotype in C2C12 mouse skeletal muscle cells with *Dpf3* siRNA knockdown.

### ***Mef2a regulates Dpf3 expression***

Mef2a deficient mice and zebrafish embryos are phenotypically similar to the observed myofibrillar disarray in *dpf3* knockdown embryos (Naya et al., 2002; Wang et al., 2005; Potthoff et al., 2007). Consequently, the *Dpf3* proximal promoter was screened for potential Mef2 binding sites. Within a conserved 1.2kbp promoter region three Mef2 matrices using TRANSFAC MATCH with stringent settings (Kel et al., 2003) were observed. Mef2a ChIP-chip analysis in mouse cardiomyocytes (HL-1 cells) showed a significant peak of Mef2a binding in the *Dpf3* promoter region that could also be confirmed by real-time PCR (1.8 fold change). Knockdown of *Mef2a* in HL-1 cells using two different siRNAs led to a reduction of *Dpf3* expression of up to 40%, demonstrating that Mef2a functionally binds the *Dpf3* promoter and activates its expression. Transcriptional regulation of *Dpf3* by Mef2a was also tested in luciferase reportergene assays using promoter fusion constructs of a previously characterized *DPF3* core promoter and 4 consecutive repeats of the putative Mef2 binding sites confirmed the regulation of Dpf3 by Mef2a.

Taken together, DPF3 contains the first plant-homeodomains known to bind acetylated as well as methylated histone residues. Furthermore DPF3 interacts with the BAF complex and its essential role for muscle development and function. Thus DPF3 potentially represents the missing link to explain the high impact of the histone modification status on the recruitment of the BAF complex to chromatin target sites and its consequence for cardiac function. It is tempting to speculate that DPF3a and DPF3b might serve as tissue-specific BAF subunits that regulate the transition of muscle precursors to differentiating myocytes. Moreover, it is highly suggestive that other PHD fingers might be capable to bind acetylation marks and play a yet unappreciated role in recruiting chromatin remodelling complexes.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

# Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex

**Martin Lange,[1] Bogac Kaynak,[1,8] Ulrike B. Forster,[2] Martje Tönjes,[1] Jenny J. Fischer,[1] Christina Grimm,[1] Jenny Schlesinger,[1] Steffen Just,[3] Ilona Dunkel,[1] Tammo Krueger,[1] Siegrun Mebus,[4] Hans Lehrach,[5] Rudi Lurz,[6] Johan Gobom,[7] Wolfgang Rottbauer,[3] Salim Abdelilah-Seyfried,[2] and Silke Sperling[1,9]**

[1]Group Cardiovascular Genetics, Department Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany; [2]Cell Polarity and Epithelial Development, Max Delbrück Center, Berlin 13125 Germany; [3]Molecular Cardiology, Ruprecht-Karls-Universität Heidelberg, Heidelberg 69120, Germany; [4]Department Pediatric Cardiology, German Heart Center Berlin, Berlin 13353, Germany; [5]Department Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany; [6]Microscopy Unit, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany; [7]Mass Spectrometry Group, Department Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany

**Chromatin remodeling and histone modifications facilitate access of transcription factors to DNA by promoting the unwinding and destabilization of histone–DNA interactions. We present DPF3, a new epigenetic key factor for heart and muscle development characterized by a double PHD finger. DPF3 is associated with the BAF chromatin remodeling complex and binds methylated and acetylated lysine residues of histone 3 and 4. Thus, DPF3 may represent the first plant homeodomains that bind acetylated lysines, a feature previously only shown for the bromodomain. During development Dpf3 is expressed in the heart and somites of mouse, chicken, and zebrafish. Morpholino knockdown of *dpf3* in zebrafish leads to incomplete cardiac looping and severely reduced ventricular contractility, with disassembled muscular fibers caused by transcriptional deregulation of structural and regulatory proteins. Promoter analysis identified *Dpf3* as a novel downstream target of Mef2a. Taken together, DPF3 adds a further layer of complexity to the BAF complex by representing a tissue-specific anchor between histone acetylations as well as methylations and chromatin remodeling. Furthermore, this shows that plant homeodomain proteins play a yet unexplored role in recruiting chromatin remodeling complexes to acetylated histones.**

[*Keywords*: Heart and skeletal muscle development and function; PHD finger; BAF chromatin remodeling complex; SMARCD3–BAF60; acetylated and methylated histones; Mef2]

Supplemental material is available at http://www.genesdev.org.

Received January 16, 2008; revised version accepted July 1, 2008.

Complex transcription networks mediate cell specification, proliferation, and differentiation throughout development and life. Coordinated activation and repression of different subsets of genes is regulated at several levels by genetic and epigenetic mechanisms. Genomic DNA is packaged into nucleosomes, the basic unit of chromatin structure formed by DNA wrapped around a histone octamer. Chromatin remodeling and covalent histone modifications facilitate DNA access for DNA-binding transcription factors (Simone 2006; Bernstein et al. 2007; Sperling 2007). Specific patterns of histone tail modifications attract or repel regulatory proteins of the chromatin remodeling complex. Histone modifications can influence one another and thus not just the level of modification but also the pattern may dictate biological outcome (Fischer et al. 2008).

The main histone modifications are acetylation and methylation. Recently, several transcription or remodeling factors (e.g., TFIID, BPTF, Yng2) have been identified, which bind to methylated histone lysine residues via different domains, such as WD-40, Tudor, MBT, and the plant homeodomain (PHD) (Kim et al. 2006; Ruthenburg

---

[8]**Present address: Gladstone Institute of Cardiovascular Disease, San Francisco, CA 94158, USA.**
[9]**Corresponding author.**
**E-Mail sperling@molgen.mpg.de; FAX 40-30-8413-1699.**
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.471408.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

et al. 2007; Vermeulen et al. 2007). Acetylation of histone lysine residues by histone acetyltransferases (HATs) stimulates gene expression by recruiting chromatin remodeling complexes and neutralizing positive charge, resulting in destabilization of histone–histone and histone–DNA interactions that limit access of transcription factors to DNA. The effect of HATs is counteracted by histone deacetylases (HDACs), and represents a control point of gene expression exemplified by cardiac growth in response to acute and chronic stress stimuli (Backs and Olson 2006). The recruitment of remodeling complexes is highly affected by histone acetylation and the bromodomain is the only protein domain that is presently known to recognize acetylated lysine residues of histones (Mujtaba et al. 2007). Surprisingly, bromodomains can be dispensable in vivo, which suggests functional redundancy among proteins (Elfring et al. 1998; Bourachot et al. 1999; Hassan et al. 2002; Mohrmann and Verrijzer 2005). Chromatin remodeling complexes use free energy derived from ATP hydrolysis to actively alter nucleosomal structure. These factors peel DNA from the edge of the nucleosomes forming a DNA loop or slide the histone octamer to a different position (Kassabov et al. 2003). Different chromatin remodeling complexes have been identified (e.g., SWR/NURF, CHD/NuRD, or SWI/SNF), which are defined by a unique subunit composition and the presence of a distinct ATPase (Palacios and Puri 2006; Simone 2006; Bao and Shen 2007). Mammalian SWI/SNF-like complexes (BAF complexes) are characterized by central core subunits BRG1 and BRM and 10 further subunit elements; e.g., SMARCD3 (BAF60c) representing a muscle-specific component. BRG1 and BRM contain an ATPase domain and a bromodomain that recognizes acetylated lysine in histone tails and other proteins (Sif 2004; Simone 2006). Thus, BRG1 acts as a ubiquitously expressed targeting molecule to anchor chromatin remodeling complexes on promoters with particular histone modification marks (Hassan et al. 2002, 2007). SMARCD3 is a promiscuous partner for several DNA-binding transcription factors, including nuclear receptors PPARγ, RXRα, RAR, and muscle regulatory factors like MEF2, MyoD, Nkx2.5, Tbx5, and Gata4 (Debril et al. 2004; Lickert et al. 2004; Palacios and Puri 2006; Simone 2006; Flajollet et al. 2007; Z.Y. Li et al. 2007). Tissue-specific transcription can be initiated by ligand-dependent activation of signaling cascades; e.g., phosphorylation of SMARCD3 and MEF2 through p38 MAP-kinase leads to translocation of MEF2 to the nucleus, potentially enhances their interaction, and finally, the BAF complex is targeted to muscle-specific loci (Simone et al. 2004; Rauch and Loughna 2005).

In the early mouse embryo Smarcd3 is specifically expressed in heart and somites, and is required for cardiac looping and outflow tract development. Smarcd3-deficient mice furthermore show impaired trabeculation of the heart and disorganized somites (Lickert et al. 2004; Takeuchi et al. 2007). The four Mef2 transcription factors (Mef2a, Mef2b, Mef2c, and Mef2d) regulate muscle cell differentiation, and can, in part, compensate each other's function (Karamboulas et al. 2006). Mef2s are DNA-binding transcription factors that interact with members of the MyoD family to cooperatively activate muscle specific genes. Embryonic hearts of Mef2a-deficient mice and zebrafish show myofibrillar disarray, and mice with skeletal muscle ablation of Mef2c form abnormally assembled sarcomeres (Naya et al. 2002; Wang et al. 2005; Potthoff et al. 2007).

In a genome-wide gene expression study of congenital malformed human hearts we identified *DPF3* as significantly up-regulated in the right ventricular myocardium of patients with Tetralogy of Fallot (TOF) (Kaynak et al. 2003). The study showed disease-associated expression profiles for a panel of cardiac conditions in addition to profiles specific for each cardiac chamber of the normal human heart. DPF3 contains a double PHD finger containing protein and a putative transcription factor. We show that DPF3 is associated with the BAF complex, and binds methylated and acetylated lysine residues of histone 3 and 4. Thus, DPF3 contains the first PHD that binds acetylated lysines, a feature previously only shown for bromodomains. Furthermore, *Dpf3* shows tissue-specific expression in heart and somites during development of mouse, chicken, and zebrafish. Promoter analysis identified *Dpf3* as a novel downstream target of Mef2a. Morpholino (MO) knockdown of *dpf3* in zebrafish lead to severely reduced cardiac contractility, incomplete cardiac looping and defective organization of cardiac and skeletal muscle fibers caused by transcriptional deregulation of structural and regulatory proteins essential for muscle fibers. Taken together, DPF3 adds a further layer of complexity to the BAF complex by representing a tissue-specific anchor between histone acetylations as well as methylations and chromatin remodeling.

## Results

### *DPF3 is a muscle expressed member of the D4, zinc, and double PHD finger family*

DPF3 is an evolutionary highly conserved member of the d4-protein family characterized by an N-terminal 2/3 domain unique to this protein family, a C2H2-type zinc finger, and a C-terminal PHD zinc finger (Supplemental Table S1; Natalia et al. 2001). *DPF3* gives rise to two splice variants (*DPF3a* and *DPF3b*) in human and mouse, four in chicken, and one in zebrafish, with human and mouse DPF3 differing only by one amino acid (Supplemental Fig. S1). The human *DPF3b* variant and the *DPF3* full-length ortholog in zebrafish had not been identified previously, and were cloned from human heart and zebrafish cDNA (AY803021, NM_001111169). DPF3 variants differ at the C terminus such that DPF3a encodes a 357-amino-acid protein containing a single truncated PHD finger, while DPF3b consists of 378 amino acids and a double PHD finger (Fig. 1A).

The other members of the d4 family are DPF1 and DPF2. In the mouse, Dpf1 (Neud4) is expressed predominantly in the brain, and may have an important role in developing neurons through regulation of cell survival as

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

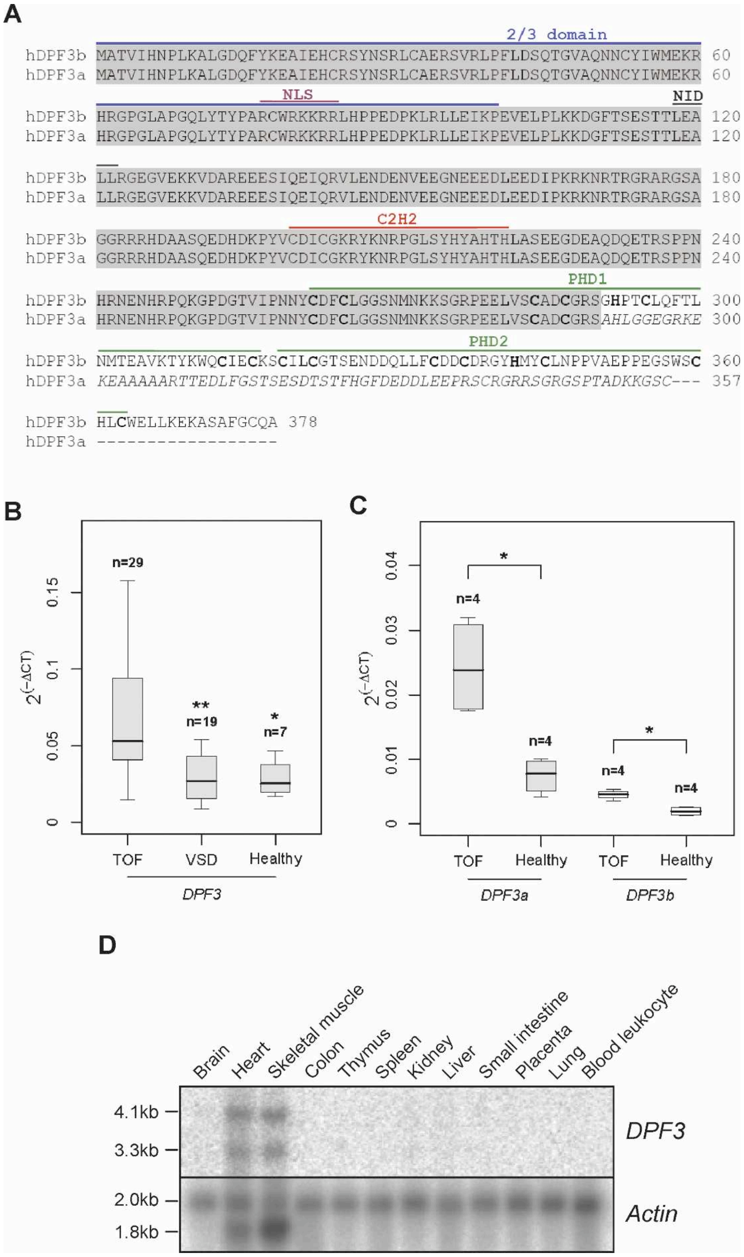


**Figure 1.** DPF3—a zinc and double PHD finger protein. (*A*) Sequence conservation and divergence of human DPF3 isoforms. DPF3a (AAX20019.1) and DPF3b (NP_036206) contain an N-terminal 2/3 domain, a putative nuclear localization signal (NLS), a nuclear receptor interaction domain (NID), and a C2H2-Krüppel-like zinc finger. Note that the C-terminal double PHD is truncated in DPF3a. Cysteine and histidine residues of the PHDs are marked in bold. (*B*,*C*) Expression of *DPF3* mRNA in malformed and normal human hearts. Real-time PCR analysis of *DPF3* mRNA levels in myocardial, right ventricular tissue from patients with TOF, ventricular septal defect (VSD), and healthy controls. Analysis of splice variant-specific expression of *DPF3a* and *DPF3b* in TOF patients and healthy controls. Expression values normalized to the housekeeping gene HPRT. Statistically significant differences analyzed by two-sided Wilcoxon test are indicated with asterisks ([*] $P < 0.01$; [**] $P < 0.01$). Scale bars represent ±SEM. (*D*) Tissue-specific expression of *DPF3* mRNA in humans analyzed by Northern blot. *DPF3* mRNA expression is restricted to heart and skeletal muscle. The blot containing mRNA from the indicated tissues was probed with $^{32}$P-*DPF3* cDNA (*top* panel), stripped, and reprobed with $^{32}$P-*Actin* cDNA (*bottom* panel).

a neurospecific transcription factor (Lessard et al. 2007). Dpf2 (ubi-d4/requiem) is ubiquitously expressed (Mertsalov et al. 2000) and implicated to be required for cell death after deprivation of trophic factors (Gabig et al. 1994).

We found both splice variants of *DPF3* to be significantly up-regulated in human right ventricular myocardial tissue of TOF hearts compared with age- and gender-matched samples obtained from hearts with single ventricular septal defects as well as healthy donors (Fig. 1B,C). TOF represents a defect in heart looping and outflow tract formation characterized by a ventricular septal defect, an overriding aorta, right ventricular outflow tract stenosis and right ventricular hypertrophy secondary to hemodynamic stress, mainly due to increased right ventricular systolic pressure. Using a multiple human tissue Northern blot we observed *DPF3* to be specifically expressed in cardiac and skeletal muscle (Fig. 1D).

## *DPF3a and DPF3b associate with BAF chromatin remodeling complexes*

DPF3 contains two PHD fingers, domains frequently found in nuclear proteins whose substrate tend to be nucleosomes (Bienz 2006). Using tandem affinity purification technique (TAP) and mass spectrometry we isolated potential nuclear binding partners of DPF3a and DPF3b in HEK293T cells. We identified nearly all core

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

components of the BAF chromatin remodeling complex to be associated with both isoforms of DPF3 (Table 1). We found that a very high percentage of proteins purified with DPF3 correspond to the BAF complex (91.2% BAF components with DPF3a and 86,8% with DPF3b as bait). Among the interactors of DPF3a and DPF3b we found SMARCD3, a heart and somite-specific subunit of the complex. To confirm the association of DPF3 with the BAF complex, we performed reverse-TAP and mass spectrometry using SMARCD3 as bait (Table 1). Thus, both DPF3 isoforms associate with the BAF chromatin remodeling complex.

### *DPF3 interacts with methylated and acetylated lysine residues of histones 3 and 4*

It has recently become evident that proteins involved in chromatin remodeling recognize specific modifications on histone tails. The recognition of the methylation state of lysine residues on histone 3 and 4 has been shown to be mediated, among others, by the PHD, whereas lysine acetylations are recognized by the bromodomain (Kouzarides 2007). To address whether DPF3 generally binds to histones, we used a glutathione-S-transferase (GST) pull-down system and tested for the

**Table 1.** *Human DPF3 protein interactions*

| Bait | Alias | HUGO ID | MW (kDa) | Length (amino acids) | Mascot score | Spectral counts | Sequence coverage (%) | NSAF | Accession number |
|---|---|---|---|---|---|---|---|---|---|
| DPF3a | BAF250A | ARID1A | 242.8 | 2285 | 1844 | 47 | 24 | 0.038 | O14497 |
| | BAF250B | ARID1B | 237.1 | 2236 | 1171 | 33 | 14 | 0.027 | Q8NFD5 |
| | BRG1 | SMARCA4 | 185.0 | 1647 | 2196 | 51 | 29 | 0.057 | P51532 |
| | BRM | SMARCA2 | 181.3 | 1586 | 959 | 27 | 16 | 0.031 | P51531 |
| | BAF170 | SMARCC2 | 133.2 | 1214 | 2160 | 48 | 30 | 0.073 | Q8TAQ2 |
| | BAF155 | SMARCC1 | 123.2 | 1105 | 2401 | 49 | 38 | 0.081 | Q92922 |
| | BAF60A | SMARCD1 | 55.2 | 476 | 80 | 4 | 11 | 0.014 | Q96GM5 |
| | BAF60B | SMARCD2 | 52.7 | 456 | 846 | 23 | 44 | 0.093 | Q92925 |
| | BAF60C | SMARCD3 | 55.2 | 483 | 644 | 19 | 36 | 0.072 | Q6STE5 |
| | BAF57 | SMARCE1 | 46.7 | 411 | 982 | 19 | 49 | 0.085 | Q969G3 |
| | BAF53 | ACTL6A | 47.9 | 429 | 826 | 17 | 39 | 0.073 | O96019 |
| | BAF47 | SMARCB1 | 44.4 | 385 | 659 | 14 | 39 | 0.067 | Q12824 |
| | β-actin | ACTB | 42.1 | 375 | 697 | 18 | 38 | 0.088 | P60709 |
| | CERD4 | DPF3 | 26.1 | 224 | 844 | 15 | 63 | 0.123 | Q92784 |
| DPF3b | BAF250A | ARID1A | 242.8 | 2285 | 1461 | 49 | 27 | 0.036 | O14497 |
| | BAF250B | ARID1B | 237.1 | 2236 | 457 | 16 | 9 | 0.012 | Q8NFD5 |
| | BRG1 | SMARCA4 | 185.0 | 1647 | 1912 | 43 | 22 | 0.044 | P51532 |
| | BRM | SMARCA2 | 181.3 | 1586 | 792 | 21 | 12 | 0.022 | P51531 |
| | BAF170 | SMARCC2 | 133.2 | 1214 | 1676 | 42 | 28 | 0.058 | Q8TAQ2 |
| | BAF155 | SMARCC1 | 123.2 | 1105 | 1511 | 42 | 29 | 0.064 | Q92922 |
| | BAF60A | SMARCD1 | 55.2 | 476 | 1062 | 26 | 45 | 0.091 | Q96GM5 |
| | BAF60B | SMARCD2 | 52.7 | 456 | 957 | 21 | 43 | 0.077 | Q92925 |
| | BAF60C | SMARCD3 | 55.2 | 483 | 616 | 19 | 37 | 0.066 | Q6STE5 |
| | BAF57 | SMARCE1 | 46.7 | 411 | 969 | 19 | 44 | 0.077 | Q969G3 |
| | BAF53 | ACTL6A | 47.9 | 429 | 699 | 15 | 35 | 0.059 | O96019 |
| | BAF47 | SMARCB1 | 44.4 | 385 | 880 | 18 | 56 | 0.078 | Q12824 |
| | β-actin | ACTB | 42.1 | 375 | 604 | 18 | 43 | 0.080 | P60709 |
| | CERD4 | DPF3 | 26.1 | 224 | 909 | 16 | 63 | 0.120 | Q92784 |
| BAF60c | BAF250A | ARID1A | 242.8 | 2285 | 680 | 46 | 23 | 0.046 | O14497 |
| | BAF250B | ARID1B | 237.1 | 2236 | 375 | 30 | 12 | 0.031 | Q8NFD5 |
| | BAF180 | PBRM1 | 194.1 | 1689 | 91 | 9 | 5 | 0.012 | Q86U86 |
| | BRG1 | SMARCA4 | 185.0 | 1647 | 523 | 39 | 21 | 0.054 | P51532 |
| | BRM | SMARCA2 | 181.3 | 1586 | 195 | 20 | 10 | 0.029 | P51531 |
| | BAF170 | SMARCC2 | 133.2 | 1214 | 656 | 35 | 25 | 0.066 | Q8TAQ2 |
| | BAF155 | SMARCC1 | 123.2 | 1105 | 1069 | 42 | 36 | 0.087 | Q92922 |
| | BAF60B | SMARCD2 | 52.7 | 456 | 91 | 5 | 9 | 0.025 | Q92925 |
| | BAF60C | SMARCD3 | 55.2 | 483 | 202 | 14 | 19 | 0.066 | Q6STE5 |
| | BAF57 | SMARCE1 | 46.7 | 411 | 711 | 23 | 38 | 0.127 | Q969G3 |
| | BAF53 | ACTL6A | 47.9 | 429 | 258 | 15 | 36 | 0.080 | O96019 |
| | BAF47 | SMARCB1 | 44.4 | 385 | 195 | 14 | 39 | 0.083 | Q12824 |
| | β-actin | ACTB | 42.1 | 375 | 398 | 13 | 37 | 0.079 | P60709 |
| | CERD4 | DPF3 | 26.1 | 224 | 72 | 2 | 11 | 0.021 | Q92784 |

Peptides associated with DPF3a, DPF3b, and BAF60c identified by TAP and mass spectrometry. (MW) Calculated molecular weight; (NSAF) Normalized spectral abundance factor (Florens et al. 2006).

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

ability of recombinant full-length GST-DPF3 to pull down histones from calf thymus extracts followed by Western analyses using histone specific antibodies against H2A, H2B, H3, and H4. DPF3b was able to pull down histones H3 and H4 but not histones H2A and H2B, whereas DPF3a did not bind any histones (Fig. 2A). To further analyze if DPF3b binds specific histone modifications through its PHD fingers, we tested a broad panel of histone 3 and 4 peptides harboring specific modifications such as methylations, acetylations, or phosphorylations on different residues with pull-down assays (Fig. 2B). Surprisingly, we observed specific binding of DPF3b to acetylated lysines on histone 3 and 4 (H3K14ac, H3K9ac, H4K5ac, H4K8ac, H4K12ac, H4K16ac) besides binding to mono- and dimethylated Lys 4 on histone 3 (H3K4me1/2). Unmodified histone 3 and 4 and other modifications were detected at the background level.

Since DPF3b contains a double PHD finger, we asked whether the PHD1 or PHD2 alone is sufficient to recognize histone lysine modifications. Pull-down assays revealed that single DPF3-PHD fingers were sufficient for the interaction with lysine acetylations on histone 4, whereas histone 3 acetylations and methylations were only recognized by the double PHD finger (Fig. 2B). Furthermore, DPF3a, which only contains a truncated PHD finger, did not bind any of the studied peptides. To substantiate these findings, we generated point mutations of residues essential for the structural integrity of the aromatic cage formed by the PHD finger (W358E) as well as residues that contribute to zinc-complexing (C360R/C363R). These mutations lead to the abolishment of single and double PHD finger binding to H3 and H4 modified peptides showing the specificity of the interactions (Fig. 3B). The binding properties of DPF3-PHD fingers were furthermore compared with the known methyl- and acetyllysine recognition of the BPTF-PHD finger and the BRG1 bromodomain.

