Functional and phylogenetic analyses of chromosome 21 promoters and hominid-specific transcription factor binding sites

DISSERTATION

zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von

Dipl.-Biol. Robert Querfurth
aus Hamburg

Oktober 2009

1. Gutachter: Prof. Dr. R. Mutzel,
Institut für Biologie, Freie Universität Berlin,
Königin-Luise-Str. 12-16, D-14195 Berlin

2. Gutachter: Prof. Dr. H. Lehrach,
Max-Planck-Institut für molekulare Genetik,
Ihnestr.73, D-14195 Berlin

Disputation am 09.12.2009

TABLE OF CONTENTS

1. Summary

The focus of this work addresses functional studies of human and primate promoters, and the genome-wide localization and validation of human-specific transcription factor binding sites of the essential transcription factor GABPa. In this context, the development of an improved PCR protocol, including the careful adjustment of PCR additives to compose an efficient enhancer mix, was central to the amplification of large GC-rich promoter fragments used as source for the functional studies. Based on this, part of the work assessed the potential of promoter-reporter constructs to drive transcription in human HEK cells, in order to capture regulatory regions corresponding to a large fraction of the human chromosome 21 genes. The results obtained in this study demonstrated the usefulness of transient transfection assays. The high correlations of reporter activities with endogenous expression levels of the corresponding genes, and with the presence of DNA sequence elements important for transcription initiation, indicate that transient reporter gene assays are capable of depicting endogenous transcription regulation for individual promoters in living cells. This finding was further underlined by the results obtained after either truncation and/or external stimulation of promoters, showing that especially distal promoter regions of reporter constructs are capable of integrating endogenous response signaling pathways into reporter activity. Thus, we applied this technology in a comparative genomics approach specially designed for identifying and testing human-specific transcription factor binding sites (TFBSs). To find TFBSs specific to human and hominids, a new approach was implemented combining leading tools in sequence analysis and comparative genomics. The established pipeline was applied to analyze ChIP-seq data capturing endogenous binding sites of the human transcription factor GABPa in HEK293 cells. Among the genes with human-specific binding sites, several functionally related groups were found, which can be linked without difficulties to human-specific traits. Functional testing showed consistent impacts of orthologous promoters of human, chimpanzee and rhesus macaque on the transcriptional outputs. Mutational analyses of candidate sites strongly supported these findings. In particular, the TMBIM6 (transmembrane BAX inhibitor motif containing 6) promoter, harboring several uncharacterized human-specific mutations and a hominid-specific GABPa binding site, represents an interesting candidate for follow-up studies, as TMBIM6 is involved in oxidative stress reduction and has been implicated in diabetes, atherosclerosis and in many of the aging-related neurodegenerative diseases, such as Alzheimer's and Parkinson's. This work presents the first successful implementation of a genome-wide approach to the identification of newly evolved *cis*-regulatory elements showing a specific function in human cells lines in comparison to our closest living relatives, the chimpanzees.