### *Mapping of DPF3-binding sites reveals global colocalization with histone modifications by chromatin immunoprecipitation (ChIP)–chip*

To obtain a global overview of potential downstream targets of DPF3, we used ChIP followed by array detection

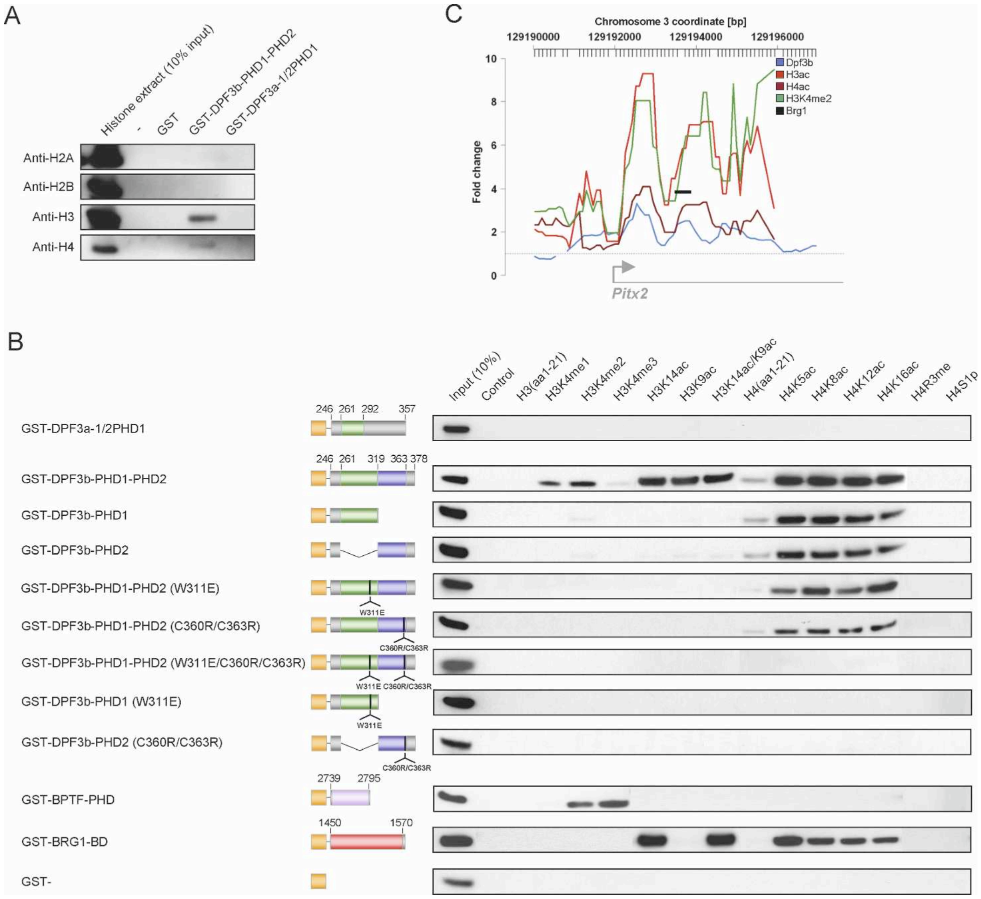


**Figure 2.** The PHD fingers of DPF3b interact with modified histone tails on histone 3 and histone 4. (*A*) Pull-down assays followed by Western blotting and immunodetection of indicated histones using GST-DPF3 fusion proteins and calf thymus histone extracts. (*B*) Western blot analysis of histone peptide pulldowns with indicated GST-DPF3 fusion proteins and biotinylated peptides. GST-BPTF and GST-BRG1 fusion proteins are shown as positive/negative controls. (Orange) GST tag; (green) DPF3-PHD1; (blue) DPF3-PHD2; (purple) BPTF-PHD; (red) BRG1 bromodomain. (*C*) Co-occurrence of Dpf3, BRG1, H3K4me2, and H3ac/H4ac modifications on the murine *Pitx2* locus. Normalized and smoothed relative ChIP–chip intensities and position of real-time PCR primer are shown.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

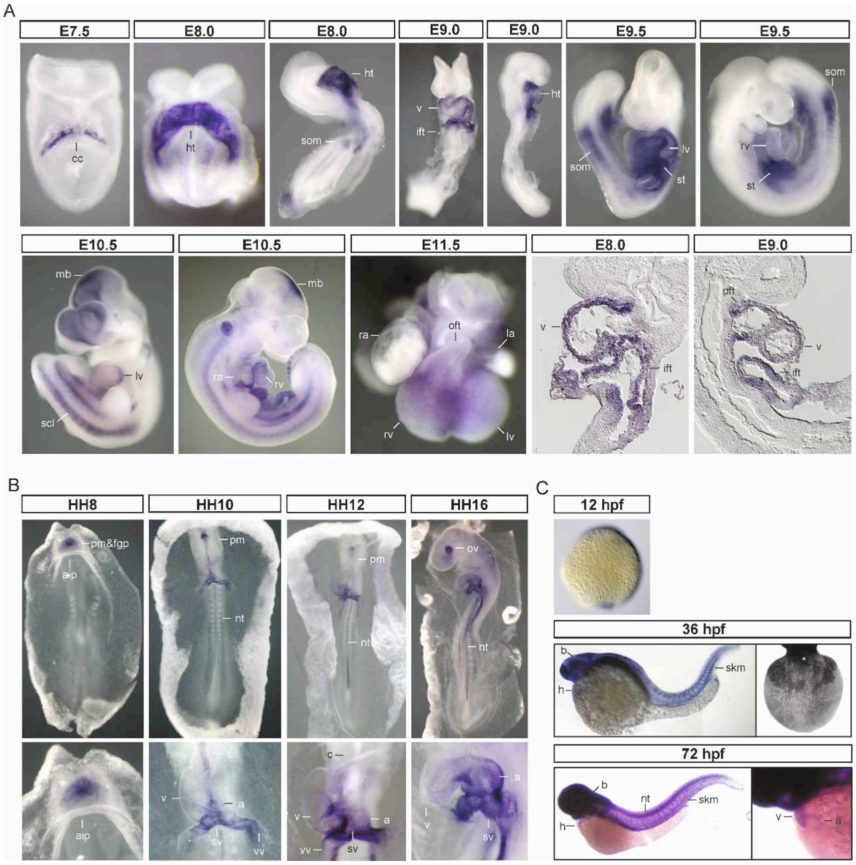


**Figure 3.** Expression patterns of *Dpf3* mRNA during embryonic development analyzed by in situ hybridization. Expression pattern of *Dpf3* mRNA during mouse (*A*), chicken (*B*), and zebrafish (*C*) development is shown (ventral and lateral views and closeups). (cc) Cardiac crescent; (ht) heart tube; (v) ventricle; (ift) inflow tract; (oft) outflow tract; (lv) left ventricle; (rv) right ventricle; (la) left atrium; (ra) right atrium; (som) somites; (st) septum transversum; (mb) midbrain; (scl) sclerotome; (pm) prechordal mesoderm; (fgp) foregut pocket; (aip) anterior intestinal portal; (sv) sinus venosus; (vv) vitelline veins; (a) atrium; (nt) neural tube; (c) conus; (h) heart; (v) ventricle; (skm) skeletal muscle; (ov) optic vesicle.

(ChIP–chip) and mapped the genomic localization of DPF3-binding sites in C2C12 skeletal muscle cells. We designed a custom muscle specific promoter array with 740,000 probes covering 10 kb upstream of and 3 kb downstream from ~12,000 transcripts. This array enabled analysis of our genes of interest with a much higher degree of tiling and sequence coverage than standard whole-genome arrays would provide. We found a total of 1201 transcripts in close distance to DPF3a- or DPF3b-binding sites (460 and 979 respectively; 238 shared) (Supplemental Table S2). To gain first insight into the functional role of downstream targets of DPF3, we performed an analysis of GO terms and found that DPF3 targets particularly play a role in cell proliferation, nucleosome assembly, and chromatin remodeling (Supplemental Table S3). Moreover, DPF3b targets are especially important for cardiovascular development and cytoskeleton organization. A number of DPF3 targets are structural genes, like α *actinin* (*Actn1*), *cardiomyopathy associated 3* (*Cmya3*), *myosin light chain* (*Myl1*), and *troponin C* (*Tnnc1*); muscle-regulating transcription factors and cofactors such as *myocyte enhancer factor* (*Mef2c/d*), *Cbp/p300-interacting transactivator* (*Cited2*), *paired-like homeodomain transcription factor 2* (*Pitx2*), *four and a half LIM domains 2* (*Fhl2*), *inhibitor of DNA binding 2* (*Id2*); as well as genes essential for muscle differentiation like *bone morphogenetic protein 2* (*Bmp2*).

Figure 2C exemplifies observed binding sites of DPF3b in the vicinity of target genes such as *Pitx2*. Moreover, *Pitx2* represents an example of co-occurrence of DPF3b-binding sites with acetylated/methylated lysine residues of histone 3 and 4, which have been analyzed by us previously (Fischer et al. 2008). To gain insight into the frequency and relevance of DPF3b binding to histone 3 and 4 modification marks, we compared the two ChIP–chip data sets. Out of 546 DPF3b-binding sites, 265 overlapped with histone 3 acetylation, 220 overlapped with histone 4 acetylation, and 294 overlapped with histone 3 methylation marks. Thus, 66% of DPF3b-binding sites overlap with acetylation marks and 54% with methylation marks, which is significantly more than one could expect from random permutations (mininum 26%, maximum 39%).

### *Co-occurrence of modified histones with DPF3 and BRG1 genomic binding sites*

Considering that DPF3 is a member of the BAF chromatin remodeling complex and binds modified histones, we additionally analyzed the co-occurrence of DPF3-binding sites with those of BRG1, a core component of the BAF complex. To select potentially shared targets, we performed ChIP–chip analysis for BRG1 in C2C12 cells (data not shown). Using real-time PCR we screened 21 muscle relevant downstream targets for co-occurring binding sites of modified histones, DPF3 and BRG1, and observed a high degree of overlap (Table 2). This suggests

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

**Table 2.** *Co-occurrence of DPF3b with histone modifications and BRG1 binding*

| Gene | | Peak position | | | Brg1 | | Dpf3b | | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Ensemble transcript ID | Chr. | Start | End | Fold change | SD | Fold change | SD | Histone modifications |
| Jmjd1c | ENSMUST00000095573 | 10 | 66581196 | 66581787 | 18.00 | 0.12 | 7.12 | 0.10 | H3ac |
| Ctnnb1 | ENSMUST00000007130 | 9 | 120783627 | 120783820 | 5.84 | 0.09 | 3.79 | 0.08 | H3ac, H3K4me2 |
| Musk | ENSMUST00000098059 | 4 | 58380142 | 58380250 | 5.25 | 0.17 | 3.27 | 0.06 | H3ac, H3K4me2 |
| Flrt2 | ENSMUST00000057324 | 12 | 96093975 | 96094272 | 3.81 | 0.13 | 2.75 | 0.04 | H3ac, H3K4me2 |
| Gsk3b | ENSMUST00000023507 | 16 | 38010138 | 38010438 | 6.07 | 0.14 | 2.54 | 0.04 | H3ac, H3K4me2 |
| Cald1 | ENSMUST00000079391 | 6 | 34529598 | 34530598 | 11.44 | 0.19 | 2.36 | 0.12 | H3ac, H3K4me2 |
| Pten | ENSMUST00000013807 | 19 | 32825230 | 32825724 | 2.16 | 0.14 | 2.16 | 0.30 | H3ac, H3K4me2 |
| Creb1 | ENSMUST00000049932 | 1 | 64468558 | 64468747 | 2.09 | 0.12 | 1.84 | 0.05 | H3ac, H3K4me2 |
| Arpc2 | ENSMUST00000006467 | 1 | 74172324 | 74172916 | 3.72 | 0.18 | 1.50 | 0.10 | H3ac, H3K4me2 |
| Sema3a | ENSMUST00000095012 | 5 | 13405946 | 13406745 | 4.50 | 0.87 | 5.67 | 0.15 | H3ac, H4ac |
| Zeb2 | ENSMUST00000028229 | 2 | 44933122 | 44933520 | 6.23 | 0.30 | 6.71 | 0.20 | H3ac, H4ac, H3K4me2 |
| Trim23 | ENSMUST00000022225 | 13 | 105298442 | 105298742 | 49.82 | 0.14 | 4.82 | 0.15 | H3ac, H4ac, H3K4me2 |
| Pitx2 | ENSMUST00000029657 | 3 | 129193542 | 129193945 | 3.97 | 0.05 | 4.77 | 0.08 | H3ac, H4ac, H3K4me2 |
| Asb5 | ENSMUST00000033918 | 8 | 56048828 | 56049632 | 13.41 | 0.48 | 4.05 | 0.11 | H3ac, H4ac, H3K4me2 |
| Foxp1 | ENSMUST00000074346 | 6 | 99060857 | 99061758 | 10.99 | 0.02 | 2.94 | 0.11 | H3ac, H4ac, H3K4me2 |
| Csrp2 | ENSMUST00000020403 | 10 | 110335474 | 110335974 | 8.03 | 0.17 | 2.06 | 0.09 | H3ac, H4ac, H3K4me2 |
| Cxcr7 | ENSMUST00000065587 | 1 | 92036250 | 92036550 | 22.29 | 0.29 | 1.93 | 0.23 | H3ac, H4ac, H3K4me2 |
| Daam1 | ENSMUST00000085299 | 12 | 72801340 | 72801838 | 10.00 | 0.09 | 2.73 | 0.12 | H3K4me2 |
| Igfbp5 | ENSMUST00000027377 | 1 | 72811484 | 72812274 | 4.59 | 0.05 | 2.31 | 0.04 | H4ac |
| Lamc1 | ENSMUST00000027752 | 1 | 155062722 | 155063616 | 1.51 | 0.18 | 2.01 | 0.09 | H4ac |
| Mtss1 | ENSMUST00000080371 | 15 | 58894578 | 58895274 | 2.43 | 0.17 | 1.64 | 0.10 | H4ac |

Real-time PCR analysis showing cobinding of DPF3b and BRG1 at genomic sites that are further characterized by histone modifications. (SD) Standard deviation.

that DPF3 potentially serves as an anchor between the BAF complex and modified histones.

### *Dpf3 expression patterns during embryonic development*

As *DPF3* was up-regulated in hypertrophic cardiac tissue of TOF patients, we were interested in its spatiotemporal expression pattern during embryogenesis and performed in situ hybridization in mouse, chicken, and zebrafish embryos. Whole-mount in situ hybridization in mouse embryos revealed cardiac and somite expression of *Dpf3a* starting in the first differentiating cardiomyocytes of the cardiac crescent at embryonic day 7.5 (E7.5) and in the first somites at E8.0 (Fig. 3A). A detailed description is provided in the Supplemental Material. Section in situ hybridization revealed that *Dpf3a* expression was restricted to the myocardial compartment of the heart (Fig. 3A). Further in situ hybridizations using a common *Dpf3* probe revealed a similar expression pattern (data not shown).

In order to analyze expression profiles of *Dpf3a* and *Dpf3b* during later stages of heart development, real-time PCR analysis was performed using cDNA obtained from embryonic hearts extracted between E9.5 and E16.5 as well as P0 and adult hearts. Expression of *Dpf3a* and *Dpf3b* was detectable from E9.5 onward, although *Dpf3a* showed substantially higher expression until E11.5, where both splice variants subsequently reached a similar level of expression that remained stable until birth and adulthood (Supplemental Fig. S2).

The expression patterns of *Dpf1* and *Dpf2* were also analyzed by in situ hybridization in mouse embryos. *Dpf1* was predominantly expressed in the developing brain, whereas *Dpf2* was ubiquitously expressed (data not shown).

In situ hybridization experiments in chicken embryos using a probe targeting all splice variants of *Dpf3* showed conservation of the mouse *Dpf3* expression pattern (Fig. 3B; see the Supplemental Material for a detailed description). In zebrafish embryos, *dpf3* was strongly expressed within the developing brain and throughout somitic tissues along the entire length of the embryonic trunk and tail shown by in situ hybridization at 36 and 72 h postfertilization (hpf) (Fig. 3C). Within the heart, *dpf3* was strongly expressed in the ventricle and faintly in the atria. In the early embryo at 12 hpf, dpf3 is expressed unspecifically. (Fig. 3C). Expression of *dpf2* at 36 hpf is within the developing brain and spinal cord (data not shown), and in contrast to *dpf3* was not detected in heart or somites. This suggests that *dpf3* is likely the only muscle expressed d4 family member. Taken together, these data demonstrate an evolutionarily conserved expression pattern of *DPF3* orthologs.

### *Knockdown of dpf3 reveals its essential role for heart and skeletal muscle development in vivo*

To address the role of *dpf3* in vivo, we performed MO antisense oligonucleotide-mediated knockdown in zebrafish. We characterized embryos injected with MO$^{dpf3}$, which targets the exon4–intron4 boundary of *dpf3* pre-mRNA and blocks correct splicing. The specificity of the MO$^{dpf3}$ was demonstrated by coinjection of synthetic

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

and mature *dpf3* mRNA, resulting in rescue of the $\mathrm{MO}^{dpf3}$ phenotypes (Fig. 4). Efficacy of the $\mathrm{MO}^{dpf3}$ was tested by PCR, which showed that the majority of *dpf3* mRNA was incorrectly spliced leading to two truncated proteins (Fig. 4A; Supplemental Material).

To assess cardiac morphogenesis and differentiation,

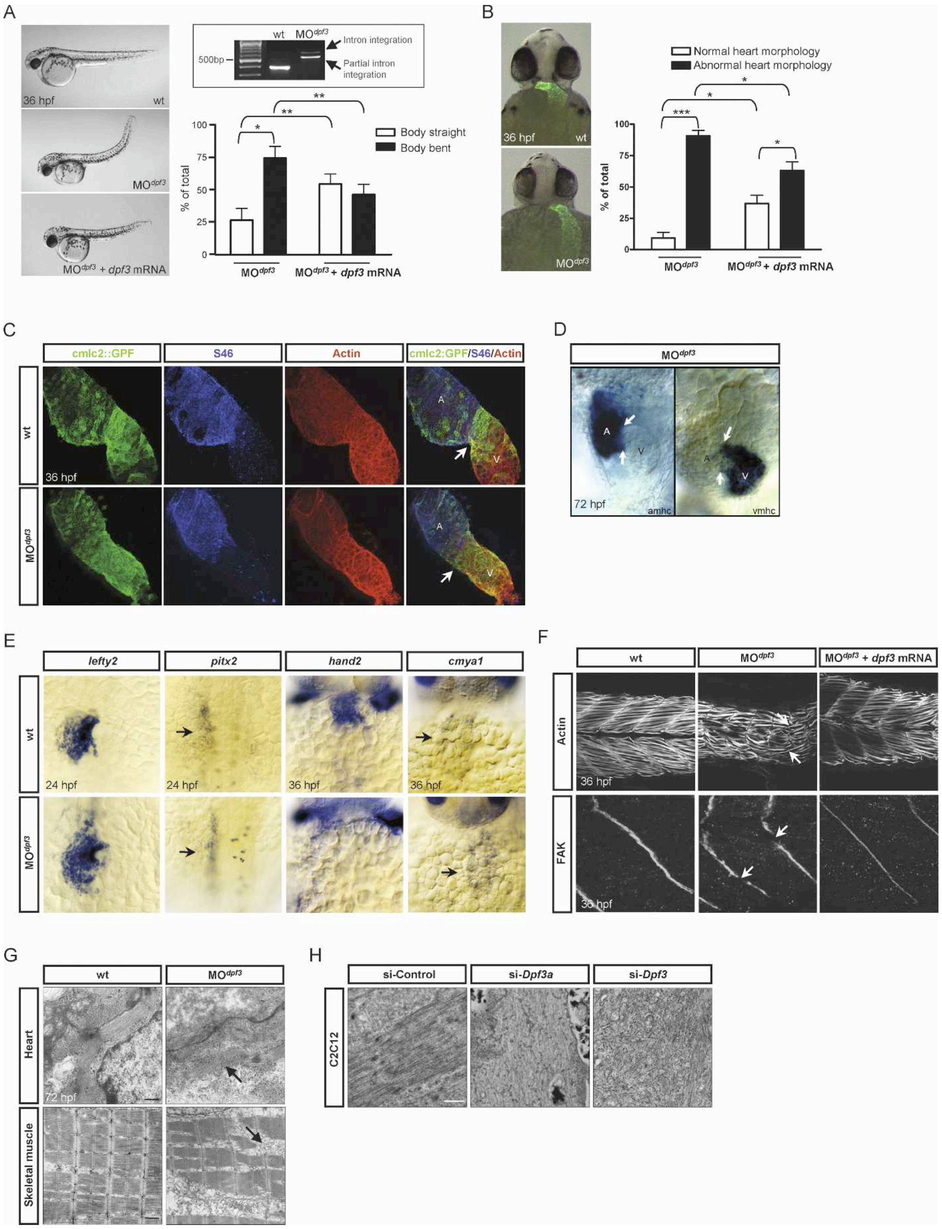


**Figure 4.** Knockdown of *dpf3* in zebrafish and in C2C12 mouse skeletal muscle cells analysis of body and heart morphology in embryos injected with $\mathrm{MO}^{dpf3}$ and controls at 36 hpf. (*A*) Knockdown of *dpf3* lead to abnormal body posture (curved tail) in zebrafish embryos. The phenotype could be rescued by coinjection of mature *dpf3*-mRNA. (Five independent experiments; embryos scored: 255 MO, 321 MO + rescue RNA; [*] $P < 0.05$; [**] $P < 0.01$.) (*B*) Analysis of heart morphology in [*Tg(cmlc::GFP)*] zebrafish embryos. Knockdown of *dpf3* lead to abnormal heart looping, which could be rescued by coinjection of mature *dpf3*-mRNA. (Three independent experiments; embryos scored: 101 MO, 145 MO + rescue RNA; [*] $P < 0.05$; [***] $P < 0.001$.) Statistical significance analyzed by two-way ANOVA with Bonferroni post hoc testing. Scale bars represent ±SEM. (*C*) Analysis of isolated zebrafish hearts by confocal microscopy. Cmlc marks all cardiomyocytes, S46 labels, the cells of the atrium, actin is predominantly expressed in the ventricle at 36 hpf. (*D*) In situ hybridization of chamber-specific markers *amhc* and *vmhc*. (*E*) In situ hybridization of left–right asymmetry marker *lefty2*, and *pitx2* as well as of differentially expressed genes *hand2* and *cmya1*. (*F*) Analysis of skeletal muscle in wild-type, *dpf3* morphant, and rescued embryos by immunohistochemistry at 36 hpf. (*Top* panel) Actin staining shows myofibrillar disarray and transversion of somite boundaries. (*Bottom* panel) FAK staining reveals disruption of somite boundaries in *dpf3* morphants. (*G*) Disrupted sarcomere integrity in heart and skeletal muscle of *dpf3* morphant embryos at 72 hpf shown by electron microscopy. Bar, 500 nm. (*H*) siRNA-mediated knockdown of *Dpf3a* and *Dpf3* lead to defects in myofibrillar assembly in C2C12 mouse skeletal muscle cells analyzed by electron microscopy. Bar, 500 nm.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

we used the $\mathrm{MO}^{dpf3}$ in a transgenic line of zebrafish that expresses green fluorescent protein (GFP) under control of the *cardiac myosin light chain 2* (*cmlc2*) promoter region [*Tg(cmlc2:GFP)*]. Injection of $\mathrm{MO}^{dpf3}$ at the one-cell stage resulted in 91% of embryos with abnormal heart morphology ($n = 101$) and in 74% of embryos with a curved tail at 36 hpf ($n = 255$) (Fig. 4A,B). Consistent with strong somitic expression of *dpf3*, $\mathrm{MO}^{dpf3}$-injected embryos frequently displayed disturbed forward swimming movements indicating skeletal muscle defects (Fig. 4A). Coinjection of synthetic full-length *dpf3* mRNA produced a significant rescue effect, with the percentage of embryos with a *dpf3* morphant body phenotype decreasing to 46% ($n = 321$, $P < 0.01$) and the heart phenotype decreasing to 63% ($n = 145$, $P < 0.05$) (Fig. 4A,B). The heart phenotype was characterized by a thin and elongated heart tube, with both ventricular and atrial portions being affected. Moreover, looping of the heart was strongly reduced and the atrioventricular boundary was poorly defined in morphants (Fig. 4A,C). The strength of ventricular and atrial contractility was weakened compared with wild type, which resulted in slower blood flow, supported also by a significantly reduced ventricular shortening fraction (VSF) ($P < 0.05$) (data not shown). Nevertheless, the heart beat rate was normal (Supplemental Movies S1, S2). Both myocardial and endocardial layers were formed in morphant embryos, excluding defects in endocard–myocard signaling (data not shown).

In order to characterize the cardiac phenotype more thoroughly, we analyzed isolated hearts using confocal microscopy and found that despite the weakly developing atrioventricular boundary and loss of heart looping, atrial and ventricular myocyte specification was grossly normal (Fig. 4C). Immunohistochemistry using the atrial specific marker S46 showed that the atrium was clearly separated from the ventricle (Yelon et al. 1999). This finding was further confirmed by normal *atrial myosin heavy chain* (*amhc*) and *ventricular myosin heavy chain* (*vmhc*) expression at 72 hpf analyzed by in situ hybridization (Fig. 4D).

### *Dpf3 morphant zebrafish embryos display muscle fiber disarray*

To identify genes deregulated in *dpf3* morphants, we performed gene expression analysis (Affymetrix GeneChip Zebrafish Genome Arrays) using RNA from whole *dpf3* morphant embryos with severely reduced ventricular contractility and control-injected stage-matched embryos ($n = 30$, two replicates). Genes differentially regulated with an adjusted *P*-value of <0.1 were selected (1210 of ~15,000 transcripts) for global functional analysis based on overrepresented Gene Ontology terms (Supplemental Table S4). The set of up-regulated genes contained many genes essential for transcriptional regulation, nucleosome assembly, and metabolic processes, whereas genes involved in ion and electron transport were overrepresented among down-regulated genes. A subset of differentially expressed genes was confirmed by real-time PCR including genes directly involved in sarcomere assembly and muscle function that could explain the cardiac and skeletal muscle phenotypes of *dpf3* morphants (Supplemental Table S5). We observed significantly increased expression of *cmya1* (fold change 2.9) and of *actin-binding protein 280-like* (*flncb*; fold change 2.5). Furthermore, we found decreased expression of *heart and neural crest derivatives expressed 2* (*hand2*; fold change 0.5), *thymosin* β (fold change 0.3), and a novel protein (zgc:101755) similar to mouse *actin filament capping protein of muscle Z-lines* (fold change 0.5). However, as gene expression profiling was performed using whole embryos, we further analyzed expression levels in situ. Figure 4E shows in situ hybridization analysis confirming the differential expression of *hand2* and *cmya1*. To test if the heart looping defects were due to disturbed establishment of left–right asymmetry in the embryo, asymmetrically expressed markers *left–right determination factor 2* (*lefty2*) and *pitx2* were analyzed revealing that left–right asymmetry was properly initiated.

To further evaluate the deregulation of sarcomeric proteins we performed immunohistochemistry of morphant muscle fibers in the zebrafish and found a grossly disturbed actin organization compared with wild-type animals. The normal chevron-shaped somite organization was lost and myofibers were misaligned. Frequently, myofibers transversed somite boundaries. Focal adhesion kinase (FAK) is a marker of somite boundaries. Immunohistochemistry using an antibody against FAK showed disruption of somite boundaries (Fig. 4F). The thickness of somites was also markedly reduced. The specificity of this phenotype was confirmed by coinjection of synthetic full-length *dpf3* mRNA together with $\mathrm{MO}^{dpf3}$, which largely restored the myofiber organization and somite boundary formation (Fig. 4F).

Using transmission electron microscopy, we found that few myofibrils were present in *dpf3* morphant ventricles and skeletal muscle, which displayed a severe disruption of sarcomere assembly. Analysis of *Dpf3* siRNA knockdown in C2C12 mouse skeletal muscle cells showed conservation of this phenotype (Fig. 4G,H) with myofibrillar disarray compared with fiber aggregation in cells treated with control siRNA.

### *Mef2a regulates Dpf3 expression in vivo*

Mef2a-deficient mice and zebrafish embryos are phenotypically similar to the observed myofibrillar disarray in *dpf3* knockdown embryos (Naya et al. 2002; Wang et al. 2005; Potthoff et al. 2007). Consequently, we screened the *Dpf3* proximal promoter for potential Mef2-binding sites. Within a conserved 1.2kbp promoter region we found three Mef2 matrices using TRANSFAC MATCH with stringent settings (Fig. 5A; Kel et al. 2003). Mef2a ChIP–chip analysis in mouse cardiomyocytes (HL-1 cells) showed a significant peak of Mef2a binding in the *Dpf3* promoter region that could also be confirmed by real-time PCR (1.8-fold change) (Fig. 5A). Knockdown of *Mef2a* in HL-1 cells using two different siRNAs led to a

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

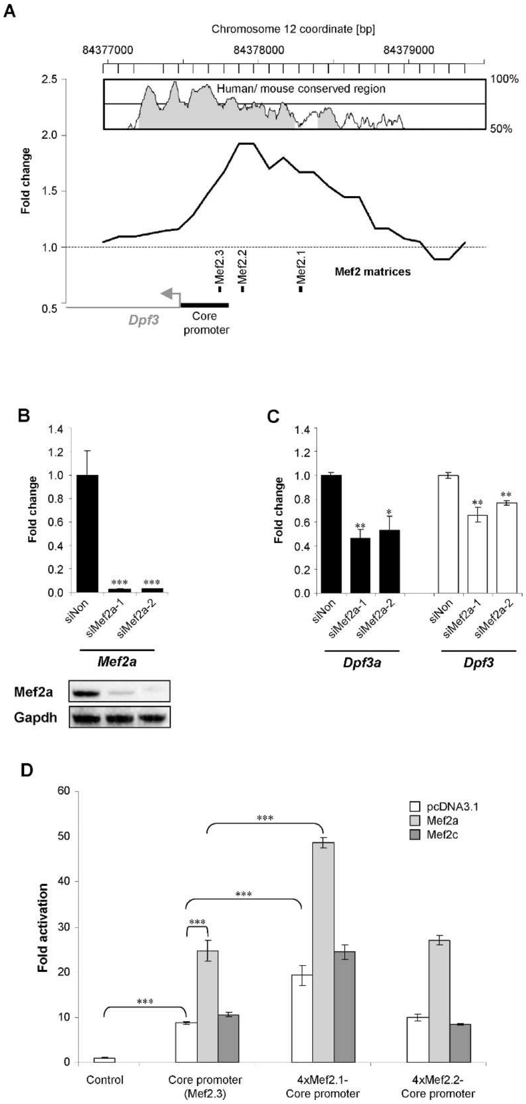


**Figure 5.** Mef2a regulates DPF3 expression. (*A*) ChIP followed by chip analysis shows binding of Mef2a to an evolutionary conserved region of the *Dpf3* promoter in vivo. Mef2a matrices obtained by TRANSFAC MATCH are indicated. Conservation of promoter sequence is shown. The *Dpf3* core promoter indicated is a minimal sufficient region required for transcriptional activity. (*B*) Knockdown of *Mef2a* in HL-1 cells using two different siRNAs. Knockdown efficiency was analyzed by real-time PCR and Western blot. (siNon) Nonspecific/scrambled. (*C*) Knockdown of *Mef2a* in HL-1 cells lead to reduced expression of *Dpf3* and *Dpf3a*. (siNon) Nonspecific/scrambled; (*) $P < 0.05$; (**) $P < 0.01$; (***) $P < 0.001$. (*D*) Luciferase reportergene assay using the *DPF3* core promoter alone or in combination with four repeats of the conserved, putative Mef2-binding sites (Mef2.1, Mef2.2, Mef2.3) fused to luciferase. Activity of the reporter was measured alone or in cotransfections with Mef2a/Mef2c expression vectors in HEK293T cells.

reduction of *Dpf3* expression of up to 40%, demonstrating that Mef2a functionally binds the *Dpf3* promoter and activates its expression (Fig. 5B,C). Transcriptional regulation of *Dpf3* by Mef2a was also tested in luciferase reportergene assays using promoter fusion constructs of a previously characterized *DPF3* core promoter (M. Lange and S. Sperling, unpubl.) and four consecutive repeats of the putative Mef2-binding sites. Cotransfections in HEK293T cells revealed an activation of the core promoter by Mef2a, which was additionally enhanced by the Mef2.1-binding site, supporting a role for Mef2a as a regulator of *Dpf3* through combinatorial effects on the Mef2.1 and Mef2.3 sites (Fig. 5D).

## Discussion

### *Targeting of the BAF chromatin remodeling complex to specific chromatin sites*

A central question regarding the action of chromatin remodeling complexes is how they are recruited to their target nucleosomes at specific positions within the genome. Most likely two mechanisms, the guidance by DNA-binding transcription factors and the binding to acetylated histone tails, play a central role (Peterson and Workman 2000; Hassan et al. 2001). Both transcription factor-binding sites as well as acetylated histones do not occur exclusively in conjunction with actively transcribed genes; thus, potentially, the interplay and co-occurrence of both might be essential for directed and tissue-specific gene transcription. Here, we present DPF3, which contains the first PHD fingers shown to bind acetylated in addition to methylated histone residues. Moreover, DPF3 links these modifications to the BAF chromatin remodeling complex and displays an essential role for skeletal and cardiac muscle development and function in vivo. The tissue-specific expression of DPF3 in combination with the specific read-out of modified histone residues allows for a side-directed recruitment of the BAF chromatin remodeling complex, similar to that of DNA-binding transcription factors.

The high impact of the modification status of histones (acetylation/deacetylation) on transcription and on the phenotype is well characterized; e.g., class II HDACs control cardiac growth and gene expression in response to stress stimuli (Backs and Olson 2006). DPF3 potentially represents the missing link to explain the high impact of the histone modification status on the recruitment of the BAF complex to chromatin target sites. So far, only bromodomains, frequently found in core and subunit proteins of chromatin remodeling complexes, have been shown to recognize histone acetylation marks.

Using ChIP we show on a global scale that DPF3 binds distinct chromatin sites in vivo, which are furthermore essential for muscle development and function, and marked by acetylated and/or methylated histones. It would be interesting to analyze if DPF3 is also associated with histone-modifying enzymes through the BAF complex or other interaction partners. Thus, the binding of DPF3 would be followed by a change in the histone modification status, building a regulatory feedback loop.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

### *The PHD of DPF3 binds modified histone lysine residues*

PHDs are frequently found in nuclear proteins, and are defined by a stretch of ~60 amino acids containing conserved cysteine and histidine residues (C4-H-C3) that coordinate two zinc ions forming interweaved zinc fingers bridged by two small β-strands (Bienz 2006). They are known to serve as a protein–protein interaction domain and bind nuclear phosphoinositides as well as nucleosomes (Bienz 2006; Ruthenburg et al. 2007). Moreover, in a proteome-wide screen, only eight out of 18 PHD fingers showed specific histone methyl-lysine interactions, indicating additional roles for the PHD (Shi et al. 2007). We report that the double PHD finger of DPF3 interact with acetylated as well as methylated histone tail residues, namely acetylated lysines on histones 3 and 4 (H3K9ac, H3K14ac, H4K5ac, H4K8ac, H4K12ac, H4K16ac) and mono- and dimethylated lysine on histone 3 (H3K4me1/me2). Interestingly, single PHD fingers of DPF3 only recognize histone 4 acetylation and an intact PHD finger is necessary for histone interactions, as the truncated PHD1 of DPF3a is not capable of binding any histones.

So far, single PHD fingers have been shown to recognize methylated histones; e.g., the PHD fingers of BPTF and ING2 (Shi et al. 2006; Wysocka et al. 2006) bind H3K4me with increasing affinity according to methylation status, while BHC80-PHD recognizes unmodified H3K4 (Lan et al. 2007). Moreover, methylation at different residues, namely H3 methylated at both Lys 4 and Arg 2, can be read simultaneously by a single PHD of RAG2, revealing additional complexity in the readout of combinatorial modifications (Ramon-Maiques et al. 2007). Further experiments are needed to answer the question if binding of acetylation and methylation marks by the double PHD finger of DPF3 can occur simultaneously, which would allow a combinatorial readout of different modifications.

The finding that H3 modifications are only recognized by the double PHD finger may be due to the interweaved nature of the PHD finger. The domain necessary for H3K4me1/me2 and H3ac recognition might be a compound in which amino acids from PHD1 and PHD2 contribute to the three-dimensional structure.

Histone methyl-lysine-binding properties similar to DPF3 have been described for the malignant brain tumor (MBT) domain of L3MBTL1 and a mutated form of BPTF-PHD, which also specifically recognize H3K4me1/me2. Although structurally unrelated, both domains achieve methyl-lysine binding through formation of a cage consisting of aromatic residues (H. Li et al. 2007; Min et al. 2007). The PHD fingers of DPF3 contain several aromatic residues that can potentially contribute to the formation of an aromatic cage, although a conserved tryptophan is missing.

Further experiments using crystallography and NMR spectroscopy will determine the structural basis for the histone tail recognition by DPF3.

### *Role of DPF3 in heart and skeletal muscle development*

The up-regulated expression of *DPF3* in patients with TOF, a congenital heart defect in part characterized by muscular hypertrophy, prompted us to investigate its role during development and muscle differentiation. Knockdown in zebrafish embryos and RNAi in mouse skeletal muscle cells revealed an essential role of Dpf3 in muscle cell differentiation.

In morphant embryos, we frequently observed myofibrillar disarray, transversion of the somite boundary by actin filaments, and disruption of somite boundary formation. In particular, the z-disc of sarcomeres representing the lateral boundaries where titin, nebulin, and the thin filaments are anchored (Clark et al. 2002), appeared to be affected. This phenotype could be explained by the deregulation of several genes essential for muscle fiber function shown by our expression studies—e.g., *capZ α-1* (zgc:101755) and *tropomodulin 4* (Schafer et al. 1995; Sussman et al. 1998)—the actin-binding protein *filamin c γ b* (flncb) and its interaction partner *cmya1*. *Filamin C*-knockout mice display severe defects in myogenesis, including loss of distinct z-discs (Dalkilic et al. 2006), while *Cmya1*-α-null mouse hearts show intercalated disc disruption and myofilament disarray (Gustafson-Wagner et al. 2007). Further, *dpf3* morphants frequently displayed impaired cardiac contractility, which may be due to the strong up-regulation of *troponin I*. Notably, mice expressing mutated versions of Troponin I display hypercontractility (James et al. 2000), mirror imaging the *dpf3* morphant phenotype.

The morphant phenotype was also characterized by disturbed heart looping and a poorly defined AV boundary. Initial microarray analyses point to the deregulation of transcription factors and extracelluar matrix molecules implicated in heart looping and left–right asymmetry (data not shown). These molecules will be subject to further studies on the role of *dpf3* in early heart development. Notably, knockdown of *Smarcd3*, the DPF3 interaction subunit of the BAF complex, also affects heart looping in mouse and zebrafish by influencing Notch signaling (Takeuchi et al. 2007). Moreover, *Bmp2*, a gene essential for development of the AV cushions (Ma et al. 2005) is a target of Dpf3 in C2C12 cells analyzed by ChIP, and has been shown to be upstream of *mef2a* in zebrafish in a pathway controlling cardiac contractility (Wang et al. 2007).

Interestingly, the *dpf3* morphant phenotype resembles in part the defects seen in *mef2a* morphants and Mef2a-deficient mice (Naya et al. 2002; Wang et al. 2005). As our experiments show that Mef2a regulates *Dpf3*, it is suggestive that the Mef2a phenotypes are partially caused by loss of Dpf3 function. In the future, it will be interesting to test the influence of Dpf3 on the Mef2a phenotypes in mouse and zebrafish in detail.

Despite the strong expression of *dpf3* in neuronal cells, we did not observe any obvious malformations of the brain. It has been shown recently that Dpf3a and Dpf1 seem to have overlapping functions during differentia-

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

tion of neurons (Lessard et al. 2007). It is likely that Dpf1 may compensate for the loss of Dpf3 there, while expression in striated muscle appears exclusive to Dpf3.

We report that DPF3 contains the first PHDs known to bind acetylated as well as methylated histone residues, interacts with the BAF complex, and has an essential role for muscle development and function. It is tempting to speculate that DPF3a and DPF3b might serve as tissue-specific BAF subunits that regulate the transition of muscle precursors to differentiating myocytes. Moreover, it is highly suggestive that other PHD fingers might be capable to bind acetylation marks and play a yet unappreciated role in recruiting chromatin remodeling complexes.

## Materials and methods

Detailed procedures are provided in the Supplemental Material

### *Samples and preparation*

Human cardiac samples were obtained from the German Heart Center and treated as described (Kaynak et al. 2003). Mouse embryonic and adult hearts were dissected from the rest of the body at indicated stages and handled as human samples.

### *Gene expression analyses*

Real-time PCR analysis was performed using SYBR Green I PCR Master Mix (Abgene) and the ABI PRISM 7900HT Sequence Detection System. Primer sequences are given in Supplemental Table S6. In situ hybridization in mouse, chicken, and zebrafish embryos was carried out as described (Wilkinson and Nieto 1993; Jowett and Lettice 1994). A multiple tissue human Northern blot (NTM 12, Clontech) was hybridized with a $^{32}$P-labeled cDNA probe against *DPF3* (AY803021; 7–423 bp) according to the manufacturers' instructions.

Affymetrix GeneChip Zebrafish Genome Arrays were hybridized with labeled cDNA obtained from total RNA of MO-*dpf3* and MO-control-injected zebrafish embryos 72 hpf. Four chips were hybridized (two MO-control, two MO-$^{dpf3}$, 30 embryos each) (www.ebi.ac.uk/arrayexpress, E-TABM-354). Data were normalized via qspline after MAS background correction using the Bioconductor affy package and the zebrafish annotation package. Differentially expressed genes were calculated via the limma package. *P*-values were adjusted for multiple testing using the Benjamini and Hochberg method. Genes with an adjusted *P*-value of <0.1 were defined as differentially expressed.

### *Antisense oligonucleotide MO and rescue experiments*

Full-length zebrafish *dpf3* (NM_001111169) was cloned into the pCS2$^+$ expression vector and used as rescue construct. Constructs were transcribed using the SP6 MessageMachine kit (Ambion). For functional and rescue experiments, wild-type Tuebingen LF/AB hybrids; Tg(*cmlc2:GFP*) transgenic fish embryos were injected with ~75 pg of mRNA. MOs (GeneTools) were injected at a concentration of 100 µmol/L.

### *Confocal and live-stream imaging*

Confocal images and z-stacks were obtained using the Zeiss Meta 510 confocal microscope with a 40× lens and 1× zoom. For live-stream imaging, *Tg(cmlc2:GFP)* transgenic embryos were prepared as described (Westerfield 1994). Myocardial contraction and beating of the developing heart tube was imaged with a CoolSnap ES camera (Photometrics) on an Axioplan2 microscope.

### *Immunohistochemistry and transmission electron microscopy*

Antibody staining was performed as described previously (Huang et al. 2003). Zebrafish electron micrographs were obtained essentially as described (Rottbauer et al. 2001). C2C12 cells were grown on Thermanox coverslips (13 mm ø; Nunc) and embedded in Spurr's resin. Sixty-nanometer sections were observed using Philips CM100 at 100 kV (FEI Company) with a TVIPS Fastscan CCD camera (Tietz Systems).

### *Proteomic analyses*

GST-DPF3 fusion proteins were created using the pGEX3x vector, expressed in *Escherichia coli* BL21 DE3 pRARE and purified using glutathione-sepharose matrix (Amersham) according to the manufacturer's instructions.

For histone peptide-binding assays, 1 µg of biotinylated histone peptide (Upstate Biotechnologies, and kind gifts of D. Patel and D. Allis) was incubated with 1 µg of purified GST fusion protein in binding buffer (50 mM Tris-HCl 7.5, 300 mM NaCl, 0.1% NP-40, 50 µM ZnAc) overnight at 4°C with rotation. Streptavidin beads (Dynabeads) were added and incubated for 1 h at 4°C with rotation followed by four rounds of 15 min washing in binding buffer. Bound proteins were analyzed on SDS-PAGE gels and subjected to immunoblotting analysis.

TAP was performed essentially as described (Gingras et al. 2005). Full-length DPF3a, DPF3b, and SMARCD3 was cloned into the pcDNA3-NTAP vector, verified by sequencing, and transfected into HEK293T cells.

### *siRNA knockdown experiments*

C2C12 or HL-1 cells were seeded in six-well plates and transfected with 4.4 µL of 20 µm siRNA (Supplemental Table S7). siRNAs targeting *Dpf3a* (Invitrogen), both splice variants of *Dpf3* (Qiagen), or a control siRNA (AllStars Negative Control siRNA, Qiagen) were used in C2C12, and siRNAs targeting *Mef2a* in HL-1 cells. XtremeGene (Roche) and Lipofectamine Plus (Invitrogen) were used for transfection according to manufacturer's protocol and cultivated for 48 h. Cells were subsequently subjected to electron microscopy or microarray gene expression analysis.

### *ChIP with chip detection (ChIP–chip)*

C2C12 myoblasts cells were used either untransfected or transfected with Flag-empty, Flag-DPF3a or Flag-DPF3b expression vectors using Lipofectamine Plus (Invitrogen) according to manufacturers' instructions (Supplemental Fig. S3). ChIP experiments were performed in duplicate essentially as described (Horak et al. 2002). For immunoprecipitation, mouse-M2-anti-Flag (Sigma) antibody and Brg1 (Santa Cruz Biotechnologies, sc-10768) antibodies were used at 10 and 5 µg/mL for C2C12 cells and rabbit anti-Mef2A (Santa Cruz Biotechnologies) at 2 µg/mL for HL-1 cells. Samples were labeled and hybridized according to NimbleGen standard procedures on custom designed muscle arrays (www.ebi.ac.uk/arrayexpress, A-MEXP-893). Array analysis was performed as described (Toedling et al. 2007). Enriched targets (23 sites) of the negative control (Flag-empty) were subtracted from DPF3 ChIP data. Data are deposited at www.ebi.ac.uk/arrayexpress (E-TABM-362).

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

*Reporter gene assays*

Reporter constructs were made by cloning four repeats of the putative Mef2-binding site upstream of a 385-bp (chr14:72,430,563–72,430,943) *DPF3* minimal promoter into the pGL3 basic vector (Promega). Transient cotransfections were carried out in triplicates in 96-well plates in HEK293T cells by transfecting 45 ng of reporter vector, 5 ng of Firefly luciferase vector for internal normalization of transfection efficiency, and 100 ng of the respective expression vectors. Activity was measured by Dual-Luciferase Assay (Promega) after 48 h.

## Acknowledgments

We gratefully acknowledge Drs. Bruneau, Gingras, Black, and Patel for providing plasmids; Drs. Allis and Patel for modified histone peptides; Dr. Yelon for amhc and vmhc probes; Dr. Schmidt for C2C12 cell line; Dr. Claycomb for the HL-1 cell line; and Dr. Huber and J. Toedling for bioinformatic help. We are indebted to B. Lukaszewska-McGreal and G. Lueder for technical assistance, M. Schueler for bioinformatic support, and Dr. Nordhoff for mass spectrometry support. We are very grateful to all patients and family members with congenital heart disease supporting our work. This work was supported by a grant to S.S. from the European Community's Sixth Framework Programme contract ("HeartRepair") LSHM-CT-2005-018630. Furthermore, M.T. receives a Ph.D. scholarship of the Studienstiftung des Deutschen Volkes. The work of J.G. was funded by the Structural Fonds of the European Union (project 2D/3D-ProteinChips) and the BMBF/NGFN project SMP Protein.

## References

Backs, J. and Olson, E.N. 2006. Control of cardiac growth by histone acetylation/deacetylation. *Circ. Res.* **98:** 15–24.

Bao, Y. and Shen, X. 2007. SnapShot: Chromatin remodeling complexes. *Cell* **129:** 632.e1–632.e2. doi: 10.1016/j.cell.2007.04.018.

Bernstein, B.E., Meissner, A., and Lander, E.S. 2007. The mammalian epigenome. *Cell* **128:** 669–681.

Bienz, M. 2006. The PHD finger, a nuclear protein-interaction domain. *Trends Biochem. Sci.* **31:** 35–40.

Bourachot, B., Yaniv, M., and Muchardt, C. 1999. The activity of mammalian brm/SNF2α is dependent on a high-mobility-group protein I/Y-like DNA binding domain. *Mol. Cell. Biol.* **19:** 3931–3939.

Clark, K.A., McElhinny, A.S., Beckerle, M.C., and Gregorio, C.C. 2002. Striated muscle cytoarchitecture: An intricate web of form and function. *Annu. Rev. Cell Dev. Biol.* **18:** 637–706.

Dalkilic, I., Schienda, J., Thompson, T.G., and Kunkel, L.M. 2006. Loss of FilaminC (FLNc) results in severe defects in myogenesis and myotube structure. *Mol. Cell. Biol.* **26:** 6522–6534.

Debril, M.B., Gelman, L., Fayard, E., Annicotte, J.S., Rocchi, S., and Auwerx, J. 2004. Transcription factors and nuclear receptors interact with the SWI/SNF complex through the BAF60c subunit. *J. Biol. Chem.* **279:** 16677–16686.

Elfring, L.K., Daniel, C., Papoulas, O., Deuring, R., Sarte, M., Moseley, S., Beek, S.J., Waldrip, W.R., Daubresse, G., DePace, A., et al. 1998. Genetic analysis of brahma: The *Drosophila* homolog of the yeast chromatin remodeling factor SWI2/SNF2. *Genetics* **148:** 251–265.

Fischer, J.J., Toedling, J., Krueger, T., Schueler, M., Huber, W., and Sperling, S. 2008. Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics* **91:** 41–51.

Flajollet, S., Lefebvre, B., Cudejko, C., Staels, B., and Lefebvre, P. 2007. The core component of the mammalian SWI/SNF complex SMARCD3/BAF60c is a coactivator for the nuclear retinoic acid receptor. *Mol. Cell. Endocrinol.* **270:** 23–32.

Florens, L., Carozza, M.J., Swanson, S.K., Fournier, M., Coleman, M.K., Workman, J.L., and Washburn, M.P. 2006. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40:** 303–311.

Gabig, T.G., Mantel, P.L., Rosli, R., and Crean, C.D. 1994. Requiem: A novel zinc finger gene essential for apoptosis in myeloid cells. *J. Biol. Chem.* **269:** 29515–29519.

Gingras, A.-C., Aebersold, R., and Raught, B. 2005. Advances in protein complex analysis using mass spectrometry. *J. Physiol.* **563:** 11–21.

Gustafson-Wagner, E.A., Sinn, H.W., Chen, Y.L., Wang, D.Z., Reiter, R.S., Lin, J.L., Yang, B., Williamson, R.A., Chen, J., Lin, C.I., et al. 2007. Loss of mXinα, an intercalated disk protein, results in cardiac hypertrophy and cardiomyopathy with conduction defects. *Am. J. Physiol. Heart Circ. Physiol.* **293:** H2680–H2692.

Hassan, A.H., Neely, K.E., and Workman, J.L. 2001. Histone acetyltransferase complexes stabilize swi/snf binding to promoter nucleosomes. *Cell* **104:** 817–827.

Hassan, A.H., Prochasson, P., Neely, K.E., Galasinski, S.C., Chandy, M., Carrozza, M.J., and Workman, J.L. 2002. Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* **111:** 369–379.

Hassan, A.H., Awad, S., Al-Natour, Z., Othman, S., Mustafa, F., and Rizvi, T.A. 2007. Selective recognition of acetylated histones by bromodomains in transcriptional co-activators. *Biochem. J.* **402:** 125–133.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. 2002. GATA-1 binding sites mapped in the β-globin locus by using mammalian chIp–chip analysis. *Proc. Natl. Acad. Sci.* **99:** 2924–2929.

Huang, C.J., Tu, C.T., Hsiao, C.D., Hsieh, F.J., and Tsai, H.J. 2003. Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin light chain 2 promoter of zebrafish. *Dev. Dyn.* **228:** 30–40.

James, J., Zhang, Y., Osinska, H., Sanbe, A., Klevitsky, R., Hewett, T.E., and Robbins, J. 2000. Transgenic modeling of a cardiac troponin I mutation linked to familial hypertrophic cardiomyopathy. *Circ. Res.* **87:** 805–811.

Jowett, T. and Lettice, L. 1994. Whole-mount in situ hybridizations on zebrafish embryos using a mixture of digoxigenin- and fluorescein-labelled probes. *Trends Genet.* **10:** 73–74.

Karamboulas, C., Dakubo, G.D., Liu, J., De Repentigny, Y., Yutzey, K., Wallace, V.A., Kothary, R., and Skerjanc, I.S. 2006. Disruption of MEF2 activity in cardiomyoblasts inhibits cardiomyogenesis. *J. Cell Sci.* **119:** 4315–4321.

Kassabov, S.R., Zhang, B., Persinger, J., and Bartholomew, B. 2003. SWI/SNF unwraps, slides, and rewraps the nucleosome. *Mol. Cell* **11:** 391–403.

Kaynak, B., von Heydebreck, A., Mebus, S., Seelow, D., Hennig, S., Vogel, J., Sperling, H.P., Pregla, R., Alexi-Meskishvili, V., Hetzer, R., et al. 2003. Genome-wide array analysis of normal and malformed human hearts. *Circulation* **107:** 2467–2474.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA se-

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

quences. *Nucleic Acids Res.* **31:** 3576–3579.

Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y., and Bedford, M.T. 2006. Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep.* **7:** 397–403.

Kouzarides, T. 2007. Chromatin modifications and their function. *Cell* **128:** 693–705.

Lan, F., Collins, R.E., De Cegli, R., Alpatov, R., Horton, J.R., Shi, X., Gozani, O., Cheng, X., and Shi, Y. 2007. Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **448:** 718–722.

Lessard, J., Wu, J.I., Ranish, J.A., Wan, M., Winslow, M.M., Staahl, B.T., Wu, H., Aebersold, R., Graef, I.A., and Crabtree, G.R. 2007. An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron* **55:** 201–215.

Li, H., Fischle, W., Wang, W., Duncan, E.M., Liang, L., Murakami-Ishibe, S., Allis, C.D., and Patel, D.J. 2007. Structural basis for lower lysine methylation state-specific readout by MBT repeats of L3MBTL1 and an engineered PHD finger. *Mol. Cell* **28:** 677–691.