2. Zusammenfassung

Der Fokus dieser Arbeit liegt in der funktionellen Charakterisierung von Menschen- und Primatenpromotoren, einschließlich der genomweiten Lokalisierung und Validierung von humanspezifischen Transkriptionsfaktor-Bindestellen (TFBS) des essentiellen Transkriptionsfaktors GABPa. In diesem Kontext war die Etablierung eines verbesserten PCR Protokolls, einschließlich der Entwicklung eines *PCR enhancers*, zur Amplifikation langer und GC-reicher Promotoren ein zentraler Bestandteil. Darauf aufbauend befasst sich ein Teil dieser Arbeit mit der Analyse eines Großteils der Promotoren des humanen Chromosoms 21 in Hinblick auf ihr Potential, Transkription in HEK293-Zellen anzutreiben, und regulatorische Regionen zu charakterisieren. Die beobachtete hohe Korrelation von Reportergenaktivität und endogener Expression, wie auch die Korrelation mit DNS-Sequenzelementen von wichtiger Funktion während der Transkriptionsinitiation, zeigen, daß transiente Reportergenassays dazu geeignet sind, endogene Generegulation an individuellen Promotoren wiederzuspiegeln. Diese Aussage wird unterstützt sowohl durch Versuche mit verkürzten Promotoren wie auch durch externe Stimulation der Reporterkonstrukte, mit dem Ergebnis, daß vor allem distale Promoterregionen in der Lage sind, endogen ablaufende Signalkaskaden in Reporteraktivität zu integrieren. In dieser Hinsicht wurde die Technik in einem Ansatz komparativer Genomanalyse angewandt, um human-spezifische TFBS funktionell zu testen. Zur Identifikation human- und hominiden-spezifischer TFBS wurde ein neuer Ansatz implementiert, der führende Programme und Algorithmen aus den Bereichen der Sequenzanalyse und komparativen Genomanalyse vereint. Diese Implementation wurde auf ChIP-seq-Daten von endogenen Bindestellen des humanen Transkriptionsfaktors GABPa angewandt. Unter den Genen mit human-spezifischen Bindestellen finden sich einige funktionell verwandte Gruppen von Genen, die ohne Schwierigkeiten mit human-spezifischen Eigenschaften in Verbindung gebracht werden können. Die funktionelle Analyse von Kandidatenbindestellen zeigte in konsistenter Weise den unterschiedlichen Einfluß von orthologen Promotoren aus Mensch, Schimpanse und Rhesusaffe auf die Reporteraktivitäten. Mutationsanalysen mit ausgewählten Bindestellen bekräftigten diese Ergebnisse. Insbesondere repräsentiert der TMBIM6-Promoter (transmembrane BAX inhibitor motif containing 6), der neben mehreren uncharakterisierten human-spezifischen Mutationen eine hominiden-spezifische GABPa-Bindestelle enthält, einen interessanten Kandidaten für Folgestudien, denn TMBIM6 ist beteiligt an der Reduktion von oxidativem Stress und ebenso an Diabetes, Arthereosklerose und verschiedenen altersbedingten neuro-degenerativen Erkrankungen, wie Alzheimer und Parkinson. Diese Arbeit stellt die erste erfolgreiche Implementation eines genomweiten Ansatzes zur Identifizierung von jüngst evolvierten *cis*-regulatorischen Elementen dar, die einen messbaren Einfluß in einer humanen Zelllinie haben, auch im Vergleich zu unseren nächsten Verwandten, den Schimpansen.

3. Introduction

3.1. Transcriptional regulation

Starting from a single cell, multi-cellular organisms develop into complex systems composed of various different cell types all equipped with the same set of genes. Yet each cell employs only a part of the genes at any given moment. During development and life, the proportion and composition of expressed genes changes considerably among cell types and in response to physiological and environmental conditions [1-5]. Eukaryotic genomes contain on the order of $0.5–5x10^4$ genes. To allow precise spacio-temporal gene expression, a particularly complex system of regulatory mechanisms is necessary. Today, a number of contributing mechanisms are known, including chromatin condensation, histone modification, DNA methylation, transcription initiation, transcription elongation, alternative splicing, mRNA stability, translational control, different forms of posttranslational modifications, intracellular trafficking, and protein degradation [6, 7].

In eukaryotic cells, two meters of DNA fit into the nucleus of about 5 µm in diameter. This can only be achieved by higher-order packaging of DNA. Such packed DNA is referred to as chromatin, and in the most compacted form, chromatin is visible in the form of the chromosomes. Prior to transcription, chromatin needs to de-condense so that proteins necessary for transcription gain access to the DNA. This step, even though tightly controlled, represents a general switch that turns genes on or off, rather than regulating the levels of gene activity [8]. Of all the mechanisms mentioned above, for most genes, transcription initiation was thought to be the principal determinant of gene expression levels [9-12]. Meanwhile, also the regulation of transcription elongation has turned out to be of central importance for a large fraction of genes [13, 14]. Hence, the principal determinants of gene expression levels are involved in two mechanisms, transcription initiation and elongation [15].