Li, Z.Y., Yang, J., Gao, X., Lu, J.Y., Zhang, Y., Wang, K., Cheng, M.B., Wu, N.H., Wu, Z., and Shen, Y.F. 2007. Sequential recruitment of PCAF and BRG1 contributes to myogenin activation in 12-O-tetradecanoylphorbol-13-acetate-induced early differentiation of rhabdomyosarcoma-derived cells. *J. Biol. Chem.* **282:** 18872–18878.

Lickert, H., Takeuchi, J.K., von Both, I., Walls, J.R., McAuliffe, F., Lee Adamson, S., Mark Henkelman, R., Wrana, J.L., Rossant, J., and Bruneau, B.G. 2004. Baf60c is essential for function of BAF chromatin remodelling complexes in heart development. *Nature* **432:** 107–112.

Ma, L., Lu, M.F., Schwartz, R.J., and Martin, J.F. 2005. Bmp2 is essential for cardiac cushion epithelial–mesenchymal transition and myocardial patterning. *Development* **132:** 5601–5611.

Mertsalov, I.B., Kulikova, D.A., Alimova-Kost, M.V., Ninkina, N.N., Korochkin, L.I., and Buchman, V.L. 2000. Structure and expression of two members of the d4 gene family in mouse. *Mamm. Genome* **11:** 72–74.

Min, J., Allali-Hassani, A., Nady, N., Qi, C., Ouyang, H., Liu, Y., MacKenzie, F., Vedadi, M., and Arrowsmith, C.H. 2007. L3MBTL1 recognition of mono- and dimethylated histones. *Nat. Struct. Mol. Biol.* **14:** 1229–1230.

Mohrmann, L. and Verrijzer, C.P. 2005. Composition and functional specificity of SWI2/SNF2 class chromatin remodeling complexes. *Biochim. Biophys. Acta* **1681:** 59–73.

Mujtaba, S., Zeng, L., and Zhou, M.M. 2007. Structure and acetyl-lysine recognition of the bromodomain. *Oncogene* **26:** 5521–5527.

Natalia, N.N., Ilja, B.M., Dina, A.K., Maria, V.A.-K., Olga, B.S., Leonid, I.K., Sergey, L.K., and Vladimir, L.B. 2001. Cerd4, third member of the d4 gene family: Expression and organization of genomic locus. *Mamm. Genome* **V12:** 862–866.

Naya, F.J., Black, B.L., Wu, H., Bassel-Duby, R., Richardson, J.A., Hill, J.A., and Olson, E.N. 2002. Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. *Nat. Med.* **8:** 1303–1309.

Palacios, D. and Puri, P.L. 2006. The epigenetic network regulating muscle development and regeneration. *J. Cell. Physiol.* **207:** 1–11.

Peterson, C.L. and Workman, J.L. 2000. Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev.* **10:** 187–192.

Potthoff, M.J., Arnold, M.A., McAnally, J., Richardson, J.A., Bassel-Duby, R., and Olson, E.N. 2007. Regulation of skeletal muscle sarcomere integrity and postnatal muscle function by Mef2c. *Mol. Cell. Biol.* **27:** 8143–8151.

Ramon-Maiques, S., Kuo, A.J., Carney, D., Matthews, A.G., Oettinger, M.A., Gozani, O., and Yang, W. 2007. The plant homeodomain finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proc. Natl. Acad. Sci.* **104:** 18993–18998.

Rauch, C. and Loughna, P.T. 2005. Static stretch promotes MEF2A nuclear translocation and expression of neonatal myosin heavy chain in C2C12 myocytes in a calcineurin- and p38-dependent manner. *Am. J. Physiol. Cell Physiol.* **288:** C593–C605. doi: 10.1152/ajpcell.00346.2004.

Rottbauer, W., Baker, K., Wo, Z.G., Mohideen, M.A., Cantiello, H.F., and Fishman, M.C. 2001. Growth and function of the embryonic heart depend upon the cardiac-specific L-type calcium channel α1 subunit. *Dev. Cell* **1:** 265–275.

Ruthenburg, A.J., Allis, C.D., and Wysocka, J. 2007. Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25:** 15–30.

Schafer, D.A., Hug, C., and Cooper, J.A. 1995. Inhibition of CapZ during myofibrillogenesis alters assembly of actin filaments. *J. Cell Biol.* **128:** 61–70.

Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F., Kaadige, M.R., et al. 2006. ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442:** 96–99.

Shi, X., Kachirskaia, I., Walter, K.L., Kuo, J.H., Lake, A., Davrazou, F., Chan, S.M., Martin, D.G., Fingerman, I.M., Briggs, S.D., et al. 2007. Proteome-wide analysis in *Saccharomyces cerevisiae* identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *J. Biol. Chem.* **282:** 2450–2455.

Sif, S. 2004. ATP-dependent nucleosome remodeling complexes: Enzymes tailored to deal with chromatin. *J. Cell. Biochem.* **91:** 1087–1098.

Simone, C. 2006. SWI/SNF: The crossroads where extracellular signaling pathways meet chromatin. *J. Cell. Physiol.* **207:** 309–314.

Simone, C., Forcales, S.V., Hill, D.A., Imbalzano, A.N., Latella, L., and Puri, P.L. 2004. p38 pathway targets SWI–SNF chromatin-remodeling complex to muscle-specific loci. *Nat. Genet.* **36:** 738–743.

Sperling, S. 2007. Transcriptional regulation at a glance. *BMC Bioinformatics* **8:** S2. doi: 10.1186/1471-2105-8-S6-S2.

Sussman, M.A., Baque, S., Uhm, C.S., Daniels, M.P., Price, R.L., Simpson, D., Terracio, L., and Kedes, L. 1998. Altered expression of tropomodulin in cardiomyocytes disrupts the sarcomeric structure of myofibrils. *Circ. Res.* **82:** 94–105.

Takeuchi, J.K., Lickert, H., Bisgrove, B.W., Sun, X., Yamamoto, M., Chawengsaksophak, K., Hamada, H., Yost, H.J., Rossant, J., and Bruneau, B.G. 2007. Baf60c is a nuclear Notch signaling component required for the establishment of left–right asymmetry. *Proc. Natl. Acad. Sci.* **104:** 846–851.

Toedling, J., Skylar, O., Krueger, T., Fischer, J.J., Sperling, S., and Huber, W. 2007. Ringo—An R/bioconductor package for analyzing ChIP–chip readouts. *BMC Bioinformatics* **8:** 443. doi: 10.1186/1471-2105-8-221.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T. 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131:** 58–69.

Wang, Y.X., Qian, L.X., Yu, Z., Jiang, Q., Dong, Y.X., Liu, X.F., Yang, X.Y., Zhong, T.P., and Song, H.Y. 2005. Requirements of myocyte-specific enhancer factor 2A in zebrafish cardiac contractility. *FEBS Lett.* **579:** 4843–4850.

Downloaded from genesdev.cshlp.org on September 10, 2008 - Published by Cold Spring Harbor Laboratory Press

Wang, Y.X., Qian, L.X., Liu, D., Yao, L.L., Jiang, Q., Yu, Z., Gui, Y.H., Zhong, T.P., and Song, H.Y. 2007. Bone morphogenetic protein-2 acts upstream of myocyte-specific enhancer factor 2a to control embryonic cardiac contractility. *Cardiovasc. Res.* **74:** 290–303.

Westerfield, M. 1994. *The zebrafish book.* University of Oregon Press, Eugene, OR.

Wilkinson, D.G. and Nieto, M.A. 1993. Detection of messenger RNA by in situ hybridization to tissue sections and whole mounts. *Methods Enzymol.* **225:** 361–373.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. 2006. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442:** 86–90.

Yelon, D., Horne, S.A., and Stainier, D.Y.R. 1999. Restricted expression of cardiac myosin genes reveals regulated aspects of heart tube assembly in zebrafish. *Dev. Biol.* **214:** 23–37.

## 2.3 Advanced biotechniques to analyze patient material

Single-nucleotide polymorphisms (SNPs) as well as gene expression levels have become popular markers for a wide range of molecular studies, including both diagnostic and research settings for analysis of human disease. However, the use of these markers requires appropriate methods, which are suitable for routinely performed diagnostics outside of well-equipped research institutions. Thus they should be robust, reliable, easy to handle, and low cost, without the need of expensive instrumentation. Along with this line, the LDR-TaqMan genotyping assays and the cr-real-time PCR were developed and applied in the patient studies described.

### 2.3.1 A novel technique for quantitative PCR analysis – cr-real-time PCR

> Rickert AM, Lehrach H, **Sperling S**. Multiplexed real-time PCR using universal reporters. *Clin Chem* 2004;**50**:1680-1683.

Real-time quantitative PCR is widely accepted as the method of choice for gene expression studies, being more sensitive and accurate than alternative methods used to analyze mRNA levels, such as Northern blotting, RNase protection assay, in situ hybridization, reverse transcription (RT-) PCR, and cDNA arrays (Giulietti et al., 2001). The detection chemistries of all real-time PCR procedures are based on one of two principles for monitoring amplification products: binding to double-stranded (ds) DNA or hybridization to single-stranded (ss) DNA. Small molecules bind to dsDNA either as intercalators or as minor groove binders, like ethidium bromide, Hoechst 33258 or SYBR Green I. Several approaches using target-specific hybridization to ssDNA have been introduced, including Molecular Beacons (Tyagi et al., 1998), Scorpions (Whitcombe et al., 1999), the TaqMan or hydrolysis/5'-nuclease assay (Holland et al., 1991), AEGIS probe system (Moser et al., 2003), labelled primers (Nazarenko et al., 2002) or light-up probes (Svanvik et al., 2000). In contrast to dsDNA binding dyes, these methods are suitable for multiplexing approaches by using differentially labelled fluorescent dyes. However, since requiring a unique probe or modified primer for each target, currently used hybridization based methods for real-time quantitative PCR suffer from high costs and establishing efforts per assay.

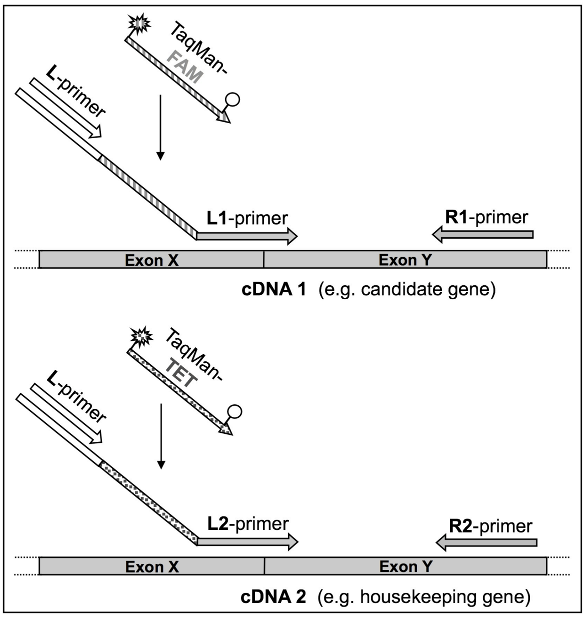


**Figure 15. Scheme of the multiplexed *cr-real-time PCR* assay.** Each gene of interest is amplified using a set of three primers: One tailed locus-specific forward oligonucleotide (L*1*/*2*-primer), one locus-specific reverse primer (R*1*/*2*-primer), and one common primer (L-primer) annealing to a common sequence stretch of the tails. Due to varying sequence tags of the tails, amplification products of different cDNAs are specifically detected via TaqMan probes corresponding to the appropriate tags. As illustrated, the design of exon-junction spanning primers appears advisable, though it is not mandatory for the assay. White arrows indicate common sequences and primers; gray arrows indicate locus-specific sequences and primers; patterned arrows indicate TaqMan probes and their corresponding sequence stretches in the tailed oligonucleotides; arrow endings indicate 5'-ends of sequences; and blunt endings indicate 3'-ends of sequences.

The presented cr-real-time PCR assay uses universal hybridization based probe sets suitable for any target. Since using tailed locus-specific non-modified amplification primers, PCR products can be monitored via *c*ommon *r*eporters (*cr*) hybridizing to the common tails. The general principle of combining tailed PCR primers with universal probes has been introduced for other genetic applications like SNP genotyping or in situ amplification, however these methods had not been applied for quantitative gene expression studies so far. Due to the application of differentially labelled universal reporting reagents, *cr-real-time PCR* enables a multiplexing setup for simultaneous analysis of target gene and internal control (housekeeping gene), as an accepted method for normalizing sample-to-sample variation. The cr-real-time PCR assay was compared to SYBR Green I assays with respect to robustness and sensitivity; and its similar specificity due to the use of primer-specific detection. As a proof of principle, the new assay was applied on a study comprising three previously published differentially expressed candidate genes for congenital heart defects (Kaynak et al., 2003).

The principle of the cr-real-time PCR assay is as follows: Target DNA is subjected to PCR in the presence of four oligonucleotides; one low concentrated tailed locus-specific primer, one locus-specific primer without tail, one common primer annealing to a common sequence stretch of the tail, and one universal TaqMan probe corresponding to another common part of the tail (**Figure 15**). Choosing locus-specific primers that span at least one intron of the genomic sequence minimizes problems associated with DNA contamination. During the first amplification cycle, the tailed locus-specific primer initiates the polymerase reaction, leading to the synthesis of a fragment with the tail at the 5'-end. Cycle two leads to the synthesis of the complement of the tail. Starting from cycle three, the common amplification primer primes synthesis on the fragment including the tail sequence. From this step on, the TaqMan probe anneals to the products resulting from the amplification with the common primer and the locus-specific primer without tail. Due to the 5'-nuclease activity of the DNA polymerase, the TaqMan probe bound to the tailed target amplicon is hydrolyzed, leading to a physical separation of the reporter and quencher dye and release of fluorescence emission. The introduction of further common tails and corresponding differentially labelled TaqMan probes opens up the possibility for multiplexing. Though, the degree of multiplexing is limited by the number of different dyes and the restrictions of currently available instruments. Since employing universal tails and TaqMan probes, they can be easily transferred to other targets in combination with two locus-specific primer sequences. In cases, where locus-specific primer are already designed, ligation to the corresponding common tails using the ligation-based synthesis method could be envisaged. Furthermore, the amplification of primer dimers or pseudogenes can be tested by a melting analysis using SYBR Green I instead of the TaqMan probe for amplicon detection.

In summary, cr-real-time PCR is a single-step method that is sensitive, robust, and requires minimal optimization effort. Since the system utilizes non-modified tailed amplification primers and universal reporting reagents, it is characterized by a flexible and low-cost format. Due to the use of differentially labelled reporting reagents, multiplexing approaches can be performed monitoring more than one target per well, e.g. candidate and housekeeping gene. Therefore, the cr-real-time PCR assay appears suitable for the broad spectrum of all real-time PCR applications.

**Table 1. Absolute reverse transcription yields for RNA.**

| | Mean (SD) yields[a] (%) at external RNA input (in molecules) of: | | | | Mean (SD)[b] yield for RNA MultiStandard, % |
|---|---|---|---|---|---|
| | $10^6$ | $10^5$ | $10^4$ | $10^3$ | |
| MMLVH | 22 | 50 | 48 | 125 | 40 (16) |
| Omniscript | 7.2 | 3.1 | 11.5 | 66 | 7.3 (4.2) |
| AMV | 0.4 | 0.6 | 4.9 | 44 | 2.0 (2.5) |
| MMLV | 32 | 49 | 50 | 110 | 44 (10) |
| Improm-II | 32 | 22 | 12 | 98 | 22 (10) |
| cAMV | 6.3 | 17 | 35 | 88 | 19 (15) |
| ThermoScript | 1.1 | 9.0 | 14 | 46 | 8.0 (6.6) |
| SuperScript III | 87 | 72 | 90 | 43 | 83 (10) |
| Mean (SD) | 24 (29) | 28 (26) | 33 (29) | 78 (32) | 28 (27) |

[a] Reverse transcription yields of RNA prepared from liver and spleen. The samples were diluted 30-fold before QPCR measurements, giving initial copy numbers of 33–33 333 molecules/sample. Note the markedly higher yields at an input of $10^3$ RNA molecules.

[b] Reverse transcription yield for samples containing $10^4$-$10^6$ RNA MultiStandard molecules.

shown to be significantly improved when carrier is used *(1, 15)*. The reverse transcription yields for the RNA MultiStandard varied more than 100-fold. The lowest yield (0.4%) was obtained with AMV for $10^6$ RNA molecules, and the highest yield (90%) was obtained with SuperScript III for $10^4$ RNA molecules (Table 1). The latter was overall the most efficient reverse transcriptase, with a mean yield of 83%. MMLV and MMLVH gave mean yields of 44% and 40%, respectively, whereas the mean yields of the other reverse transcriptases were <25%. The yield obtained with MMLVH was comparable to that reported in a previous study *(15)*.

In conclusion, we show that reverse transcription yields vary up to 100-fold with the choice of reverse transcriptase and that the variation is gene dependent. Previously, we also reported a dependence on priming strategy *(1)*. Hence, for quantitative gene expression measurements based on reverse transcription to be comparable among laboratories, the same enzyme, priming strategy, and experimental conditions must be used.

This work was supported by Deutsche Forschungsgemeinschaft, the Chalmers Bioscience effort, and the Crafoord Foundation.

**References**

1. Ståhlberg A, Håkansson J, Xian X, Semb H, Kubista M. Properties of the reverse transcription reaction in mRNA quantification. Clin Chem 2004;50: 509–15.
2. Polumuri SK, Ruknudin A, Schulze DH. RNase H and its effects on PCR. Biotechniques 2002;32:1224–5.
3. Mayers TW, Gelfand DH. Reverse transcription and DNA amplification by *Thermus thermophilus* DNA polymerase. Biochem 1991;30:7661–6.
4. Brooks EM, Sheflin LG, Spaulding SW. Secondary structure in the 3′UTR of EGF and the choice of reverse transcriptases affect the detection of message diversity by RT-PCR. Biotechniques 1995;19:806–15.
5. Kuo KW, Leung M, Leung WC. Intrinsic secondary structure of human TNFR-I mRNA influences the determination of gene expression by RT-PCR. Mol Cell Biochem 1997;177:1–6.
6. Pfaffl MW, Hageleit M. Validities of mRNA quantification using recombinant RNA and recombinant DNA external calibration curves in real-time RT-PCR. Biotechnol Letters 2001;23:275–82.
7. Reist M, Pfaffl MW, Morel C, Meylan M, Hirsbrunner G, Blum JW, et al. Quantitative mRNA analysis of bovine 5-HT receptor subtypes in brain, abomasums, and intestine by real-time PCR. J Recept Signal Transduct Res 2003;23:271–87.
8. Ståhlberg A, Åman P, Ridell B, Mostad P, Kubista M. Quantitative real-time PCR method for detection of B-lymphocyte monoclonality by comparison of $\kappa$ and $\lambda$ immunoglobulin light chain expression. Clin Chem 2003;49:51–9.
9. Köhler T. Design of suitable primers and competitor fragments for quantitative PCR. In: Köhler T, Lassner D, Rost AK, Thamm B, Pustowoit B, Remke H, eds. Quantitation of mRNA by polymerase chain reaction—nonradioactive PCR methods. Heidelberg: Springer-Verlag, 1995;1.2:15–26.
10. Köhler T, Lerche D, Meye A, Weisbrich C, Wagner O. Automated analysis of nucleic acids by quantitative PCR using DNA coated ready-to-use reaction tubes. J Lab Med 1999;23:408–14.
11. Southern EM, Kalim UM. Determining the influence of structure on hybridization using oligonucleotide arrays. Nat Biotechnol 1999;17:788–92.
12. Sohail M, Southern EM. Hybridization of antisense reagents to RNA. Curr Opin Mol Ther 2000;2:264–71.
13. Peccoud J, Jacob C. Theoretical uncertainty of measurements using quantitative polymerase chain reaction. Biophys J 1996;71:1001–8.
14. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 2001;29:e45.
15. Curry J, McHale C, Smith MT. Low efficiency of the Moloney murine leukemia virus reverse transcriptase during reverse transcription of rare t(8;21) fusion gene transcripts. Biotechniques 2002;32:768–75.

DOI: 10.1373/clinchem.2004.035469

---

## Multiplexed Real-Time PCR Using Universal Reporters,

*Andreas M. Rickert, Hans Lehrach, and Silke Sperling** (Max-Planck-Institute for Molecular Genetics, Berlin, Germany; * address correspondence to this author at: Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; fax 49-30-8413-1128, e-mail sperling @molgen.mpg.de)

Real-time quantitative PCR is a sensitive and accurate method for gene expression studies *(1)*. The detection chemistries of all real-time PCR procedures are based on one of two principles for monitoring amplification products: binding to double-stranded DNA or hybridization to single-stranded DNA. Small molecules bind to double-stranded DNA either as intercalators or as minor groove binders, e.g., ethidium bromide *(2)*, Hoechst 33258 *(3)*, or SYBR® Green I *(4)*. Several approaches using target-specific hybridization to single-stranded DNA have been introduced, including Molecular Beacons *(5)*, Scorpions *(6, 7)*, the TaqMan or hydrolysis/5′-nuclease assay *(8, 9)*, the AEGIS probe system *(10)*, labeled primers *(11, 12)*, and light-up probes *(13)*. In contrast to binding of dyes to double-stranded DNA, these methods are suitable for multiplexing approaches because they use differentially labeled fluorescent dyes. However, as they require a unique probe or modified primer for each target, currently used hybridization-based methods for real-time quantitative PCR have high reagent costs and require large developmental efforts.

Here we present a real-time PCR assay that uses universal hybridization-based probe sets suitable for any target. Because the assay uses tailed locus-specific non-modified amplification primers, PCR products can be

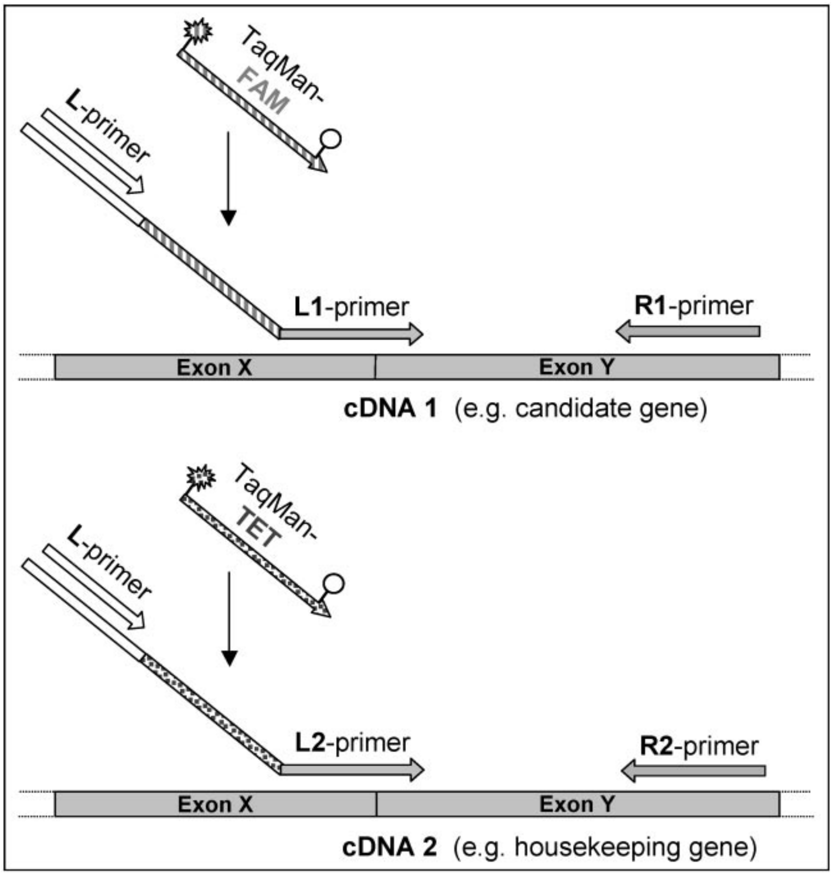


Fig. 1. Scheme of the multiplexed cr-real-time PCR assay.

Each gene of interest is amplified with a set of three primers: one tailed locus-specific forward oligonucleotide (L*1/2*-primer), one locus-specific reverse primer (R*1/2*-primer), and one common primer (L-primer) annealing to a common sequence stretch of the tails. Because the sequence tags of the tails vary, amplification products of different cDNAs are specifically detected by use of TaqMan probes corresponding to the appropriate tags. As illustrated, the design of exon-junction-spanning primers appears advisable, although it is not mandatory for the assay. *Open arrows* indicate common sequences and primers; *gray arrows* indicate locus-specific sequences and primers; *patterned arrows* indicate TaqMan probes and their corresponding sequence stretches in the tailed oligonucleotides; the *heads* of the *arrows* indicate the 5′ ends of sequences, and the *tails* of the *arrows* indicate the 3′ ends of sequences. *FAM*, 6-carboxyfluorescein; *TET*, tetrachloro-6-carboxyfluorescein.

monitored via common reporters (cr) hybridizing to the common tails. The general principle of combining tailed PCR primers with universal probes has been introduced for other genetic applications, such as single-nucleotide polymorphism genotyping *(14, 15)* and in situ amplification *(16)*, but these methods have not been applied to quantitative gene expression studies. Our system, which is similar to the method developed by Whitcombe et al. *(15)*, leads to a more flexible and low-cost setup than conventional hybridization-based approaches. Using differentially labeled universal reporting reagents, we have developed a multiplex setup for simultaneous analysis of target gene and internal control (housekeeping gene), which is an accepted method for normalizing sample-to-sample variation. We then compared the cr-real-time PCR assay with SYBR Green I assays with respect to robustness and sensitivity. As a proof of principle, we applied the new assay to a study of three previously published, differentially expressed candidate genes for congenital heart defects *(17)*.

The principle of the cr-real-time PCR assay is as follows: Target DNA is subjected to PCR in the presence of four oligonucleotides; one low-concentration, tailed, locus-specific primer; one locus-specific primer without a tail; one common primer annealing to a common sequence stretch of the tail; and one universal TaqMan probe corresponding to another common part of the tail (Fig. 1 and Table 1). The use of locus-specific primers that span at least one intron of the genomic sequence minimizes problems associated with DNA contamination. During the first amplification cycle, the tailed locus-specific primer initiates the polymerase reaction, leading to the synthesis of a fragment with the tail at the 5′ end. In the second cycle, the complement of the tail is synthesized. Starting from cycle three, the common amplification primer primes synthesis on the fragment including the tail sequence. From this step on, the TaqMan probe anneals to the products resulting from the amplification with the common primer and the locus-specific primer without a tail. The TaqMan probe bound to the tailed target amplicon is hydrolyzed by the 5′-nuclease activity of the DNA polymerase, leading to a physical separation of the reporter and quencher dye and release of fluorescence emission *(8)*.

Introduction of further common tails and corresponding, differentially labeled TaqMan probes opens up the possibility for multiplexing (Fig. 1), although the degree of multiplexing is limited by the number of different dyes and the restrictions of currently available instruments *(1)*. The universal tails and TaqMan probes can be easily transferred to other targets in combination with two locus-specific primer sequences. In cases in which locus-specific primers are already designed, ligation to their corresponding common tails by use of the ligation-based synthesis method *(18)* could be envisaged. Furthermore, the amplification of primer-dimers or pseudogenes can be

**Table 1. Oligonucleotides used for cr-real-time PCR assays.**[a]

| Name | Sequence, 5→3′ | Gene |
|---|---|---|
| L1_PIPPIN_F | [tail-L1]-ACCAGGACCTATTCAGCGACA | *PIPPIN* |
| PIPPIN_R | AACTGCTTACAGACGCCCTTG | |
| L1-FLJ10350 | [tail-L1]-CTCAGTGGAGTCTCCCAAGCAA | *FLJ10350* |
| FLJ10350_R | TGTTCGGCTCAGACTCTTGTCC | |
| L1_TNNL1_F | [tail-L1]-TGGATGAGGAGCGATACGACAT | *TNNL1* |
| TNNL1_R | GGTCCTTAATCTCCCTGGTGTTG | |
| L2_B2M_F | [tail-L2]-TGCTGTCTCCATGTTTGATGTATCT | *B2M* |
| B2M_R | TCTCTGCTCCCCACCTCTAGGT | |
| tail-L1 | TGCACAATTCACGACTCACGAT*CCACACGGTCTCGCACTGGC*ACGGG | |
| tail-L2 | TGCACAATTCACGACTCACGAT*CATCCGCTCCGACGACACGA*ACGGG | |
| TaqMan-fam | FAM-*CCACACGGTCTCGCACTGGC*-TAMRA | |
| TaqMan-tet | TET-*CATCCGCTCCGACGACACGA*-TAMRA | |
| L-primer | GCACAATTCACGACTCACGA | |

[a] Each assay uses two locus-specific amplification primers, one common primer, and one universal TaqMan probe (see Fig. 1). Common primers and their corresponding sequence stretches of the tails are underlined, TaqMan probes and their corresponding sequence stretches of the tails are in italics. TaqMan probes were labeled with the reporter dyes 6-carboxyfluorescein (FAM) or tetrachloro-6-carboxyfluorescein (TET) at the 5′ end and with the quencher dye 6-carboxytetramethylrodamine (TAMRA) at the 3′ end.

tested by a melting analysis using SYBR Green I instead of the TaqMan probe for amplicon detection.

Because the cr-real-time PCR assay requires three interacting amplification primers per target gene (Fig. 1), the optimum ratio had to be adjusted. For separate analysis, the 20-μL reactions contained 1× TaqMan Universal PCR Master Mix (Applied Biosystems), 0.3 μM appropriate TaqMan probe, 0.4 μM each of the reverse amplification primer (R*1*/*2*-primer) and common forward primer (L-primer), 0.004–0.2 μM tailed forward primer (L*1*/*2*-primer), and various amounts of template. In the multiplexed analysis, the 20-μL assay mixture contained 1× TaqMan Universal PCR Master Mix (Applied Biosystems), 0.3 μM each of both universal TaqMan probes, 0.4 μM each of both locus-specific reverse primers (R-primers), 0.04–0.2 μM both tailed forward primers (L*1*/*2*-primer), 0.4–1.0 μM the common forward primer (L-primer), and various amounts of plasmid of cDNA templates. Cr-real-time PCRs were performed and measured on an ABI Prism 7900HT system. The thermocycling protocol consisted of an initial denaturation at 95 °C for 10 min, followed by 45 cycles of 95 °C for 15 s and 60 °C for 1 min. Subsequently, a dissociation curve was generated in the range of 60–95 °C.

A concentration of 0.016 μM of the tailed locus-specific primer was sufficient as this primer is required only during the initial steps. No signal improvement was achieved with higher concentrations, whereas amplifications were less efficient with lower concentrations. The concentrations of the other two amplification primers were optimally kept at 0.4 μM each. Transferring these assay conditions for separate analyses to multiplexing approaches revealed preferential amplification of one target gene over the other. Given the competitive nature of multiplexed reactions, an increase in concentration of the universal primer necessary for all targets (L-primer) to 0.8 μM appeared crucial. When we used these conditions, the same amplification results were obtained independently from the uniplex or multiplex background of the reaction (see the Data Supplement that accompanies the online version of this Technical Brief at http://www.clinchem.org/content/vol50/issue9/).

The same assay conditions could be transferred to all loci of interest without any further optimization effort (see the online Data Supplement). This successful application of the same reaction conditions demonstrated the robustness and flexibility of the cr-real-time PCR assay. Because each reaction was performed in triplicate, showing only marginal variations (see the online Data Supplement), the cr-real-time PCR demonstrated high reproducibility and accuracy, which were in the same range as the results obtained here (see the online Data Supplement) and the results reported previously *(19, 20)* for the SYBR Green I assay.

Two different setups were performed using SYBR Green I for real-time detection of amplification products. The two-primer setup contained 1× SYBR Green PCR Master Mix (Applied Biosystems), 0.4 μM forward and reverse primer (L*1*/*2*-primer without common tail and R*1*/*2*-primer), and various amounts of plasmid or cDNA templates. To simulate the amplification conditions of the cr-real-time PCRs, the three-primer setup was performed with an amplification mixture containing 1× SYBR Green PCR Master Mix, 0.4 μM each reverse primer (R*1*/*2*-primer) and common forward primer (L-primer), 0.016 μM tailed forward primer (L*1*/*2*-primer), and various amounts of template. All 20-μL amplification reactions were carried out and measured on an ABI Prism 7900HT system (Applied Biosystems), using the same thermal profile as described for the cr-real-time PCR assay.

To evaluate the sensitivity and dynamic range of the cr-real-time PCR assay, we prepared two serial dilutions and subjected various amounts to cr-real-time PCRs as well as to two- and three-primer SYBR Green I assays. A dilution series of total cDNA ranging from 1 (corresponding to 100 ng of reverse-transcribed RNA) to 1:50 000 (corresponding to 2 pg of reverse-transcribed RNA) was used to quantitatively target the housekeeping gene *B2M*. The three assays were highly linear over the examined range (see the online Data Supplement). Likewise, in a 10-fold dilution series of a *FLJ10350* cDNA clone ranging from $10^9$ to 10 copies per reaction, assayed by all three assays, real-time quantification was linear, although it showed more deviation among the triplicates performed on low amounts of template (see the online Data Supplement).

We analyzed expression of three genes that had been found, by SYBR Green I analyses, to be differentially expressed in patients with congenital heart defects *(17)*. The results confirmed the previously published data, with *FLJ10350* and *TNNI1* being significantly up-regulated and *PIPPIN* being significantly down-regulated (see the online Data Supplement). Throughout our assays, we saw no amplification of the no-template controls (see the online Data Supplement).

For normalization of the target genes analyzed in the course of this study, the housekeeping gene *B2M* was simultaneously assayed with the genes of interest. The obvious sample-to-sample variations (see the online Data Supplement) stress the importance of effective systems for normalization, as achieved with the multiplexed cr-real-time PCR assay.

In summary, we have described a single-step method for real-time PCR that is sensitive, robust, and requires minimal optimization effort. Because the system uses nonmodified, tailed amplification primers and universal reporting reagents, it is characterized by a flexible and low-cost format. The use of differentially labeled reporting reagents enables multiplexing approaches for monitoring of more than one target per well, e.g., both a candidate and housekeeping gene. Therefore, the cr-real-time PCR assay appears suitable for the broad spectrum of real-time PCR applications.

This work was funded by EU FP6 Grant 503155 (Moltools). We are grateful to Martin Lange for providing the *FLJ10350* cDNA clone and Bogac Kaynak for providing the cDNA samples. We thank Maike Tribbels and Christina Grimm for critical reading of the manuscript.

## References

1. Giulietti A, Overbergh L, Valckx D, Decallonne B, Bouillon R, Mathieu C. An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. Methods 2001;25:386–401.
2. Higuchi R, Dollinger G, Walsh PS, Griffith R. Simultaneous amplification and detection of specific DNA sequences. Biotechnology 1992;10:413–7.
3. Nielsen PE. Sequence-selective DNA recognition by synthetic ligands. Bioconj Chem 1991;2:1–12.
4. Morrison TB, Weis JJ, Wittwer CT. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques 1998;24:954–8, 960, 962.
5. Tyagi S, Bratu DP, Kramer FR. Multicolor molecular beacons for allele discrimination. Nat Biotechnol 1998;16:49–53.
6. Nazarenko IA, Bhatnagar SK, Hohman RJ. A closed tube format for amplification and detection of DNA based on energy transfer. Nucleic Acids Res 1997;25:2516–21.
7. Whitcombe D, Theaker J, Guy SP, Brown T, Little S. Detection of PCR products using self-probing amplicons and fluorescence. Nat Biotechnol 1999;17:804–7.
8. Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Natl Acad Sci U S A 1991;88: 7276–80.
9. Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. PCR Methods Appl 1995;4:357–62.
10. Moser MJ, Marshall DJ, Grennier JK, Kiefer CD, Killeen AA, Ptacin JL, et al. Exploiting the enzymatic recognition of an unnatural base pair to develop a universal genetic analysis system. Clin Chem 2003;49:407–14.
11. Nazarenko I, Lowe B, Darfler M, Ikonomi P, Schuster D, Rashtchian A. Multiplexed quantitative PCR using self-quenched primers labeled with single fluorophore. Nucleic Acid Res 2002;e37.
12. Crockett AO, Wittwer CT. Fluorescein-labeled oligonucleotides for real-time PCR: using the inherent quenching of deoxyguanosine nucleotides. Anal Biochem 2001;290:89–97.
13. Svanvik N, Stahlberg A, Sehlstedt U, Sjöback R, Kubista M. Detection of PCR products in real time using light-up probes. Anal Biochem 2000;287:179–82.
14. Myakishev MV, Khripin Y, Hu S, Hamer DH. High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. Genome Res 2001;11:163–9.
15. Whitcombe D, Brownie J, Gillard HL, McKechnie D, Theaker J, Newton CR, et al. A homogeneous fluorescence assay for PCR amplicons: its application to real-time, single-tube genotyping. Clin. Chem 1998;44:918–23.
16. Nuovo GJ, Hohman RJ, Nardone GA, Nazarenko IA. In situ amplification using universal energy transfer-labeled primers. J Histochem Cytochem 1999;47: 273–80.
17. Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, et al. Genome-wide array analysis of normal and malformed human hearts. Circulation 2003;107:2467–74.
18. Borodina TA, Lehrach H, Soldatov AV. Ligation-based synthesis of oligonucleotides with block structure. Anal Biochem 2003;318:309–13.
19. Mikula M, Dzwonek A, Jagusztyn-Krynicka K, Ostrowski J. Quantitative detection for low levels of *Helicobacter pylori* infection in experimentally infected mice by real-time PCR. J Microbiol Methods 2003;55:351–9.
20. Ponchel F, Toomes C, Bransfield K, Leong FT, Douglas SH, Field SL, et al. Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. BMC Biotechnol 2003;3:18.

DOI: 10.1373/clinchem.2004.034512

---

### *ACE2* Gene Polymorphisms Do Not Affect Outcome of Severe Acute Respiratory Syndrome,

*Rossa W.K. Chiu,[1,2] Nelson L.S. Tang,[1,2] David S.C. Hui,[1,3] Grace T.Y. Chung,[1,2] Stephen S.C. Chim,[1,2] K.C. Allen Chan,[1,2] Ying-man Sung,[2] Louis Y.S. Chan,[4] Yu-kwan Tong,[1,2] Wing-shan Lee,[1,2] Paul K.S. Chan,[1,5] and Y.M. Dennis Lo[1,2*]* ([1] The Centre for Emerging Infectious Diseases, and Departments of [2] Chemical Pathology, [3] Medicine and Therapeutics, [4] Obstetrics and Gynaecology, and [5] Microbiology, The Chinese University of Hong Kong, Shatin, Hong Kong; * address correspondence to this author at: Department of Chemical Pathology, The Chinese University of Hong Kong, Room 38023, 1/F Clinical Sciences Bldg., Prince of Wales Hospital, 30-32 Ngan Shing St., Shatin, New Territories, Hong Kong Special Administrative Region, China; fax 852-2194-6171, e-mail loym@cuhk.edu.hk)

Severe acute respiratory syndrome (SARS) is the first pandemic of the 21st century *(1)*. Since its recognition, 8437 individuals have been affected and 813 have died *(2)*. Approximately 20–30% of patients required intensive care admission *(1)*. Although there was a slight predom-

### 2.3.2 Refined genotyping for sparse sample material – LDR-TaqMan

> Rickert AM, Borodina TA, Kuhn EJ, Lehrach H, **Sperling S**. Refinement of single-nucleotide polymorphism genotyping methods on human genomic DNA: amplifluor allele-specific polymerase chain reaction versus ligation detection reaction-TaqMan. *Anal Biochem* 2004;**330**:288-297.

Single-nucleotide polymorphisms (SNPs) represent the most frequent DNA sequence variations accounting for 90% of all polymorphisms (Collins et al., 1997), and they occur with an approximate frequency of one every kilobase in the human genome (Li and Sadler, 1991; Sachidanandam et al., 2001). Due to the high density of this type of genetic variations, SNPs have become popular markers for a wide range of molecular genetic studies, including both diagnostic and research settings for analysis of human diseases.

However, the valuable use of SNPs requires appropriate methods for genotyping SNPs. Thus, a vast number of SNP genotyping methods have been introduced, but no single method has been widely accepted. In general, each method comprises two elements. First, direct biological, chemical, or physical interaction with the alleles of a SNP is essential. This sequence-specific detection and allele-specific discrimination is based on one of the following procedures: hybridization, invasive cleavage, oligonucleotide ligation, or primer extension, which is either using allele-specific nucleotide incorporation or allele-specific PCR (Kwok, 2001). Secondly, allele-specific products are analyzed with mechanisms usually relying on gel separation, arrays, mass spectrometer, or fluorescence plate reader (Gut, 2001). Almost all possible combinations of techniques belonging to the two procedure elements have been introduced as SNP genotyping methods, with each of them showing specific shortcomings (Gut, 2001). Thus, the choice of the primary method varies from application to application and depends on the number of SNPs and individuals to be genotyped, expertise, time exposure, assay costs, and availability of equipment.

Recently, the LDR-TaqMan method has been introduced for genotyping 30 SNPs of the *Arabidopsis thaliana* genome (Borodina et al., 2004). This technique utilizes the specificity of ligation detection reactions (LDRs) due to their preferred ligation of perfectly matched nicks in contrast to mismatched ones (Landegren et al., 1988; Nickerson et al., 1990; Barany, 1991). During LDR, tailed detector oligonucleotides (DOs) are ligated, and subsequently the allele-specific ligation-products are identified via TaqMan PCR amplification (5'-nuclease assay) (Holland et al., 1991; Livak et al., 1995) using universal primers and probes. In the following the LDR-TaqMan method could be adjusted for genotyping SNPs on the approximately 20-fold more complex human genome (**Figure 16**).

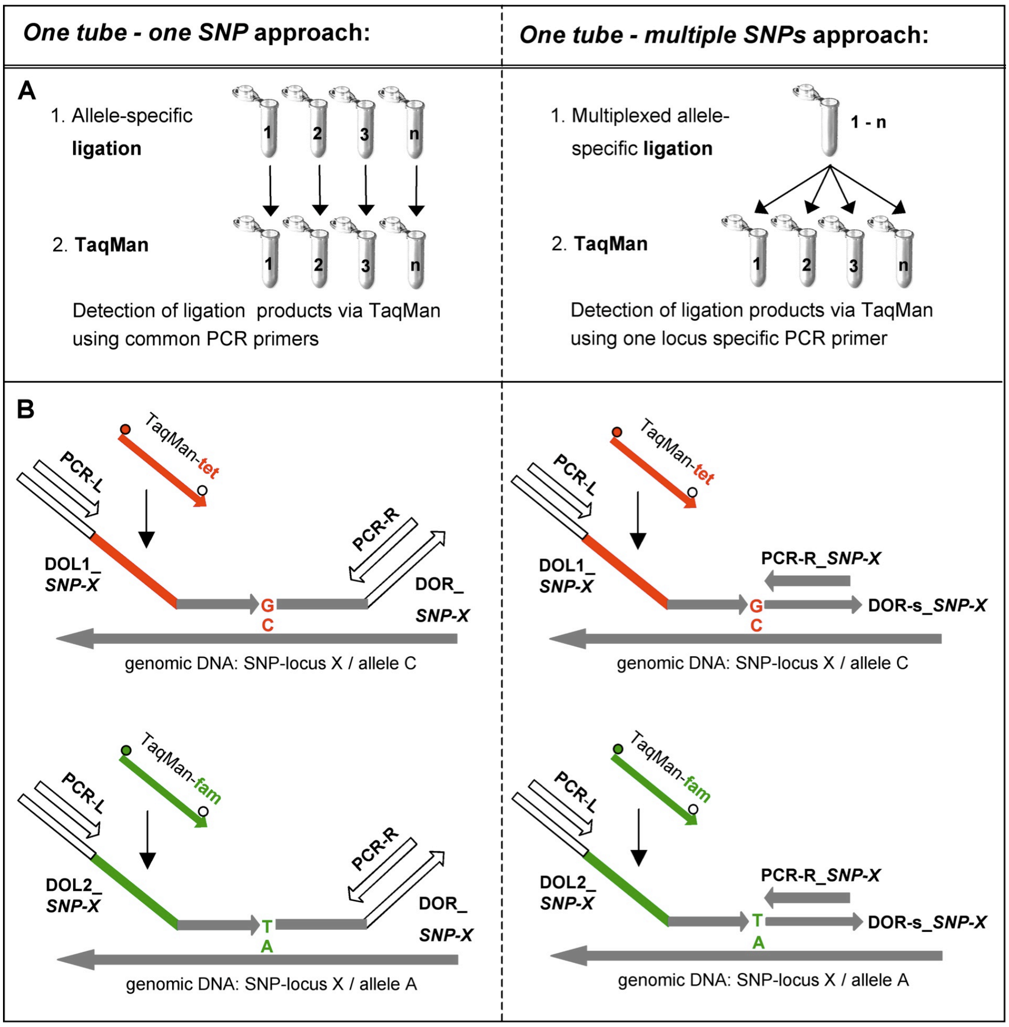

*One tube - one SNP* approach:
*One tube - multiple SNPs* approach:
A
1. Allele-specific **ligation**
1
2
3
n
2. **TaqMan**
Detection of ligation products via TaqMan using common PCR primers
1. Multiplexed allele-specific **ligation**
1 - n
Detection of ligation products via TaqMan using one locus specific PCR primer
B
TaqMan-*tet*
PCR-L
PCR-R
DOL1_ *SNP-X*
DOR_ *SNP-X*
G
C
genomic DNA: SNP-locus X / allele C
PCR-R_*SNP-X*
DOR-s_*SNP-X*
TaqMan-*fam*
DOL2_ *SNP-X*
T
A
genomic DNA: SNP-locus X / allele A

**Figure 16. LDR-TaqMan genotyping assay.** (A) The LDR-genotyping assay comprises two reaction steps. First, tailed, locus- and allele-specific detector oligonucleotides (DOs) are ligated. In case of the *one tube – one SNP* approach each SNP is separately analyzed. Applying the *one tube – multiple SNPs* approach, all SNP-loci are targeted on the same aliquot of genomic DNA in a single ligation reaction. Subsequently in a second step, the ligation-products are subjected to TaqMan PCR. During the PCR amplifications the alleles are determined due to TaqMan probes detecting the allele-specific ligation-products. Having performed separate ligation reactions, the following TaqMan PCR reactions only require common amplification primers (left). Ligation reaction products from the *one tube – multiple SNPs* approach are distributed to different tubes, and are amplified in single tubes using locus-specific PCR primers (right). (B) Required oligonucleotides are shown for separate (left) and multiplexed (right) LDR-TaqMan genotyping assays for a C/A-SNP (*SNP-X*). Corresponding to the genomic DNA a ‘left’ tailed, allele-specific DO (DOL-1_*SNP-X* or DOL-2_*SNP-X*) is ligated with a ‘right’ tailed (*one tube – one SNP* approach: DOR_*SNP-X*) or un-tailed DOR (*one tube – multiple SNPs* approach: DOR-s_*SNP-X*). During PCR ligation-products are amplified using the primer pairs PCR-L and PCR-R (*one tube – one SNP* approach) or PCR-L and PCR-R_*SNP-X* (*one tube – multiple SNPs* approach). White arrows indicate common sequences and primers; gray arrows indicate locus-specific sequences and primers; green and red arrows indicate TaqMan probes and their corresponding sequence stretches in the DOLs; arrow endings indicate 5’-ends of sequences; and blunt endings indicate 3’-ends of sequences.

To genotype human samples using LDR-TaqMan, it appeared beneficial to design detector oligonucleotides with 5°C higher melting temperatures concerning the locus-specific sequence stretches and to double the concentration of DOs to 2 nM each for the ligation reaction. These adjustments led to higher signal intensities, indicating higher ligation efficiency. Furthermore, the previously three-step multiplexing procedure for LDR-TaqMan was optimized to a less laborious one tube – multiple SNPs approach based on multiplexed allele-specific ligation directly followed by separate TaqMan PCRs (**Figure 16**). The genotyping of 18 related human DNA samples for four not preselected SNPs showed a high accuracy and robustness of the method. In conclusion, based on the evaluation and optimization, LDR-TaqMan is a easy to establish and low-cost SNP genotyping technique, suitable for research and clinical settings.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytical Biochemistry 330 (2004) 288–297
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

# Refinement of single-nucleotide polymorphism genotyping methods on human genomic DNA: amplifluor allele-specific polymerase chain reaction versus ligation detection reaction-TaqMan

Andreas M. Rickert, Tatiana A. Borodina, Eckehard J. Kuhn, Hans Lehrach, and Silke Sperling*

*Max-Planck-Institute for Molecular Genetics, 14105 Berlin, Germany*

Received 19 January 2004
Available online 10 May 2004

## Abstract

Single-nucleotide polymorphisms (SNPs) have proven to be powerful genetic markers for a variety of genetic applications, e.g., association studies leading to dissection of both monogenetic and complex diseases. However, no single SNP genotyping method has been broadly accepted. In the present study, we compared and refined two promising methods with potential for research and for diagnostic SNP genotyping: Amplifluor allele-specific polymerase chain reaction (PCR) and ligation detection reaction (LDR)-TaqMan. The methods are based on allele-specific primer extension and allele-specific ligation, respectively. Since LDR-TaqMan had previously been tested on just *Arabidopsis thaliana*, we adjusted the method for the more complex human genome. Amplifluor allele-specific PCR has a single-step and closed-tube format, whereas the LDR-TaqMan assay comprises two simple steps. Contrary to the primer-extension-based method, the ligation-based method can be multiplexed. Refining the LDR-TaqMan technique, we successfully replaced a previously suggested three-step multiplexing procedure with a less laborious two-step approach. Comparing refined LDR-TaqMan with Amplifluor allele-specific PCR in a family-based study, both techniques appeared similar with respect to high robustness and accuracy. As both approaches utilize primers with common tails, all SNPs can be assayed with the same couple of fluorescence reporting reagents, ensuring low establishing and running expenses.
© 2004 Elsevier Inc. All rights reserved.

*Keywords:* Single-nucleotide polymorphism; SNP; Genotyping; Amplifluor allele-specific PCR; LDR-TaqMan; Multiplexing

Single-nucleotide polymorphisms (SNPs)[1] represent the most frequent DNA sequence variations, accounting for 90% of all polymorphisms [1], and they occur with an approximate frequency of one per kilobase in the human genome [2,3]. To date, nearly 1.8 million SNPs have been discovered and characterized in man and deposited to public databases (http://snp.cshl.org;http://www.ncbi.nlm.nih.gov/SNP/). Due to the high density of this type of genetic variation, SNPs have become popular markers for a wide range of molecular genetic studies, including both diagnostic and research settings for analysis of human diseases.

However, the use of SNPs requires appropriate methods for genotyping SNPs. Thus, a vast number of SNP genotyping methods have been introduced, but no single method has been widely accepted [4–10]. In general, each method comprises two elements. First, direct biological, chemical, or physical interaction with the alleles of a SNP is essential. This sequence-specific detection and allele-specific discrimination is based on one of the following procedures: hybridization, invasive cleavage, oligonucleotide ligation, or primer extension using either allele-specific nucleotide incorporation or allele-specific PCR [7]. Second, allele-specific products are analyzed with techniques relying usually on gel separation, arrays, a mass spectrometer, or a fluorescence plate reader [5].

---

* Corresponding author. Fax: +49-30-8413-1380.
*E-mail address:* sperling@molgen.mpg.de (S. Sperling).

[1] *Abbreviations used:* SNP, single-nucleotide polymorphism; LDR, ligation detection reaction; DO, detector oligonucleotide; ASPs, allele-specific primers; CRPs, common reverse primers.

0003-2697/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2004.03.035

Almost all possible combinations of techniques belonging to the two procedure elements have been introduced as SNP genotyping methods, with each showing specific shortcomings [5]. Thus, the choice of the primary method varies from application to application and depends on the number of SNPs and individuals to be genotyped, expertise, time of exposure, assay costs, and availability of equipment. For example, methods considered suitable for routinely performed diagnostics outside of well-equipped research institutions should be robust, reliable, easy to handle, and low cost, without the need of expensive instrumentation.

Recently, two methods that potentially meet these demands have been published. The Amplifluor assay [11–13] combines allele-specific PCR [14–16] for interrogating SNPs, with the use of universal Amplifluor energy-transfer-labeled primers [17,18] for detection of allele-specific PCR products. More recently the LDR-TaqMan method has been introduced for genotyping 30 SNPs of the *Arabidopsis thaliana* genome [19]. This technique utilizes the specificity of ligation detection reactions due to their preferred ligation of perfectly matched nicks in contrast to mismatched nicks [20–23]. During LDR, tailed detector oligonucleotides (DOs) are ligated, and subsequently the allele-specific ligation products are identified via TaqMan PCR amplification (5′-nuclease assay) [24,25] using universal primers and probes (Fig. 1).

Here, we report the comparative application of these two promising methods for genotyping four not-preselected SNPs in a family-based analysis. In the course of this study, we adjusted the LDR-TaqMan method for genotyping SNPs on human genomic DNA and developed a simplified strategy for multiplexing the LDR-TaqMan procedure.