A key element to both regulatory mechanisms is the promoter, the region surrounding the transcriptional start site (TSS), where proteins make contact with specific DNA sequence elements to regulate transcription. These proteins are known as transcription factors, while their binding sequences are known as transcription factor binding sites (TFBSs). The term promoter describes a structural organization of several TFBSs that, when bound by transcription factors, synergistically regulate transcription. There are no clear definitions on promoter size and extension, as in different promoters, also TFBSs are distributed very differently [8]. In general, promoters are subdivided into core, proximal and distal promoter regions. This categorization is linked to the presence of different types of TFBSs as well as their relative densities.

In metazoans, the core promoter spans 50-100 bps surrounding the TSS [16] and harbors binding sites for proteins of general importance to transcription. The proximal promoter spans approximately 250 bps upstream to the TSS and harbors high densities of gene-specific binding sites [17], while distal promoters include gene-specific binding sites that reside further upstream.

Maybe the only structures of mammalian promoters that allow a categorization of genes according to their promoter structures are CG-rich regions of 200 and more base pairs with an average CG content of >50%. Such regions are referred as CpG-islands (CGIs), occur in approximately 72% of all human promoters [18, 19] and allow the classification of genes into CGI and non-CGI associated genes [16]. CGI promoters contain several TSSs dispersed over 50 to 100 nucleotides [16]. They are associated with both ubiquitously expressed 'housekeeping' genes, and with genes showing complex expression patterns, particularly those expressed during embryonic development [13, 14, 20, 21]. On the other hand, non-CGI genes are highly tissue-specific; they have focused transcriptional start sites and seem to be inactive by default [16, 21].

Apart from a categorization of genes based on CGIs, regulated genes have been classified into primary and secondary response genes. Primary response genes can be quickly activated, while secondary response genes require new protein synthesis and chromatin remodeling at their promoters [22]. Interestingly, primary response genes are generally associated with CGIs [23], indicating that both types of characteristics capture similar sets of genes. Even though primary and secondary response genes are differently regulated, the initial steps of transcription initiation are thought to function in similar ways.

In metazoans, transcription initiation is a complex mechanism involving different levels of regulation. The first level that was discovered involves formation of the preinitiation complex (PIC) composed of general transcription factors (GTFs) and the RNA synthesizing enzyme RNA polymerase II (RNAPII) [24, 25]. This large complex of interacting proteins is regarded as the general transcription machinery (GTM), transcribes all protein-coding genes, and assembles at the core promoter. Until recently, transcriptional control of most genes was thought to be achieved by regulating the recruitment of RNA polymerase II (RNAPII) to the promoters [8, 26, 27].

The first hint for another type of regulation, despite the recruitment of RNAPII, came in 1986, when Gilmour and Lis found that RNAPII interacts with the promoter of *Drosophila* hsp70 gene (heat shock protein 70), even though the gene had not been induced by heat shock [28]. One year later Wu and Wilson identified a protein that binds, upon heat shock, upstream of the transcriptional start site (TSS) of the hsp70 gene and induces transcription. They speculated

that: "*Once induced, the direct binding of activator to the heat shock promoter poised for transcription by the presence of constitutively bound TATA factor and RNA polymerase II could be sufficient for activation of the transcriptional apparatus*" [29]. Only recently, this picture has emerged as being representative, since most active genes show marks of stalled or poised RNAPII at the core promoter awaiting activation [13, 30]. Indeed, this mechanism of RNAPII stalling has been shown important for genes that are quickly induced upon endogenous or external signals [31]. Another recent finding was the connection of stalled RNAPII and CGIs in human fibroblasts, were RNAPII stalling near the TSS occurs in approximately 30% of active genes, of which 89% are associated with CGIs [32].

Despite quick gene induction, RNAPII stalling is also seen in genes that can be quickly shut down and might be used for the dual purpose of repressing gene expression and preparing genes for rapid induction [33]. Since ~70% of mammalian promoters contain CpG islands, and CGI genes are pivotal in 'housekeeping' and development, RNAPII stalling represents the most common quick regulatory mechanism affecting at least one third of all genes, but probably many more corresponding to the large percentage of CGI genes.

In addition, CGIs are subject to another mechanism involving addition of methyl groups to the 5-position of cytosine. CGI methylation is thought to stabilize chromatin structure, and thereby inhibits accessibility of the transcription machinery to the promoters [34]. Furthermore, CGIs are preassociated with ubiquitous transcription factors like Sp1, and have been shown to cause instable nucleosome assembly *in vitro*, two factors beneficial to active chromatin [23]. Taking together, CGI-gene promoters are in general accessible to the PIC and other regulatory factors, while the vast majority of active CGI-genes also show RNAPII stalling, allowing for rapid gene induction. Therefore, at least for CGI genes representing more than 2/3 of all genes, the rate-limiting step in transcription frequently occurs after RNAPII associates to the promoter and involves gene-specific transcription factors that tune basal transcription of the GTM [35]. The following introduction into transcription initiation and elongation focuses only on the most common factors involved, and is summarized as an overview figure on page 10.

3.1.1. Transcription initiation

Prior to transcription initiation, chromatin remodeling complexes are necessary to relax DNA-histone associations of critical regulatory regions including TSS and the promoter [36]. Transcription factors are capable of recruiting histone-modifying enzymes like histone-acetylases and histone-methylases (see Figure 1A) [37, 38]. In this way, induction of a single DNA-binding transcription factor can induce several downstream genes.

In 1969, Roeder and Rutter identified three distinct RNA polymerases: I, II and III [39]. One year later, they found that polymerase I was primarily involved in 18S and 28S ribosomal RNA transcription, while polymerase III transcribed 5S rRNA and tRNAs. Only polymerase II (RNAPII) is responsible for transcribing protein coding genes and some non-translated mRNAs, such as microRNAs [40, 41]. However, RNA polymerase II requires several accessory factors for the initiation of site-specific transcription.

These essential factors are termed general transcription factors (GTFs). Usually, the term “transcription factor” refers to DNA-binding proteins recognizing specific DNA motifs. GTFs, however, are protein complexes composed of many proteins with different functions, including DNA binding, co-activation, phosphorylation, histone actyltransferase activity, ATPase activity, helicase activity, DNA repair, glycosylation, ubiquitination and proteins involved in recruitment of GTFs, elongation and termination [35]. There are six GTFs that interact with RNAPII, namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. This huge complex of interacting proteins including RNAPII is referred to as the pre-initiation complex (PIC) and assembles at the core promoter (see Figure 1B)[35].

The PIC recognizes and binds to certain DNA sequence elements residing in the core promoter. The DNA motifs occurring most frequently include basic recognition element (BRE), TATA box, initiator element (INR) and downstream promoter element (DPE). However, bioinformatics analyses revealed that less than 22% of the human genes contain a TATA box, and among these TATA-containing promoters, 62% have an Inr, 24% include a DPE, and 12% hold a BRE. The same study also indicates that among the 78% TATA-deficient promoters, 45% possess an Inr, 25% have a DPE, and 28% harbor a BRE [42]. There is increasing knowledge on the function of core promoter elements. The TATA box is recognized by TFIID, which contains the TATA box-binding protein (TBP) and triggers PIC formation (see Figure 1B) [43]. The BRE is recognized by TFIIB and helps to orient the PIC [44]. The Inr element is capable of directing accurate transcription initiation, while TFIID has been implicated in Inr recognition [44] and is the primary GTF recognizing the DPE element [45]. However, this picture of transcription initiation, induced by PIC formation at core promoter elements, is likely

to change as many core promoters do not harbor any of the mentioned motifs [46]. It might be replaced by a picture that is to a lesser extent dominated by core promoter elements, while the control of transcription elongation is drifting into focus.

3.1.2. Transcription elongation

Following transcription initiation and PIC formation, TFIIH phosphorylates serine-5 of the C-terminal domain (CTD) of RNAPII [35]. This phosphorylation induces the recruitment of the 5' capping enzyme and the dissociation of RNAPII from the PIC (see Figure 1C) [35]. This step is often referred as promoter clearance, as the early elongation complex breaks contact with core-promoter elements to initiate transcription [47].

For many genes, the early elongation complex synthesizes only short fragments of approximately 40 bps before it pauses downstream to the TSS [48]. Meanwhile, 'pausing' is referred to as stalling and represents, besides PIC formation, the second major mechanism of transcriptional control.