## Materials and methods

### DNA samples and SNPs

Genomic DNA of 18 related individuals of current interest to our group (referred to as F1-family) was extracted from blood samples as described before [26]. DNA concentrations were measured using a UV spectrophotometer (Ultrospec 2100 pro, Amersham Biosciences, Freiburg, Germany).

The SNPs c14, c15, s56, and s63 (Table 1) had been identified and verified by comparative Sanger DNA sequencing of target sequences in the different F1-family members. Fragments including the polymorphic sites were PCR amplified using primers shown in Table 1. PCR amplifications and purification of products were performed as described before [27]. Using the amplimers as sequencing primers, both strands were sequenced according to the Big-Dye chemistry reaction protocol (Applied Biosystems, Foster City, CA, USA).

All oligonucleotides were purchased form MWG (Ebersberg, Germany), if not stated otherwise.

### Amplifluor allele-specific PCR

The Amplifluor genotyping assay is based on PCR amplification in the presence of five oligonucleotides (Table 2). Tailed allele-specific primers (ASPs) and common reverse primers (CRPs) were designed using the Amplifluor Assay Architect software (http://www.assayarchitect.com), so that the melting temperatures of the locus-specific stretches of the ASPs were approximately 64 °C and those of the CPRs were about 3 °C higher. Universal Amplifluor primers were labeled with the acceptor DABSYL (4-(dimethylamino)azo benzene sulfonic acid) and the fluorophores FAM (fluorescein) and JOE (6-carboxy-4′,5′-dichloro-2′,7′-dimethoxyfluorescein), respectively (Serologicals, Temecula, CA, USA).

The 10-μl amplification reaction contained 250 nM FAM- and JOE-labeled Amplifluor primers, 40 nM tailed allele-specific primers, 500 nM reverse primer, 1× reaction buffer (Tris–HCl, pH 8.3, 10 mM; KCl, 50 mM; $MgCl_2$, 1.8 mM) (Serologicals), 0.2 mM dNTPs, 0.5 U of HotStar *Taq* Polymerase (Qiagen, Hilden, Germany), and varying amounts of genomic DNA (0.5 to 100 ng).

Two alternative thermal amplification profiles were performed on a ABI Prism 7900HT system (Applied Biosystems, Foster City, CA, USA). One amplification protocol comprised two cycling units [28]: 96 °C, 10 min; (95 °C, 15 s; 56 to 64 °C, 10 s; 72 °C, 15 s) × 20 cycles; (95 °C, 15 s; 56 °C, 30 s; 72 °C, 40 s) × 20 to 25 cycles; 72 °C, 3 min; 15 °C hold. The second PCR profile included one cycling unit [11–13]: 96 °C, 10 min; (95 °C, 30 s; 56 to 64 °C, 30 s; 72 °C, 40 s) × 35 to 45 cycles; 72 °C, 3 min; 15 °C hold.

Amplification signals were analyzed via real-time and via endpoint measurements using the ABI Prism 7900HT system (Applied Biosystems).

### LDR-TaqMan

The LDR-TaqMan genotyping assay consists of two reaction steps, allele-specific ligation followed by allele-discriminatory TaqMan PCR amplification. We performed both *one tube–one SNP* and *one tube–multiple SNPs* approaches, which have most components in common (Fig. 1).

*Preparation of detector oligonucleotides.* For each SNP assay three different DOs were designed, two “left” DOs (DOLs) and one “right” DO (DOR) (Fig. 1B, Table 3). The locus-specific sequence stretches were selected so that the annealing temperatures were approximately 60 °C for the DOLs and 65 °C for the DORs for 4000 pM primers and 50 mM salt using the program Vector NTI 7.1 (InfoMax, Frederick, MD, USA). For the SNPs c15 and s56 additional DO sets were designed with lower

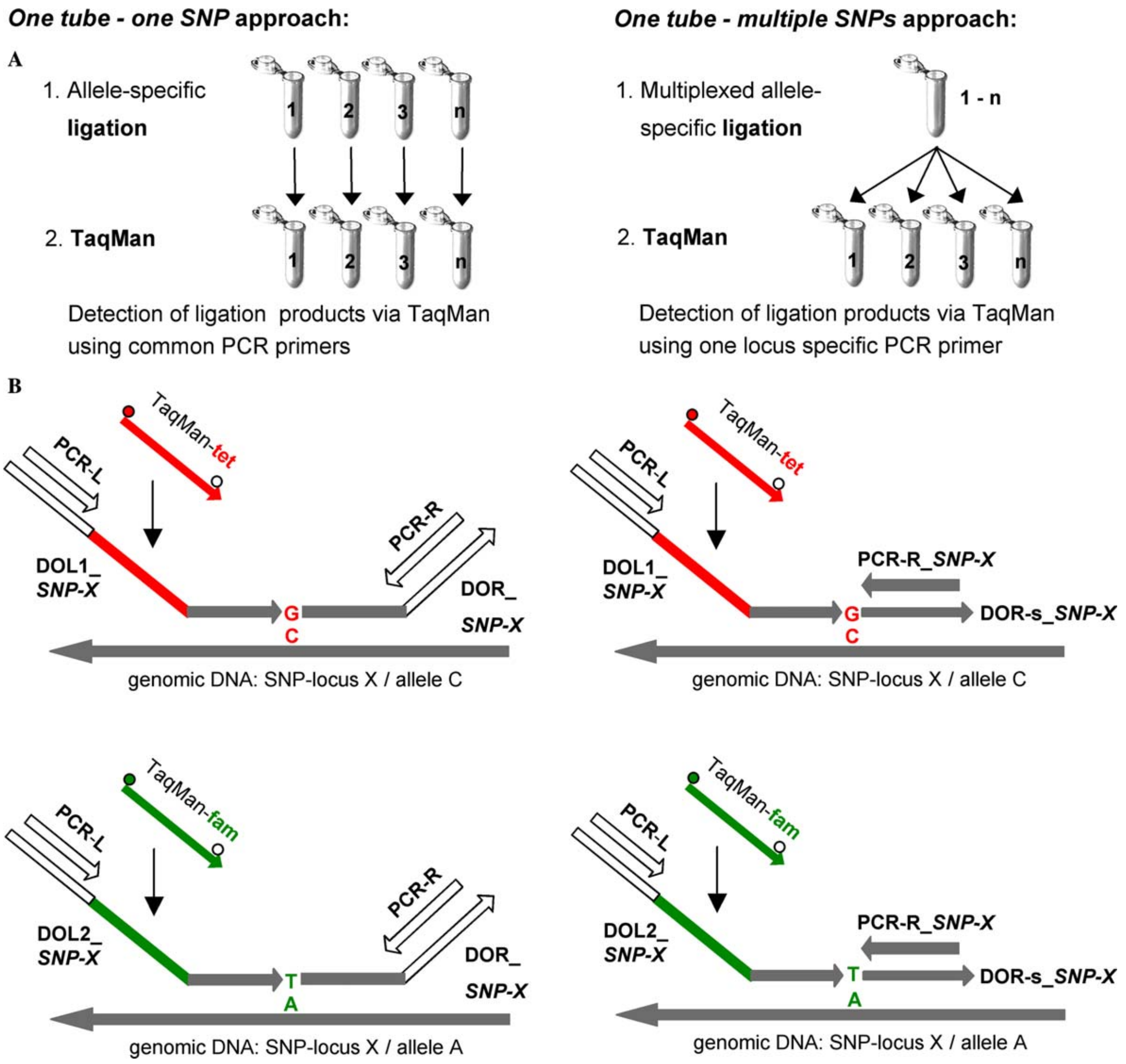

One tube - one SNP approach:
One tube - multiple SNPs approach:
A
1. Allele-specific ligation
1
2
3
n
2. TaqMan
Detection of ligation products via TaqMan using common PCR primers
1. Multiplexed allele-specific ligation
1 - n
2. TaqMan
Detection of ligation products via TaqMan using one locus specific PCR primer
B
TaqMan-tet
PCR-L
DOL1_SNP-X
PCR-R
DOR_SNP-X
G
C
genomic DNA: SNP-locus X / allele C
PCR-R_SNP-X
DOR-s_SNP-X
TaqMan-fam
DOL2_SNP-X
T
A
genomic DNA: SNP-locus X / allele A

Fig. 1. Schematic presentation of the LDR-TaqMan genotyping assay applying the *one tube–one SNP* (left) and the two-step *one tube–multiple SNPs* approach (right). (A) The LDR genotyping assay comprises two reaction steps. First, tailed, locus- and allele-specific detector oligonucleotides (DOs) are ligated. In case of the *one tube–one SNP* approach each SNP is separately analyzed. Applying the *one tube–multiple SNPs* approach, all SNP loci are targeted on the same aliquot of genomic DNA in a single ligation reaction. Subsequently in a second step, the ligation products are subjected to TaqMan PCR. During the PCR amplifications the alleles are determined due to TaqMan probes detecting the allele-specific ligation products. Having performed separate ligation reactions, the following TaqMan PCRs require only common amplification primers (left). Ligation reaction products from the *one tube–multiple SNPs* approach are distributed to different tubes and amplified in single tubes using locus-specific PCR primers (right). (B) Required oligonucleotides are shown for separate (left) and multiplexed (right) LDR-TaqMan genotyping assays for a C/A-SNP (*SNP-X*). Corresponding to the genomic DNA a "left" tailed, allele-specific DO (DOL-1_*SNP-X* or DOL-2_*SNP-X*) is ligated with a "right" tailed (*one tube–one SNP* approach: DOR_*SNP-X*) or untailed DOR (*one tube–multiple SNPs* approach: DOR-s_*SNP-X*). During PCR, ligation products are amplified using the primer pairs PCR-L and PCR-R (*one tube–one SNP* approach) or PCR-L and PCR-R_*SNP-X* (*one tube–multiple SNPs* approach). White arrows indicate common sequences and primers; gray arrows indicate locus-specific sequences and primers; green and red arrows indicate TaqMan probes and their corresponding sequence stretches in the DOLs; arrow endings indicate 5′ ends of sequences; blunt endings indicate 3′ ends of sequences.

temperatures close to 55 and 60 °C, respectively. The two DOLs of each DO set differed at the 3′ end corresponding to the two alleles of the targeted SNP and carried different tails reflecting the respective alleles (tail-L1 or tail-L2, respectively). Depending on the applied allele-specific ligation procedure (see below), the DOR was either combined with a common tail (tail-R) or remained untailed. All tailed DOs were prepared by ligating the locus-specific parts with their corresponding common tails using the ligation-based synthesis method as described before [29]. DORs were phosphorylated by incubating at 37 °C for 1 h in 10 µl of 1× T4 PNK buffer (Tris–HCl, pH 7.6, 70 mM; $MgCl_2$, 10 mM; dithiothreitol, 5 mM) with 1 mM ATP and 2.5 U of T4 PNK (New England BioLabs, Beverly, MA, USA). Finally, PNK was inactivated at 65 °C for 20 min.

*Allele-specific ligation reaction.* The ligation reactions were carried out in 5 µl containing varying amounts of genomic DNA (0.5 to 100 ng), 1× Pfu ligase buffer (Tris–HCl, pH 7.5, 20 mM; KCl, 20 mM; $MgCl_2$, 10 mM;

Table 1
SNPs and primer sequences used for PCR amplifications and sequencings of target loci

| Gene name | AC | SNP RS | Abbreviation | Primer name | Sequence 5′→3′ |
|---|---|---|---|---|---|
| CACNA1H | AE006466 | 3751664 | c14 | CACNA1H-21Fd | GAATTCACGGAGGACGACGTCCG |
| | | | | CACNA1H-21Rd | TGCTGGAAGCCTCCTGAGAC |
| CACNA1H | AE006466 | 2235634 | c15 | CACNA1H-43Fd | AGGTTCTCTCTGCGGGTGGA |
| | | | | CACNA1H-43Rd | AGGGACCAGAGAAACGAGTG |
| SCN5A | M77235 | 1805126 | s56 | SCN5A28cF | GAGCCCAGCCGTGGGCATCCT |
| | | | | SCN5A28cR | GTCCCCACTCACCATGGGCAG |
| SCN5A | M77235 | 7429945 | s63 | SCN5A28fF | GGACCGTGAGTCCATCGTGTGA |
| | | | | SCN5A28fR | AGCCCATTCACAACATATACAGTCT |

SNPs are defined by their RS numbers (NCBI SNP Cluster IDs) and for each target sequence AC numbers (NCBI accession numbers) are listed.

Table 2
Primer sequences used for Amplifluor allele-specific PCR genotyping assays

| Name | Sequence 5′→3′ |
|---|---|
| ASP-1_c14 | [tail-1-JOE] - ACTTGCTGTCCACGATGC**G** |
| ASP-2_c14 | [tail-2-FAM] - GTACTTGCTGTCCACGATGC**A** |
| CRP_c14 | CCGCCTCTGGGTTACCT |
| ASP-1_c15 | [tail-1-JOE] - AAGCTCTCCCCCAGGTAGGT**A** |
| ASP-2_c15 | [tail-2-FAM] - GCTCTCCCCCAGGTAGGT**G** |
| CRP_c15 | GGGCCTGTGCTGAGGAT |
| ASP-1_s56 | [tail-1-JOE] - TTGGCTCAGACAGGGCG**T** |
| ASP-2_s56 | [tail-2-FAM] - AGTGGCTCAGACAGGGC**A** |
| CRP_s56 | CCAGAGGCCACTCAGTTTATT |
| ASP-1_s63 | [tail-1-JOE] - TGGGAGTAAGAAATGGGCCTC**G** |
| ASP-2_s63 | [tail-2-FAM] - GGGAGTAAGAAATGGGCCTC**A** |
| CRP_s63 | CCACAGAACTCTGCCTGGTT |
| tail-1-JOE | *GAAGGTCGGAGTCAACGGATT* |
| tail-2-FAM | *GAAGGTGACCAAGTTCATGCT* |
| Amplifuor-JOE | JOE-AGGACGCTGAGATGCGTCCT$^*$*GAAGGTCGGAGTCAACGGATT* |
| Amplifuor-FAM | FAM-AGCGATGCGTTCGAGCATCGCT$^*$*GAAGGTGACCAAGTTCATGCT* |

For each SNP assay five oligonucleotides are required: two tailed allele-specific primers (ASP), one common reverse primer (CRP), and two universal Amplifluor primers. Nucleotides corresponding to the allelic variances are indicated in boldface. Corresponding sequence stretches of the Amplifluor primers and the tails are written in italics. The quencher DABSYL is attached to the nucleotide T marked by $^*$ in the sequences of the Amplifluor primers.

Igepal, 0.1%; ATP, 0.001 mM; dithiothreitol, mM), 15% PEG 6000, 2 U of Pfu DNA ligase (Stratagen, La Jolla, CA, USA), and different amounts of DOs (5, 10, or 50 fmol of each DO). When applying the *one tube–one SNP* approach, the three appropriate DOs for one SNP were added. Both the two DOLs and the DOR were carrying tails. Following the *one tube–multiple SNPs* approach, ligation reactions were performed in the presence of DO sets for all SNPs of interest (Fig. 1A). For the multiplexing strategy, DO sets comprising tailed DOLs and tailed DOR were used; alternatively, DO sets including tailed DOLs and DOR without the common tail were used (Table 3). With the following temperature regime all ligation reactions were carried out in a DNA Engine Tetrad (MJ Research, Waltham, MA, USA): 95 °C, 2 min, [(65 °C, 30 s, 74 °C, 10 s) × 6 subcycles] × 20 cycles.

*TaqMan PCR.* Each 20-μl amplification reaction contained 1× TaqMan Universal PCR Master Mix (Applied Biosystems), 300 nM TaqMan probes labeled with fluorescein (TaqMan-FAM) and with tetrachlorofluorescein (TaqMan-TET) (Table 3), 400 nM amplification primers (Table 3), and ligation products. Products of allele-specific ligation reactions were subjected to TaqMan PCRs without purification. In the case of the *one tube–one SNP* approach, all 5 μl of the ligation reaction was added. Due to the use of tailed primers, common amplification primers (PCR-L and PCR-R) could be used. Having performed the multiplexed ligation reaction, the reaction mixture was filled up with $H_2O$ to a volume of 100 μl, from which 5 μl each was distributed to individual tubes for separate analyses of SNP loci (Fig. 1A). For amplification of ligation products originating from the *one tube–multiple SNPs* approach, one common primer (PCR-L) and one locus-specific primer (PCR-R_*SNP-X*) were applied. Amplifications were carried out on an ABI Prism 7900HT system (Applied Biosystems) using the following thermal profile: 50 °C, 2 min; 95 °C, 10 min; (95 °C, 15 s; 60 °C, 30 s) × 40 cycles.

Alternatively to the *one tube–multiple SNPs* approach described above, a three-step multiplexing procedure was tested as suggested before [19]. Tailed DORs were used for a multiplexed allele-specific ligation reaction. Prior to the separate locus-specific amplifications,

Table 3
Oligonucleotides used for LDR-TaqMan genotyping assays

| Name | Sequence 5′→3′ |
|---|---|
| DOL-1_c14 | [tail-L1] - GCGGCAAGCTGCGC**C** |
| DOL-2_c14 | [tail-L2] - AGCGGCAAGCTGCGC**T** |
| DOR_c14 | [tail-R] - GCATCGTGGACAGCAAGTACTTCAG |
| DOR-s_c14 | GCATCGTGGACAGCAAGTACTTCAG |
| DOL-1-a_c15 | [tail-L1] - AAGCTCTCCCCCAGGTAGGT**G** |
| DOL-2-a_c15 | [tail-L2] - CAAGCTCTCCCCCAGGTAGGT**A** |
| DOR_a_c15 | [tail-R] - GAGCCCGCGCCATCCTCAG |
| DOR-s_c15 | GAGCCCGCGCCATCCTCAG |
| DOL-1-b_c15 | [tail-L1] - CTCTCCCCCAGGTAGGT**G** |
| DOL-2-b_c15 | [tail-L2] - GCTCTCCCCCAGGTAGGT**A** |
| DOR_b_c15 | [tail-R] - GAGCCCGCGCCATCCT |
| DOL-1-a_s56 | [tail-L1] - TCCTGTCTGACTTTGCCGA**C** |
| DOL-2-a_s56 | [tail-L2] - TCCTGTCTGACTTTGCCGA**T** |
| DOR_a_s56 | [tail-R] - GCCCTGTCTGAGCCACTCCGTAT |
| DOR-s_s56 | GCCCTGTCTGAGCCACTCCGTAT |
| DOL-1-b_s56 | [tail-L1] - CCTGTCTGACTTTGCCGA**C** |
| DOL-2-b_s56 | [tail-L2] - CCTGTCTGACTTTGCCGA**T** |
| DOR_b_s56 | [tail-R] - GCCCTGTCTGAGCCACTCCG |
| DOL-1_s63 | [tail-L1] - TGGGAGTAAGAAATGGGCCTC**A** |
| DOL-2_s63 | [tail-L2] - GGGAGTAAGAAATGGGCCTC**G** |
| DOR_s63 | [tail-R] - GCCCCGCGGATCAACCAG |
| DOR-s_s63 | GCCCCGCGGATCAACCAG |
| tail-L1 | TGCACAATTCACGACTCACGAT*CCACACGGTCTCGCACTGGC*ACGGG |
| tail-L2 | TGCACAATTCACGACTCACGAT*CATCCGCTCCGACGACACGA*ACGGG |
| tail-R | CGGTCTAACGGGATAGCGTGGTGGTA |
| TaqMan-fam | FAM-*CCACACGGTCTCGCACTGGC*-TAMRA |
| TaqMan-tet | TET-*CATCCGCTCCGACGACACGA*-TAMRA |
| PCR-L | GCACAATTCACGACTCACGA |
| PCR-R | CCACCACGCTATCCCGTTAGAC |
| PCR-R_c14 | GCTGAAGTACTTGCTGTCCAC |
| PCR-R_c-15 | GCTGAGGATGGCGCGGGC |
| PCR-R_s56 | TACGGAGTGGCTCAGACAG |
| PCR-R_s63 | CTGGTTGATCCGCGGGGC |

Each set of detector oligonucleotides consists of two tailed "left" detector oligonucleotides (DOLs) and one "right" detector oligonucleotide (DOR). Depending on the assay approach (see Fig. 1) the DOR is either tailed or without the common tail. Nucleotides corresponding to the allelic variances are indicated in boldface. Common primers and their corresponding sequence stretches of the tails are underlined. TaqMan probes and their corresponding sequence stretches of the tails are written in italics.

a preamplification was performed in a thermocycler with common primers using the same PCR mixture as described for the *one tube–one SNP* approach. The preamplification comprised the following cycling protocol: 50 °C, 2 min; 95 °C, 10 min; (95 °C, 15 s; 60 °C, 30 s) × 10 cycles. After having diluted the preamplification mixture with 80 μl of $H_2O$, 1 μl each was subjected to TaqMan PCRs with locus-specific primers as described above for the two-step *one tube–multiple SNPs* approach.

For all assays fluorescence signals were measured in real time during TaqMan PCR amplification and by an endpoint detection using the ABI Prism 7900HT system.

## Results

### *Refinement of amplifluor allele-specific PCR*

A variety of setups for the Amplifluor assay was compared with respect to Amplifluor probe concentrations and thermal amplification profiles. Different annealing temperatures for PCR amplification ranging from 56 to 64 °C were all allele discriminatory and provided sufficient amplifications for all four SNPs. Best allele discriminations for all SNP loci were achieved with 20 plus 23 cycles for the amplification protocol comprising two cycling units and 40 cycles for the protocol comprising one cycling unit (see Materials and methods). Genotyping results obtained with the two alternative temperature regimes were practically the same (data not shown). For further analyses (see below), consensus conditions of the Amplifluor method comprising an amplification protocol with 20 plus 23 cycles and annealing temperatures of 58 and 56 °C were applied.

### *Adjustment of LDR-TaqMan method for human genomic DNA*

The LDR-TaqMan method was optimized for genotyping SNP on human genomic DNA using the *one* *tube–one SNP* procedure. For two SNPs, different DO sets concerning annealing temperatures of the locus-specific sequence stretches were tested (Table 3). Comparative analyses of both SNPs revealed that amplification plateaus during real-time measurements were reached about two cycles earlier and their final signals were slightly higher if ligations had been performed using longer DOs with annealing temperatures of approximately 60 °C for DOLs and 65 °C for DOR. For all test SNPs, allele-specific ligation reactions deploying various concentrations of DOs were performed, achieving consistently the highest endpoint signals in the presence of 10 fmol of each DO in 5-μl ligation reactions (data not shown).

*Sensitivity of genotyping assays*

In general, fluorescence signals from the Amplifluor assay were approximately 1.5 times as strong as signals from the LDR-TaqMan assay comparing results obtained on the same samples for the same SNPs (example shown in Fig. 4). Comparing allelic fluorescence signals from corresponding homozygous genotypes, FAM signals appeared consistently stronger than signals from the second fluorophore (JOE and TET, respectively), independent from the method used. The FAM–JOE ratio and the FAM–TET ratio were about 2.0 (Fig. 4).

For testing the sensitivity of the genotyping methods, different DNA dilutions (50, 25, 10, 5, 2.5, and 0.5 ng per reaction) of two known genotypes of SNP c14 (C/C and C/T) were genotyped. Allele-specific PCR using Amplifluor probes produced clear genotype calls on 0.5 ng per reaction. LDR-TaqMan reactions performed with 5 ng genomic DNA per reaction gave signals that could be strictly assigned to the genotypes. Signals obtained on reactions with less than 5 ng were not clearly above signals from the no-DNA controls, and thus no accurate genotyping was possible with these low concentrations (data not shown).

*Development of multiplexed LDR-TaqMan assay*

We compared two different strategies for multiplexing the allele-specific ligation of the LDR-TaqMan method (see Materials and methods). The three-step *one tube–multiple SNPs* procedure involved a preamplification reaction after the ligation and prior to the TaqMan PCR. Applying the two-step *one tube–multiple SNPs* procedure, the ligation products were directly subjected to locus-specific TaqMan amplifications; therefore no preparation of tailed DORs was required (Fig. 1). With both strategies, we obtained clear allele discriminations for all four SNPs on all genomic DNA samples of the F1-family (examples shown in Figs. 2 and 3). The two-step multiplexing strategy was comparatively carried out using tailed and untailed DORs for the ligation reactions, showing that efficiency and signal intensity is not influenced by presence of a tail (Figs. 2b and c). Thus *one tube–one SNP* and *one tube–multiple SNPs* can alternatively be performed using the same DO sets. Approximately the same signal intensity was achieved whether the TaqMan PCR was performed with 1 μl taken from preamplification mix (filled up to 100 μl after preamplification) in the course of the three-step procedure or with 5 μl directly taken from multiplexed ligation reaction (diluted to 100 μl after ligation) during the two-step procedure (Figs. 2a and c).

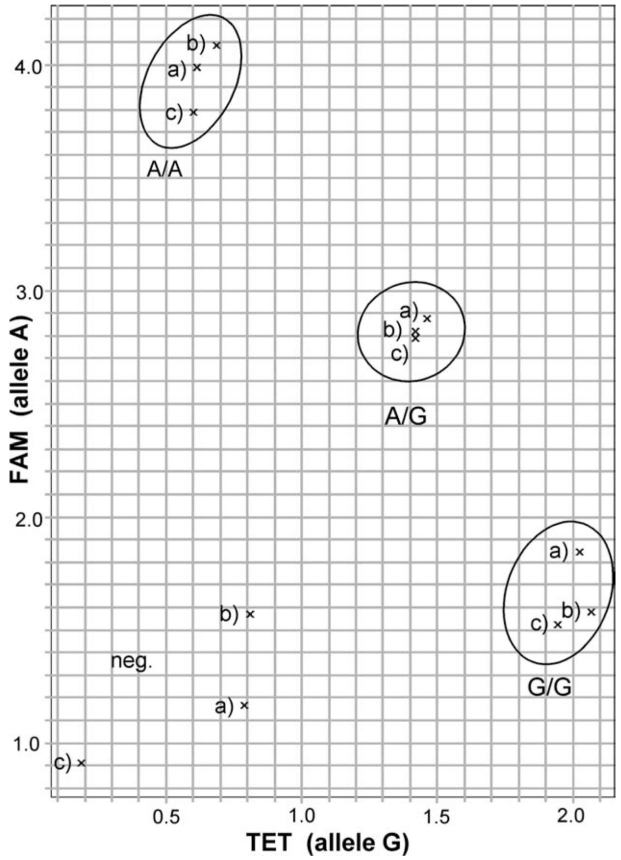


Fig. 2. Comparison of genotyping results using the LDR-TaqMan method with different setups for multiplexed ligation. Scatter plot shows endpoint-determined FAM and TET signals of SNP s63 obtained for three different genotypes plotted on the *x* and *y* axes, respectively. (a) Three-step *one tube–multiple SNPs* procedure: Ligation reaction was performed with a tailed DOR (DOR_s63). Subsequently, the ligation products were preamplified using common primers. Finally, the products were subjected to TaqMan PCR amplifications using the locus-specific reverse primer (PCR-R_s63). (b and c) Two-step *one tube–multiple SNPs* procedures: Ligation reactions were performed with either (b) a tailed DOR (DOR_s63) or (c) an untailed DOR (DOR-s_s63). Subsequently, ligation products were directly subjected to TaqMan PCR amplifications using the locus-specific reverse primer (PCR-R_s63). *neg.*, signals for the no-DNA controls.

*Genotyping results of family-based study*

Prior to testing the genotyping methods of interest, allelic states of the SNPs c14, c15, s56, and s63 were determined in the F1-family consisting of 18 members

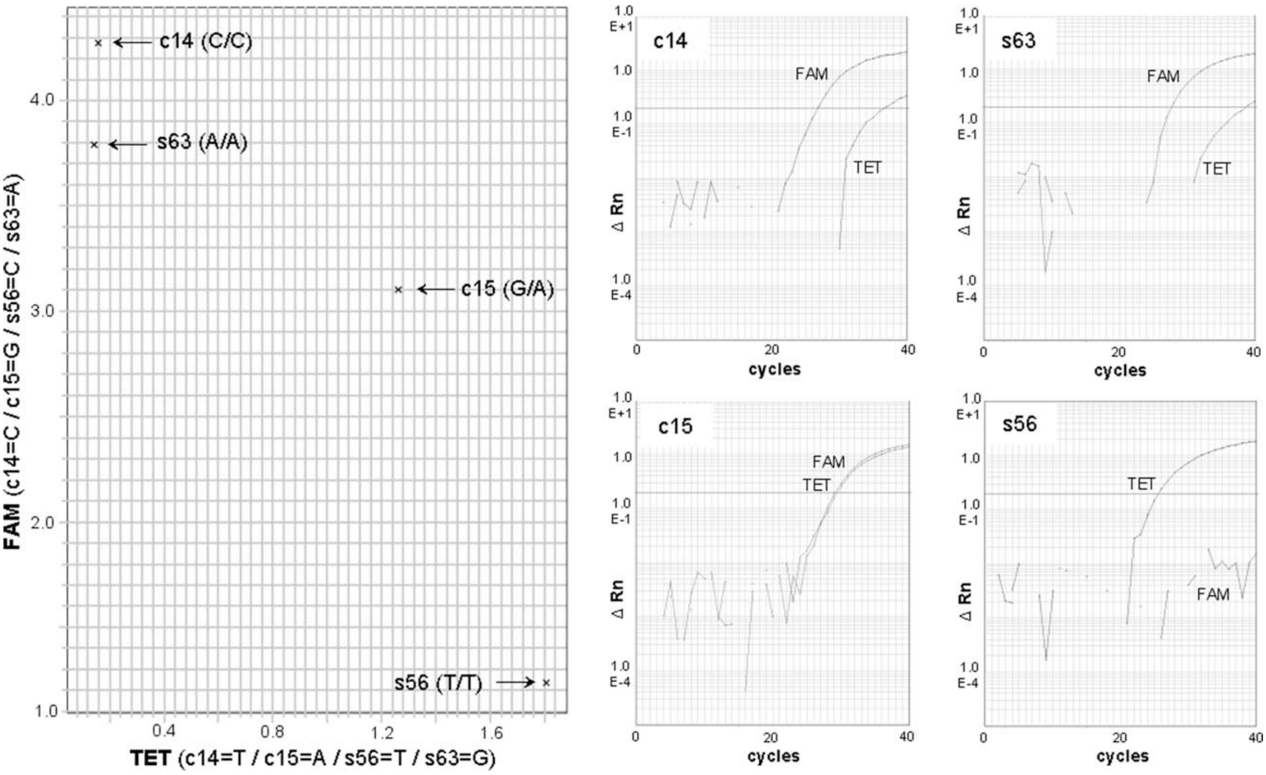


Fig. 3. Genotyping results of LDR-TaqMan method applying the two-step *one tube–multiple SNPs* approach without preamplification. Genotyping data are based on multiplexed ligation reactions of the SNPs c14, c15, s56, and s63 on the same aliquot of genomic DNA, followed by separate allele determinations for each SNP via TaqMan PCR. Scatter plots show endpoint-determined FAM and TET signals plotted on the *x* and *y* axes, respectively. Amplification plots show the log of the change in fluorescence plotted versus cycle number.

using conventional dye-terminator cycle sequencing. Knowing the precise genotypes, Amplifluor allele-specific PCR and both LDR-TaqMan strategies (*one tube–one SNP* and two-step *one tube–multiple SNPs*) were applied for genotyping these four SNPs in the complete F1-family. For each method, all genotyping reactions were performed under a single set of optimized conditions (see above). Individual amounts of DNA from the test panel were subjected to reactions varying between 15 and 100 ng. Data were evaluated via amplification plots of real-time analyses and scatter plots of endpoint measurements. For each of the 74 individual data points, there was complete correspondence between the genotyping results obtained by conventional sequencing, Amplifluor, and the two LDR-TaqMan procedures. All genotypes could be clearly assayed without the need for repetition of any single analysis. For all four SNPs, each individual allelic score fell into one of two or three defined clusters. The slight sample-to-sample variations represented differences in DNA amounts (Fig. 4). Scoring results showed 13 homozygous C/C and 5 heterozygous C/A genotypes for SNP c14; 5 homozygous G/G and 13 heterozygous C/A genotypes for SNP c15; 7 homozygous T/T, 2 homozygous C/C, and 9 heterozygous T/C genotypes for s56; and 7 homozygous A/A, 2 homozygous G/G, and 9 heterozygous A/G genotypes for s63 (Fig. 4). For both methods, the no-DNA controls clearly fell away from clusters of genomic DNA samples and clustered near the *xy* origin in scatter plots (example shown in Fig. 4).

## Discussion

We improved and comparatively applied two recently introduced promising methods for SNP genotyping directly on genomic DNA: Amplifluor allele-specific PCR [11–13] and LDR-TaqMan [19].

In addition to the differing initial steps of both methods compared—allele-specific PCR and allele-specific ligation—they share the same principle of using SNP-specific oligonucleotides with common tails, allowing discrimination of allele-specific products of all SNPs with just two universal fluorescence probes. Consequently, there is no need for expensive unique couples of fluorescence reporting reagents for each SNP, as is required for TaqMan [25], Molecular Beacons [30], and Invader [31]. This drastically reduces the development expenses for assays targeting new SNPs. Due to the use of fluorescence probes, endpoint measurement and final discrimination of allele-specific products can be conveniently performed on a standard fluorescence microplate reader. Therefore, establishment of these methods even outside of well-equipped research institutions seems feasible due to relatively low instrumentation costs. With

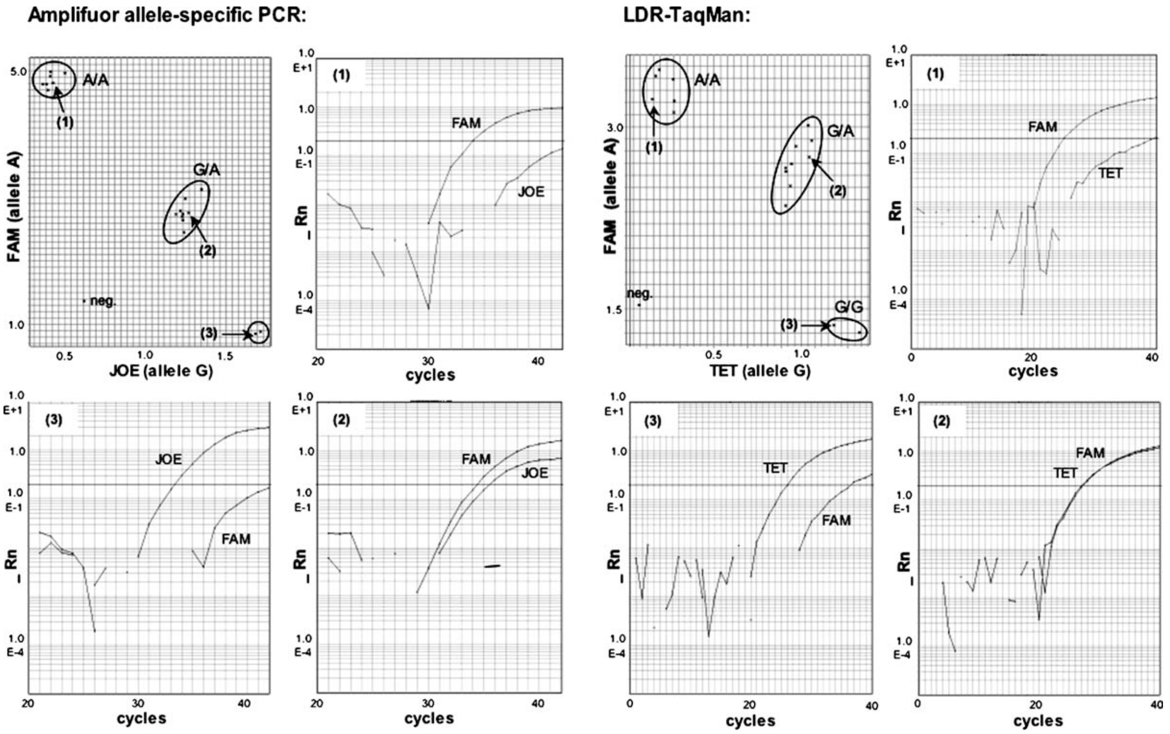


Fig. 4. Comparison of genotyping results obtained on different genomic DNA samples by Amplifluor allele-specific PCR (left) and *one tube–one SNP* LDR-TaqMan (right). For both genotyping methods scatter plots including data of all F1-family members and amplification plots for three genotypes representing the three different allelic states of the SNP s63 are presented. Genotyping reactions were performed on the same panel of genomic DNA showing varying DNA concentrations between 15 and 100 ng. Scatter plots show endpoint-determined FAM and JOE or FAM and TET signals plotted on the *x* and *y* axes, respectively. *neg.*, signals for the no-DNA controls. Amplification plots show the log of the change in fluorescence versus cycle number. Due to the use of two consecutive cycling units during the Amplifluor allele-specific PCR, the amplification plots start with cycle number 20.

regard to running expenses, final genotyping costs per SNP and sample were in the same low range for both methods, making them equally attractive for applications from small to large scales.

Since LDR-TaqMan had previously been tested only on the *A. thaliana* genome, we adjusted the method for the approximately 20-fold more complex human genome. It appeared beneficial to design detector oligonucleotides with 5 °C higher melting temperatures with regard to the locus-specific sequence stretches and to double the concentration of DOs to 2 nM each for the ligation reaction. These adjustments led to higher signal intensities, indicating higher ligation efficiency.

Previously, a three-step multiplexing procedure for LDR-TaqMan including multiplexed ligation, multiplexed preamplification, and separate TaqMan PCR amplifications had been employed [19]. We developed a less laborious *one tube–multiple SNPs* approach based on multiplexed allele-specific ligation directly followed by separate TaqMan PCRs (Fig. 1B). Both strategies provided the same accuracy and allele-discriminating quality (Fig. 2). Since preamplification using two common primers was dropped, preparation of tailed “right” DOs was not necessary for the two-step multiplexing strategy (Fig. 1B). Due to the missing preamplification, different numbers of simultaneously targeted SNPs starting from the same ligation volume appear achievable, with our data suggesting up to 20 markers using the two-step approach and up to 100 markers for the three-step procedure based on 5-μl ligation reaction. In general, multiplexing strategies are useful if the amount of genomic DNA available becomes a limiting factor as many SNPs are analyzed on the same aliquot of DNA. Furthermore, multiplexing further reduces assay costs since fewer ligation reactions using expensive ligase enzyme are performed. However, a drawback of the *one tube–multiple SNPs* ligation reaction is that, for all SNPs, locus-specific primers are required for TaqMan PCRs (Fig 1B), whereas with the *one tube–one SNP* procedure, analyses of all allele-specific ligation products can be conveniently performed with a common pair of primers (Fig 1A).

Testing the limits of sensitivity, precise allele discriminations could be performed on 5 ng of genomic DNA using the LDR-TaqMan method. The Amplifluor method demonstrated a higher sensitivity with reliable genotyping results obtained on 0.5 ng of genomic DNA, being in agreement with previous studies [13]. This difference in sensitivity might be explained by the underlying principles for allele-specific interaction with the alleles of a SNP, since—unlike the allele-specific PCR—the ligation-based interrogation of alleles does not comprise amplification itself. This assumption is also supported by overall stronger fluorescence signals obtained with the Amplifluor method than with LDR-TaqMan. The generally stronger signal intensity from FAM-labeled probes had no influence on the accuracy of allele discriminations, as it has also been described before [11].

For further analysis of accuracy and robustness, four not-preselected SNPs were assayed in a panel of 18 related human DNA samples. In general, both methods did not require intensive optimization efforts and worked for a wide range of reaction parameters; thus consensus conditions appropriate for all SNPs could be easily determined. Although all SNPs of this study could be correctly genotyped with both methodologies, it has been reported that allele-specific PCR—even the more discriminatory approach of competitively scoring both alleles in the same reaction [12]—is more error prone due to sequence-specific limitations than allele-specific ligation [7]. In addition to the "vertical" robustness observed in this study, i.e., the robust application for different SNPs, both methods demonstrated a reliable "horizontal" robustness leading to accurate allelic discriminations on all DNA samples without the need for any replication. We deliberately performed our analyses on a test panel with varying DNA concentrations (15 to 100 ng), since in the course of large-scale projects DNA samples are usually not adjusted to equal concentrations after DNA extraction. Showing these aspects of robustness, both the LDR-TaqMan and the Amplifluor methods fulfill essential demands for SNP genotyping methods meant for broad applications.

A significant advantage of the Amplifluor method is its single-step and closed-tube format. Since no post-reaction products are released into the laboratory, the potential for contamination is reduced to a minimum. The simple single-step reaction procedure makes the Amplifluor assay more suitable for high-throughput approaches using automation. In comparison, the LDR-TaqMan method is slightly more laborious as it requires an additional pipetting step. However, the two-step procedure should not restrict its potential for automated up-scaling strategies.

In conclusion, based on our evaluations and optimizations, we suggest that both methods, Amplifluor allele-specific PCR and LDR-TaqMan, are accurate, robust, easy to establish, and low-cost SNP genotyping techniques, suitable for research and for clinical diagnostic settings.

## Acknowledgments

This work was funded by BMBF Grant 0313040A-C. The authors are grateful to Ilona Dunkel for technical assistance and Aleksey Soldatov for helpful discussions. We thank Maike Tribbels, Lesie Hazelwood, and Christina Grimm for critical reading of the manuscript.

## References

[1] F.S. Collins, M.S. Guyer, A. Chakravarti, Variations on a theme: cataloging human DNA sequence variation, Science 278 (1997) 1580–1581.

[2] W.-H. Li, L.A. Sadler, Low nucleotide diversity in man, Genetics 129 (1991) 513–523.

[3] The International SNP Map Working Group, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928–933.

[4] M.C. Ellis, 'Spot-On' SNP genotyping, Genome Res. 10 (7) (2000) 895–897.

[5] I.G. Gut, Automation in genotyping of single nucleotide polymorphisms, Hum. Mut. 17 (2001) 475–492.

[6] P.Y. Kwok, Approaches to allele frequency determination, Pharmacogenomics 1 (2) (2000) 231–235.

[7] P.Y. Kwok, Methods for genotyping single nucleotide polymorphisms, Annu. Rev. Genomics Hum. Genet. 2 (2001) 235–258.

[8] M.M. Shi, Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies, Clin. Chem. 47 (2001) 164–172.

[9] A.C. Syvänen, Accessing genetic variation: genotyping single nucleotide polymorphisms, Nat. Rev. Genet. 2 (12) (2001) 940–942.

[10] R.M. Twyman, S.B. Primrose, Techniques patents for SNP genotyping, Pharmacogenomics 4 (1) (2003) 67–79.

[11] C. Bengra, T.E. Mifflin, Y. Khripin, P. Manunta, S.M. Williams, P.A. Jose, R.A. Felder, Genotyping of essential hypertension single-nucleotide polymorphisms by a homogeneous PCR method with universal energy transfer primers, Clin. Chem. 48 (12) (2002) 2131–2140.

[12] J.R. Hawkins, Y. Khripin, A.M. Valdes, T.A. Weaver, Miniaturized sealed-tube allele-specific PCR, Hum. Mut. 19 (2002) 543–553.

[13] M.V. Myakishev, Y. Khripin, S. Hu, D.H. Hamer, High-throughput SNP genotyping by allele-specific pcr with universal energy-transfer-labeled primers, Genome Res. 11 (2001) 163–169.

[14] C.D.K. Bottema, G. Sarkar, J.D. Cassay, S. Ii, C.M. Dutton, S.S. Sommer, PCR-amplification of specific alleles: a general method of rapidly detecting mutations, polymorphisms, and haplotypes, Methods Enzymol. 218 (1993) 388–402.

[15] C.R. Newton, A. Graham, L.E. Heptinstall, S.J. Powell, C. Summers, N. Kalsheker, J.C. Smith, A.F. Markham, Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS), Nucleic Acids Res. 17 (1989) 2503–2516.

[16] H. Okayama, D.T. Curiel, M.L. Brantly, M.D. Holmes, R.G. Crystal, Rapid, nonradioactive detection of muations in the human genome by allele-specific amplification, J. Lab. Clin. Med. 114 (1989) 105–113.

[17] I.A. Nazarenko, S.K. Bhatnagar, R.J. Hohman, A closed tube format for amplification and detection of DNA based on energy transfer, Nucleic Acids Res. 25 (1997) 2516–2521.

[18] G.J. Nuovo, R.J. Hohman, G.A. Nardone, I.A. Nazarenko, In situ amplification using universal energy transfer-labeled primers, J. Histochem. Cytochem. 47 (1999) 273–280.

[19] T.A. Borodina, H. Lehrach, A.V. Soldatov, LDR-TaqMan procedure for SNP detection on genomic DNA, in submission.

[20] F. Barany, Genetic disease detection and DNA amplification using cloned thermostable ligase, Proc. Natl. Acad. Sci. USA 88 (1991) 189–193.

[21] U. Landegren, R. Kaiser, J. Sanders, L. Hood, A ligase-mediated gene detection technique, Science 241 (1988) 1077–1080.

[22] D.A. Nickerson, R. Kaiser, S. Lappin, J. Stewart, L. Hood, U. Landegren, Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay, Proc. Natl. Acad. Sci. USA 87 (1990) 8923–8927.

[23] M. Samiotaki, M. Kwiatkowski, M. Parik, U. Landegren, Dual-color detection of DNA sequence variants by ligase-mediated analysis, Genomics 20 (1994) 238–242.

[24] P.M. Holland, R.D. Abramson, R. Watson, D.H. Gelfand, Detection of specific polymerase chain reaction product by utilizing the 5′-3′ exonuclease activity of Thermus aquaticus DNA polymerase, Proc. Natl. Acad. Sci. USA 88 (1991) 7276–7280.

[25] K.J. Livak, S.J. Flood, J. Marmaro, W. Giusti, K. Deetz, Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization, PCR Methods Appl. 4 (1995) 357–362.

[26] S.W.M. John, G. Weitzner, R. Rozen, C.R. Scriver, A rapid procedure for extracting genomic DNA from leukocytes, Nucleic Acids Res. 19 (2) (1991) 408.

[27] A.M. Rickert, A. Premstaller, C. Gebhardt, P.J. Oefner, Genotyping of SNPs in a polyploid genome by pyrosequencing, BioTechniques 32 (3) (2002) 592–593.

[28] Serologicals Corporations. Amplifluor SNPs Genotyping Systems. Manual.

[29] T.A. Borodina, H. Lehrach, A.V. Soldatov, Ligation-based synthesis of oligonucleotides with block structure, Anal. Biochem. 318 (2) (2003) 309–313.

[30] S. Tyagi, D.P. Bratu, F.R. Kramer, Multicolor molecular beacons for allele discrimination, Nat. Biotechnol. 16 (1998) 49–53.

[31] V. Lyamichev, B. Neri, Invader assay for SNP genotyping, Methods Mol. Biol. 212 (2003) 229–240.

# 3 Summary

The described approach to discover transcriptional networks for cardiac development, function and disease is based in the integration of clinical, phenotypic and molecular data at a global scale; a systems biology approach. To enable the integration of different relevant levels, a panel of molecular studies and computational developments have been performed, such as:

- Collection of biomaterial and detailed clinical phenotypes of congenital heart disease
- Set-up of a CardioVascular Genetics database and d-matrix as front-end/analysis tool
- Genome-wide gene expression profiling of normal and malformed human hearts
- Analysis of cardiac transcription factor binding in vivo and its functional consequences
- Characterization of epigenetic marks, such as histone modification at a global scale
- Studying the influence of the genomic organization on transcription
- Optimization of transcription factor binding prediction

Finally, transcription networks based on the integration of obtained molecular and clinical data had been predicted.

In time, methods for quantitative real-time PCR and the detection of single nucleotide polymorphism have been developed and optimized for sparse material, according to the needs of the used clinical samples.

In the genome-wide studies, three transcription factors raised particular interest and have been studied in more detail. Thus CITED2 and TBX20 could be associated with congenital heart disease in human for the first time; mutations in CITED2 were shown to be potentially disease causative. DPF3 could be discovered as a novel epigenetic transcription factor of particular importance for cardiac and skeletal muscle development and function. Moreover, functional studies showed that DPF3 represents the first plant-homeodomains known to bind histone acetylation marks, which displays a novel protein-domain function.

# 4 References

Akazawa H, Komuro I. Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases. *Pharmacol Ther* 2005;**107**:252-268.

Balza RO, Jr., Misra RP. Role of the serum response factor in regulating contractile apparatus gene expression and sarcomeric integrity in cardiomyocytes. *J Biol Chem* 2006;**281**:6498-6510.

Bamforth SD, Braganca J, Eloranta JJ, Murdoch JN, Marques FI, Kranc KR, Farza H, Henderson DJ, Hurst HC, Bhattacharya S. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. *Nat Genet* 2001;**29**:469-474.

Bamforth SD, Braganca J, Farthing CR, Schneider JE, Broadbent C, Michell AC, Clarke K, Neubauer S, Norris D, Brown NA, Anderson RH, Bhattacharya S. Cited2 controls left-right patterning and heart development through a Nodal-Pitx2c pathway. *Nat Genet* 2004;**36**:1189-1196.

Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proc Natl Acad Sci U S A* 1991;**88**:189-193.

Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J, Renault B, Kucherlapati R, Seidman JG, Seidman CE. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet* 1997;**15**:30-35.

Bauer EP, Kuki S, Zimmermann R, Schaper W. Upregulated and downregulated transcription of myocardial genes after pulmonary artery banding in pigs. *Ann Thorac Surg* 1998;**66**:527-531.

Baumgarten G, Knuefermann P, Kalra D, Gao F, Taffet GE, Michael L, Blackshear PJ, Carballo E, Sivasubramanian N, Mann DL. Load-dependent and -independent regulation of proinflammatory cytokine and cytokine receptor gene expression in the adult mammalian heart. *Circulation* 2002;**105**:2192-2197.

Benson DW, Silberbach GM, Kavanaugh-McHugh A, Cottrill C, Zhang Y, Riggs S, Smalls O, Johnson MC, Watson MS, Seidman JG, Seidman CE, Plowden J, Kugler JD. Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways. *J Clin Invest* 1999;**104**:1567-1573.

Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell* 2007;**128**:669-681.

Bhattacharya S, Macdonald ST, Farthing CR. Molecular mechanisms controlling the coupled development of myocardium and coronary vasculature. *Clin Sci (Lond)* 2006;**111**:35-46.

Bhattacharya S, Michels CL, Leung MK, Arany ZP, Kung AL, Livingston DM. Functional role of p35srj, a novel p300/CBP binding protein, during transactivation by HIF-1. *Genes Dev* 1999;**13**:64-75.

Bienz M. The PHD finger, a nuclear protein-interaction domain. *Trends Biochem Sci* 2006;**31**:35-40.

Borodina TA, Lehrach H, Soldatov AV. Ligation detection reaction-TaqMan procedure for single nucleotide polymorphism detection on genomic DNA. *Anal Biochem* 2004;**333**:309-319.

Braganca J, Eloranta JJ, Bamforth SD, Ibbitt JC, Hurst HC, Bhattacharya S. Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2. *J Biol Chem* 2003;**278**:16021-16029.

Brewer S, Jiang X, Donaldson S, Williams T, Sucov HM. Requirement for AP-2alpha in cardiac outflow tract morphogenesis. *Mech Dev* 2002;**110**:139-149.

Brown DD, Martz SN, Binder O, Goetz SC, Price BM, Smith JC, Conlon FL. Tbx5 and Tbx20 act synergistically to control vertebrate heart morphogenesis. *Development* 2005;**132**:553-563.

Bruneau BG, Nemer G, Schmitt JP, Charron F, Robitaille L, Caron S, Conner DA, Gessler M, Nemer M, Seidman CE, Seidman JG. A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* 2001;**106**:709-721.

Cai CL, Zhou W, Yang L, Bu L, Qyang Y, Zhang X, Li X, Rosenfeld MG, Chen J, Evans S. T-box genes coordinate regional rates of proliferation and regional specification during cardiogenesis. *Development* 2005;**132**:2475-2487.

Cao D, Wang Z, Zhang CL, Oh J, Xing W, Li S, Richardson JA, Wang DZ, Olson EN. Modulation of smooth muscle gene expression by association of histone acetyltransferases and deacetylases with myocardin. *Mol Cell Biol* 2005;**25**:364-376.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. The transcriptional landscape of the mammalian genome. *Science* 2005;**309**:1559-1563.

Chambers DM, Peters J, Abbott CM. The lethal mutation of the mouse wasted (wst) is a deletion that abolishes expression of a tissue-specific isoform of translation elongation factor 1alpha, encoded by the Eef1a2 gene. *Proc Natl Acad Sci U S A* 1998;**95**:4463-4468.

Charron F, Tsimiklis G, Arcand M, Robitaille L, Liang Q, Molkentin JD, Meloche S, Nemer M. Tissue-specific GATA factors are transcriptional effectors of the small GTPase RhoA. *Genes Dev* 2001;**15**:2702-2719.

Chazaud C, Oulad-Abdelghani M, Bouillet P, Decimo D, Chambon P, Dolle P. AP-2.2, a novel gene related to AP-2, is expressed in the forebrain, limbs and face during mouse embryogenesis. *Mech Dev* 1996;**54**:83-94.

Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, Thomas NR, Scotter AJ, Caves LS, Tyrrell GP, Newbury-Ecob RA, Munnich A, Bonnet D, Brook JD. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nat Genet* 2005;**37**:423-428.

Clark KL, Yutzey KE, Benson DW. Transcription factors and congenital heart defects. *Annu Rev Physiol* 2006;**68**:97-121.

Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997;**278**:1580-1581.

Copley RR, Totrov M, Linnell J, Field S, Ragoussis J, Udalova IA. Functional conservation of Rel binding sites in drosophilid genomes. *Genome Res* 2007;**17**:1327-1335.

Cosgrove MS, Wolberger C. How does the histone code work? *Biochem. Cell Biol.* 2005;**83**:468-476.

Cross SJ, Ching YH, Li QY, Armstrong-Buisseret L, Spranger S, Lyonnet S, Bonnet D, Penttinen M, Jonveaux P, Leheup B, Mortier G, Van Ravenswaaij C, Gardiner CA. The mutation spectrum in Holt-Oram syndrome. *J Med Genet* 2000;**37**:785-787.

Debril MB, Gelman L, Fayard E, Annicotte JS, Rocchi S, Auwerx J. Transcription factors and nuclear receptors interact with the SWI/SNF complex through the BAF60c subunit. *J Biol Chem* 2004;**279**:16677-16686.

Dodou E, Verzi MP, Anderson JP, Xu SM, Black BL. Mef2c is a direct transcriptional target of ISL1 and GATA factors in the anterior heart field during mouse embryonic development. *Development* 2004;**131**:3931-3942.

Eckert D, Buhl S, Weber S, Jager R, Schorle H. The AP-2 family of transcription factors. *Genome Biol* 2005;**6**:246.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;**95**:14863-14868.

Ellinghaus P, Scheubel RJ, Dobrev D, Ravens U, Holtz J, Huetter J, Nielsch U, Morawietz H. Comparing the global mRNA expression profile of human atrial and ventricular myocardium with high-density oligonucleotide arrays. *J Thorac Cardiovasc Surg* 2005;**129**:1383-1390.

Falkman G. Information visualisation in clinical Odontology: multidimensional analysis and interactive data exploration. *Artif Intell Med* 2001;**22**:133-158.

Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A* 2001;**98**:10781-10786.

Fischer JJ, Krueger T, Schueler M, Schlesinger J, Lange M, Toenjes M, Sperling S. The cardiac transcription network driven by Gata4, Mef2a, Nkx2.5 and Srf and epigenetic marks. *in preparation* 2008a.

Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, Sperling S. Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics* 2008b;**91**:41-51.

Flajollet S, Lefebvre B, Cudejko C, Staels B, Lefebvre P. The core component of the mammalian SWI/SNF complex SMARCD3/BAF60c is a coactivator for the nuclear retinoic acid receptor. *Mol Cell Endocrinol* 2007;**270**:23-32.

Freedman SJ, Sun ZY, Kung AL, France DS, Wagner G, Eck MJ. Structural basis for negative regulation of hypoxia-inducible factor-1alpha by CITED2. *Nat Struct Biol* 2003;**10**:504-512.

Gabig TG, Mantel PL, Rosli R, Crean CD. Requiem: a novel zinc finger gene essential for apoptosis in myeloid cells. *J. Biol. Chem.* 1994;**269**:29515-29519.

Garcia JM, Gonzalez R, Silva JM, Dominguez G, Vegazo IS, Gamallo C, Provencio M, Espana P, Bonilla F. Mutational status of K-ras and TP53 genes in primary sarcomas of the heart. *Br J Cancer* 2000;**82**:1183-1185.

Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, Rothrock CR, Eapen RS, Hirayama-Yamada K, Joo K, Matsuoka R, Cohen JC, Srivastava D. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* 2003;**424**:443-447.

Gaussin V, Van de Putte T, Mishina Y, Hanks MC, Zwijsen A, Huylebroeck D, Behringer RR, Schneider MD. Endocardial cushion and myocardial defects after cardiac myocyte-specific conditional deletion of the bone morphogenetic protein receptor ALK3. *Proc Natl Acad Sci U S A* 2002;**99**:2878-2883.

Gerland U, Moroz JD, Hwa T. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A* 2002;**99**:12015-12020.

Ghosh TK, Packham EA, Bonser AJ, Robinson TE, Cross SJ, Brook JD. Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. *Hum Mol Genet* 2001;**10**:1983-1994.

Giulietti A, Overbergh L, Valckx D, Decallonne B, Bouillon R, Mathieu C. An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. *Methods* 2001;**25**:386-401.

Goff LA, Davila J, Jornsten R, Keles S, Hart RP. Bioinformatic analysis of neural stem cell differentiation. *J Biomol Tech* 2007;**18**:205-212.

Guo W, Li H, Aimond F, Johns DC, Rhodes KJ, Trimmer JS, Nerbonne JM. Role of heteromultimers in the generation of myocardial transient outward K+ currents. *Circ Res* 2002;**90**:586-593.

Gut IG. Automation in genotyping of single nucleotide polymorphisms. *Hum Mutat* 2001;**17**:475-492.

Hammer S, Toenjes M, Lange M, Fischer JJ, Dunkel I, Mebus S, Grimm CH, Hetzer R, Berger F, Sperling S. Characterization of TBX20 in human hearts and its regulation by TFAP2. *J Cell Biochem* 2008;**104**:1022-1033.

Harvey RP. Patterning the vertebrate heart. *Nat Rev Genet* 2002;**3**:544-556.

Hassan AH, Awad S, Al-Natour Z, Othman S, Mustafa F, Rizvi TA. Selective recognition of acetylated histones by bromodomains in transcriptional co-activators. *Biochem J* 2007;**402**:125-133.

Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M, Carrozza MJ, Workman JL. Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* 2002;**111**:369-379.

Herrgard MJ, Covert MW, Palsson BO. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res* 2003;**13**:2423-2434.

Hershberg R, Yeger-Lotem E, Margalit H. Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet* 2005;**21**:138-142.

Hiroi Y, Kudoh S, Monzen K, Ikeda Y, Yazaki Y, Nagai R, Komuro I. Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nat Genet* 2001;**28**:276-280.

Hoffman JI. Incidence of congenital heart disease: I. Postnatal incidence. *Pediatr Cardiol* 1995a;**16**:103-113.

Hoffman JI. Incidence of congenital heart disease: II. Prenatal incidence. *Pediatr Cardiol* 1995b;**16**:155-165.

Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A* 1991;**88**:7276-7280.

Iio A, Koide M, Hidaka K, Morisaki T. Expression pattern of novel chick T-box gene, Tbx20. *Dev Genes Evol* 2001;**211**:559-562.

Jenuwein T, Allis CD. Translating the histone code. *Science* 2001;**293**:1074-1080.

Karamboulas C, Dakubo GD, Liu J, De Repentigny Y, Yutzey K, Wallace VA, Kothary R, Skerjanc IS. Disruption of MEF2 activity in cardiomyoblasts inhibits cardiomyogenesis. *J Cell Sci* 2006;**119**:4315-4321.

Karibe A, Tobacman LS, Strand J, Butters C, Back N, Bachinski LL, Arai AE, Ortiz A, Roberts R, Homsher E, Fananapazir L. Hypertrophic cardiomyopathy caused by a novel alpha-tropomyosin mutation (V95A) is associated with mild cardiac phenotype, abnormal calcium binding to troponin, abnormal myosin cycling, and poor prognosis. *Circulation* 2001;**103**:65-71.

Kasahara H, Izumo S. Identification of the in vivo casein kinase II phosphorylation site within the homeodomain of the cardiac tisue-specifying homeobox gene product Csx/Nkx2.5. *Mol Cell Biol* 1999;**19**:526-536.

Kawamura T, Ono K, Morimoto T, Wada H, Hirai M, Hidaka K, Morisaki T, Heike T, Nakahata T, Kita T, Hasegawa K. Acetylation of GATA-4 is involved in the differentiation of embryonic stem cells into cardiac myocytes. *J Biol Chem* 2005;**280**:19682-19688.

Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, Alexi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S. Genome-wide array analysis of normal and malformed human hearts. *Circulation* 2003;**107**:2467-2474.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 2003;**31**:3576-3579.

Kiehl TR, Shibata H, Vo T, Huynh DP, Pulst SM. Identification and expression of a mouse ortholog of A2BP1. *Mamm Genome* 2001;**12**:595-601.

Kim SY, Kim Y. Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data. *BMC Bioinformatics* 2006;**7**:330.

Kirby ML, Cheng G, Stadt H, Hunter G. Differential expression of the L10 ribosomal protein during heart development. *Biochem Biophys Res Commun* 1995;**212**:461-465.

Kirk EP, Sunde M, Costa MW, Rankin SA, Wolstein O, Castro ML, Butler TL, Hyun C, Guo G, Otway R, Mackay JP, Waddell LB, Cole AD, Hayward C, Keogh A, Macdonald P, Griffiths L, Fatkin D, Sholler GF, Zorn AM, Feneley MP, Winlaw DS, Harvey RP. Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. *Am J Hum Genet* 2007;**81**:280-291.

Kouzarides T. Chromatin Modifications and Their Function. *Cell* 2007;**128**:693-705.

Kraus F, Haenig B, Kispert A. Cloning and expression analysis of the mouse T-box gene tbx20. *Mech Dev* 2001;**100**:87-91.

Kuo HC, Cheng CF, Clark RB, Lin JJ, Lin JL, Hoshijima M, Nguyen-Tran VT, Gu Y, Ikeda Y, Chu PH, Ross J, Giles WR, Chien KR. A defect in the Kv channel-interacting protein 2 (KChIP2) gene leads to a complete loss of I(to) and confers susceptibility to ventricular tachycardia. *Cell* 2001;**107**:801-813.

Kurdistani SK, Grunstein M. Histone acetylation and deacetylation in yeast. *Nat. Rev. Mol. Cell. Biol.* 2003;**4**:276-284.

Kwok PY. Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* 2001;**2**:235-258.

Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. *Science* 1988;**241**:1077-1080.

Lange M, Kaynak B, Forster UB, Tonjes M, Fischer JJ, Grimm C, Schlesinger J, Just S, Dunkel I, Krueger T, Mebus S, Lehrach H, Lurz R, Gobom J, Rottbauer W, Abdelilah-Seyfried S, Sperling S. Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. *Genes Dev* 2008;**22**:2370-2384.

Lessard J, Wu JI, Ranish JA, Wan M, Winslow MM, Staahl BT, Wu H, Aebersold R, Graef IA, Crabtree GR. An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron* 2007;**55**:201-215.

Leung MK, Jones T, Michels CL, Livingston DM, Bhattacharya S. Molecular cloning and chromosomal localization of the human CITED2 gene encoding p35srj/Mrg1. *Genomics* 1999;**61**:307-313.

Li W, Cornell RA. Redundant activities of Tfap2a and Tfap2c are required for neural crest induction and development of other non-neural ectoderm derivatives in zebrafish embryos. *Dev Biol* 2007;**304**:338-354.

Li WH, Sadler LA. Low nucleotide diversity in man. *Genetics* 1991;**129**:513-523.

Li ZY, Yang J, Gao X, Lu JY, Zhang Y, Wang K, Cheng MB, Wu NH, Wu Z, Shen YF. Sequential recruitment of PCAF and BRG1 contributes to myogenin activation in 12-O-tetradecanoylphorbol-13-acetate-induced early differentiation of rhabdomyosarcoma-derived cells. *J Biol Chem* 2007;**282**:18872-18878.

Lickert H, Takeuchi JK, von Both I, Walls JR, McAuliffe F, Lee Adamson S, Mark Henkelman R, Wrana JL, Rossant J, Bruneau BG. Baf60c is essential for function of BAF chromatin remodelling complexes in heart development. *Nature* 2004;**432**:107-112.

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 1995;**4**:357-362.

Lyons I, Parsons LM, Hartley L, Li R, Andrews JE, Robb L, Harvey RP. Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes Dev* 1995;**9**:1654-1666.

Margueron R, Trojer P, Reinberg D. The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* 2005;**15**:163-176.

McFadden DG, Charite J, Richardson JA, Srivastava D, Firulli AB, Olson EN. A GATA-dependent right ventricular enhancer controls dHAND transcription in the developing heart. *Development* 2000;**127**:5331-5341.

Mertsalov IB, Kulikova DA, Alimova-Kost MV, Ninkina NN, Korochkin LI, Buchman VL. Structure and expression of two members of the d4 gene family in mouse. *Mamm Genome* 2000;**11**:72-74.

Molkentin JD, Li L, Olson EN. Phosphorylation of the MADS-Box transcription factor MEF2C enhances its DNA binding activity. *J Biol Chem* 1996;**271**:17199-17204.

Montgomery RL, Davis CA, Potthoff MJ, Haberland M, Fielitz J, Qi X, Hill JA, Richardson JA, Olson EN. Histone deacetylases 1 and 2 redundantly regulate cardiac morphogenesis, growth, and contractility. *Genes Dev* 2007;**21**:1790-1802.

Morrison AC, Bray MS, Folsom AR, Boerwinkle E. ADD1 460W allele associated with cardiovascular disease in hypertensive individuals. *Hypertension* 2002;**39**:1053-1057.

Moser M, Ruschoff J, Buettner R. Comparative analysis of AP-2 alpha and AP-2 beta gene expression during murine embryogenesis. *Dev Dyn* 1997;**208**:115-124.

Moser MJ, Marshall DJ, Grenier JK, Kieffer CD, Killeen AA, Ptacin JL, Richmond CS, Roesch EB, Scherrer CW, Sherrill CB, Van Hout CV, Zanton SJ, Prudent JR. Exploiting the enzymatic recognition of an unnatural base pair to develop a universal genetic analysis system. *Clin Chem* 2003;**49**:407-414.

Natalia NN, Ilja BM, Dina AK, Maria VA-K, Olga BS, Leonid IK, Sergey LK, Vladimir LB. Cerd4, third member of the d4 gene family: expression and organization of genomic locus. *Mammalian Genome* 2001;**V12**:862.

Naya FJ, Black BL, Wu H, Bassel-Duby R, Richardson JA, Hill JA, Olson EN. Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. *Nat Med* 2002;**8**:1303-1309.

Nazarenko I, Lowe B, Darfler M, Ikonomi P, Schuster D, Rashtchian A. Multiplex quantitative PCR using self-quenched primers labeled with a single fluorophore. *Nucleic Acids Res* 2002;**30**:e37.

Nediani C, Formigli L, Perna AM, Ibba-Manneschi L, Zecchi-Orlandini S, Fiorillo C, Ponziani V, Cecchi C, Liguori P, Fratini G, Nassi P. Early changes induced in the left ventricle by pressure overload. An experimental study on swine heart. *J Mol Cell Cardiol* 2000;**32**:131-142.

Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P. Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics* 2005;**6**:68.

Nemer G, Fadlalah F, Usta J, Nemer M, Dbaibo G, Obeid M, Bitar F. A novel mutation in the GATA4 gene in patients with Tetralogy of Fallot. *Hum Mutat* 2006;**27**:293-294.

Nickerson DA, Kaiser R, Lappin S, Stewart J, Hood L, Landegren U. Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc Natl Acad Sci U S A* 1990;**87**:8923-8927.

Nishimura K, Nakatsu F, Kashiwagi K, Ohno H, Saito T, Igarashi K. Essential role of S-adenosylmethionine decarboxylase in mouse embryonic development. *Genes Cells* 2002;**7**:41-47.

Olson EN. A decade of discoveries in cardiac biology. *Nat Med* 2004;**10**:467-474.

Ornatsky OI, Cox DM, Tangirala P, Andreucci JJ, Quinn ZA, Wrana JL, Prywes R, Yu YT, McDermott JC. Post-translational control of the MEF2A transcriptional regulatory protein. *Nucleic Acids Res* 1999;**27**:2646-2654.

Palacios D, Puri PL. The epigenetic network regulating muscle development and regeneration. *J Cell Physiol* 2006;**207**:1-11.

Pallavicini A, Kojic S, Bean C, Vainzof M, Salamon M, Ievolella C, Bortoletto G, Pacchioni B, Zatz M, Lanfranchi G, Faulkner G, Valle G. Characterization of human skeletal muscle Ankrd2. *Biochem Biophys Res Commun* 2001;**285**:378-386.

Plageman TF, Jr., Yutzey KE. Differential expression and function of Tbx5 and Tbx20 in cardiac development. *J Biol Chem* 2004;**279**:19026-19034.

Potthoff MJ, Arnold MA, McAnally J, Richardson JA, Bassel-Duby R, Olson EN. Regulation of skeletal muscle sarcomere integrity and postnatal muscle function by Mef2c. *Mol Cell Biol* 2007.

Rahmann S, Muller T, Vingron M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol* 2003;**2**:Article7.

Rauch C, Loughna PT. Static stretch promotes MEF2A nuclear translocation and expression of neonatal myosin heavy chain in C2C12 myocytes in a calcineurin- and p38-dependent manner. *Am J Physiol Cell Physiol* 2005;**288**:C593-605.

Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 2007;**23**:134-141.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;**409**:928-933.

Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR, Pierpont ME, Gelb BD. Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. *Nat Genet* 2000;**25**:42-46.

Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, Maron BJ, Seidman CE, Seidman JG. Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* 1998;**281**:108-111.

Schratt G, Philippar U, Berger J, Schwarz H, Heidenreich O, Nordheim A. Serum response factor is crucial for actin cytoskeletal organization and focal adhesion assembly in embryonic stem cells. *J Cell Biol* 2002;**156**:737-750.

Searcy RD, Vincent EB, Liberatore CM, Yutzey KE. A GATA-dependent nkx-2.5 regulatory element activates early cardiac gene expression in transgenic mice. *Development* 1998;**125**:4461-4470.

Seelow D, Galli R, Mebus S, Sperling HP, Lehrach H, Sperling S. d-matrix - database exploration, visualization and analysis. *BMC Bioinformatics* 2004;**5**:168.

Shelton EL, Yutzey KE. Tbx20 regulation of endocardial cushion cell proliferation and extracellular matrix gene expression. *Dev Biol* 2007;**302**:376-388.

Sif S. ATP-dependent nucleosome remodeling complexes: enzymes tailored to deal with chromatin. *J Cell Biochem* 2004;**91**:1087-1098.

Simone C. SWI/SNF: the crossroads where extracellular signaling pathways meet chromatin. *J Cell Physiol* 2006;**207**:309-314.

Simone C, Forcales SV, Hill DA, Imbalzano AN, Latella L, Puri PL. p38 pathway targets SWI-SNF chromatin-remodeling complex to muscle-specific loci. *Nat Genet* 2004;**36**:738-743.

Singh MK, Christoffels VM, Dias JM, Trowe MO, Petry M, Schuster-Gossler K, Burger A, Ericson J, Kispert A. Tbx20 is essential for cardiac chamber differentiation and repression of Tbx2. *Development* 2005;**132**:2697-2707.

Sirotkin H, O'Donnell H, DasGupta R, Halford S, St Jore B, Puech A, Parimoo S, Morrow B, Skoultchi A, Weissman SM, Scambler P, Kucherlapati R. Identification of a new human catenin gene family member (ARVCF) from the region deleted in velo-cardio-facial syndrome. *Genomics* 1997;**41**:75-83.

Skerjanc IS, Petropoulos H, Ridgeway AG, Wilton S. Myocyte enhancer factor 2C and Nkx2-5 up-regulate each other's expression and initiate cardiomyogenesis in P19 cells. *J Biol Chem* 1998;**273**:34904-34910.

Small EM, Krieg PA. Transgenic analysis of the atrialnatriuretic factor (ANF) promoter: Nkx2-5 and GATA-4 binding sites are required for atrial specific expression of ANF. *Developmental Biology* 2003;**261**:116-131.

Solloway MJ, Harvey RP. Molecular pathways in myocardial development: a stem cell perspective. *Cardiovasc Res* 2003;**58**:264-277.

Spencer JA, Misra RP. Expression of the serum response factor gene is regulated by serum response factor binding sites. *J Biol Chem* 1996;**271**:16535-16543.

Sperling S. Transcriptional regulation at a glance. *BMC Bioinformatics* 2007;**8**:S2.

Sperling S, Grimm CH, Dunkel I, Mebus S, Sperling HP, Ebner A, Galli R, Lehrach H, Fusch C, Berger F, Hammer S. Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Hum Mutat* 2005;**26**:575-582.

Srivastava D. Genetic assembly of the heart: implications for congenital heart disease. *Annu Rev Physiol* 2001;**63**:451-469.

Stennard FA, Costa MW, Elliott DA, Rankin S, Haast SJ, Lai D, McDonald LP, Niederreither K, Dolle P, Bruneau BG, Zorn AM, Harvey RP. Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. *Dev Biol* 2003;**262**:206-224.

Stennard FA, Costa MW, Lai D, Biben C, Furtado MB, Solloway MJ, McCulley DJ, Leimena C, Preis JI, Dunwoodie SL, Elliott DE, Prall OW, Black BL, Fatkin D, Harvey RP. Murine T-box transcription factor Tbx20 acts as a repressor during heart development, and is essential for adult heart integrity, function and adaptation. *Development* 2005;**132**:2451-2462.

Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;**403**:41-45.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004;**101**:6062-6067.

Sun X, Lee J, Navas T, Baldwin DT, Stewart TA, Dixit VM. RIP3, a novel apoptosis-inducing kinase. *J Biol Chem* 1999;**274**:16871-16875.

Svanvik N, Stahlberg A, Sehlstedt U, Sjoback R, Kubista M. Detection of PCR products in real time using light-up probes. *Anal Biochem* 2000;**287**:179-182.

Szeto DP, Griffin KJ, Kimelman D. HrT is required for cardiovascular development in zebrafish. *Development* 2002;**129**:5093-5101.

Tabibiazar R, Wagner RA, Liao A, Quertermous T. Transcriptional profiling of the heart reveals chamber-specific gene expression patterns. *Circ Res* 2003;**93**:1193-1201.

Takeuchi JK, Mileikovskaia M, Koshiba-Takeuchi K, Heidt AB, Mori AD, Arruda EP, Gertsenstein M, Georges R, Davidson L, Mo R, Hui CC, Henkelman RM, Nemer M, Black BL, Nagy A, Bruneau BG. Tbx20 dose-dependently regulates transcription factor networks required for mouse heart and motoneuron development. *Development* 2005;**132**:2463-2474.

Takeuchi JK, Ohgi M, Koshiba-Takeuchi K, Shiratori H, Sakaki I, Ogura K, Saijoh Y, Ogura T. Tbx5 specifies the left/right ventricles and ventricular septum position during cardiogenesis. *Development* 2003;**130**:5953-5964.

Tanaka M, Chen Z, Bartunkova S, Yamasaki N, Izumo S. The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development. *Development* 1999;**126**:1269-1280.

Taylor MS. Characterization and comparative analysis of the EGLN gene family. *Gene* 2001;**275**:125-132.

Toenjes M, Schueler M, Hammer S, Pape UJ, Fischer JJ, Berger F, Vingron M, Sperling S. Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes. *Mol Biosyst* 2008;**4**:589-598.

Turner BM. Cellular Memory and the Histone Code. *Cell* 2002;**111**:285-291.

Tyagi S, Bratu DP, Kramer FR. Multicolor molecular beacons for allele discrimination. *Nat Biotechnol* 1998;**16**:49-53.

Volcik KA, Zhu H, Finnell RH, Shaw GM, Canfield M, Lammer EJ. Evaluation of the Cited2 gene and risk for spina bifida and congenital heart defects. *Am J Med Genet A* 2004;**126**:324-325.

Walker AJ, Cross SS, Harrison RF. Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet* 1999;**354**:1518-1521.

Wang YX, Qian LX, Yu Z, Jiang Q, Dong YX, Liu XF, Yang XY, Zhong TP, Song HY. Requirements of myocyte-specific enhancer factor 2A in zebrafish cardiac contractility. *FEBS Lett* 2005;**579**:4843-4850.

Ware SM, Peng J, Zhu L, Fernbach S, Colicos S, Casey B, Towbin J, Belmont JW. Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. *Am J Hum Genet* 2004;**74**:93-105.

Wegman EJ. Visual data mining. *Stat Med* 2003;**22**:1383-1397.

Weninger WJ, Floro KL, Bennett MB, Withington SL, Preis JI, Barbera JP, Mohun TJ, Dunwoodie SL. Cited2 is required both for heart morphogenesis and establishment of the left-right axis in mouse development. *Development* 2005;**132**:1337-1348.

Whitcombe D, Theaker J, Guy SP, Brown T, Little S. Detection of PCR products using self-probing amplicons and fluorescence. *Nat Biotechnol* 1999;**17**:804-807.

Yang J, Moravec CS, Sussman MA, DiPaola NR, Fu D, Hawthorn L, Mitchell CA, Young JB, Francis GS, McCarthy PM, Bond M. Decreased SLIM1 expression and increased gelsolin expression in failing human hearts measured by high-density oligonucleotide arrays. *Circulation* 2000;**102**:3046-3052.

Yin Z, Haynie J, Yang X, Han B, Kiatchoosakun S, Restivo J, Yuan S, Prabhakar NR, Herrup K, Conlon RA, Hoit BD, Watanabe M, Yang YC. The essential role of Cited2, a negative regulator for HIF-1alpha, in heart development and neurulation. *Proc Natl Acad Sci U S A* 2002;**99**:10488-10493.

Zhang CL, McKinsey TA, Chang S, Antos CL, Hill JA, Olson EN. Class II histone deacetylases act as signal-responsive repressors of cardiac hypertrophy. *Cell* 2002;**110**:479-488.

Zhao Y, Samal E, Srivastava D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* 2005;**436**:214-220.

# 6 Acknowledgment

I would like to thank all those who supported my research; my collaborators, colleagues and funding agents.

Especially, I am grateful to Prof. Dr. H-D. Schmidt, Prof. Dr. D. Ledbetter, Prof. Dr. P.E. Lange and Prof. Dr. H. Lehrach to drive my interest in science, for fruitful discussions and excellent research facilities during my doctoral, postdoctoral and group leader ship.

I would like to express my gratitude to all past and present members of the Sperling Lab. It was and is my great pleasure to work with you; without your spirit and intelligence, this work would have never been possible.

At the Max Planck Institute for Molecular Genetics, I would like to thank particularly Prof. Dr. M. Vingron, who supported my interest and work in computational biology.

I am deeply grateful to the patients with congenital heart disease and their families and the long-standing cooperation with the Deutsches Herzzentrum Berlin (Prof F. Berger and Prof. R. Hetzer).

Finally, I am thankful to Prof. Dr. B. Wittig for supporting this work.

# Erklärung

§ 4 Abs. 3 (k) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde.

- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden.

- mir die geltende Habilitationsordnung bekannt ist.

| | |
|---|---|
| ............................ | ..................................... |
| Datum | Unterschrift |
| | Dr. Silke Rickert-Sperling |