Two factors have been found responsible for RNAPII stalling, including DRB-sensitive inducing factor (DSIF) and negative elongation factor (NELF) (see Figure 1E) [15, 20, 49]. The phosphorylation of DSIF, NELF and serin-2 of the CTD of RNAPII is crucial to productive elongation. A single complex called positive transcription elongation factor (P-TEFb) realizes these phosphorylations (see Figure 1F) [15, 20, 49]. Therefore, the recruitment of P-TEFb represents the limiting step in activation of stalled RNAPII [49]. Several specific activators recruit P-TEFb, but also general chromatin remodeling proteins like Brd4 [50]. Specific activators include DNA-binding proteins and co-activators that interact with DNA-binding proteins [49].

3.1.3. Gene-specific transcription control

Regulation of transcription rates is believed to be influenced largely by gene-specific transcription factors (from now on referred to as TFs) that interact directly or indirectly with the PIC [31, 35]. TFs are also known to recruit P-TEFb, yet to an extent that remains to be investigated (see Figure 1E) [49].

Direct interaction with the PIC has been described for several TFs [51], but the presence of TFs that do not bind to DNA has triggered the identification of different classes of TFs termed mediator complexes. Some of these protein complexes interact with restricted sets of TFs [52, 53] while others interact with a variety of unrelated TFs [54, 55], implying a general mechanism for the transition of TF signals to the PIC [56, 57].

The complexity of TF-mediated gene regulation is underlined by the fact that more than 3,000 different TFs are encoded in the human genome, hence approximately 10% of all human genes are directly involved in gene regulation [58]. The key feature of DNA-binding TFs is their ability to bind to specific genomic regions ranging from 4-16 base pairs termed transcription factor binding sites (TFBSs). Deciphering this part of transcriptional regulation poses particular issues, since most TFs not only recognize one specific sequence, but many related sequences. Therefore, TFBSs are regarded as degenerate, as they can often tolerate one or more nucleotide substitutions without losing functionality [10, 59]. The distribution and density of TFBSs varies enormously between genes, but in general, they tend to cluster in the proximal promoter regions [17], and also in more distal regions forming enhancers, silencers or insulators.

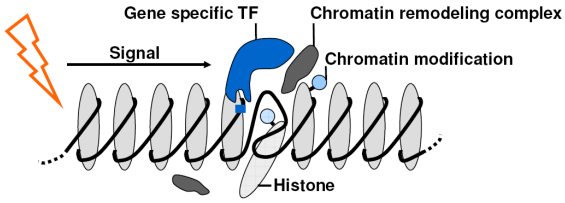
Promoters are key to transcriptional regulation as they integrate many cellular signals, delivered by TFs, into transcription levels. For example, during early development they integrate spacio-temporal signals to produce highly dynamic patterns of transcription in specific regions of the embryo [1, 60-62]. Promoters of 'housekeeping' genes that are constitutively active can shut down e.g. in response to stress conditions, such as starvation or heat shock [63]. On the other hand, promoters that are “off” by default can be activated in response to hormonal, physiological or environmental signals [8].

Therefore, the array of active TFs within the nucleus in conjunction with their target sites (TFBSs) determines which genes are expressed at what level and under which circumstances. Furthermore, the function of TFBSs is always context-dependent to some extent. For example, sites are only functional if the binding TF is present and active, the chromatin is not condensed and the TFBS is not masked by another TF occupying an adjacent TFBS. In addition, many TFs interact with cofactors or other TFs that need to be present as well.

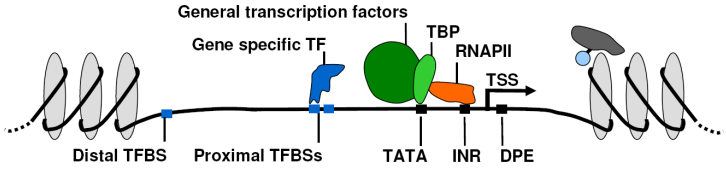
Due to the sequence degeneracy and strong context-dependence of transcriptional regulation, sequence inspection alone provides limited information about promoter function. Understanding the functional consequences of sequence differences among promoters generally requires biochemical and *in vivo* functional assays [8].

Secondary response Genes

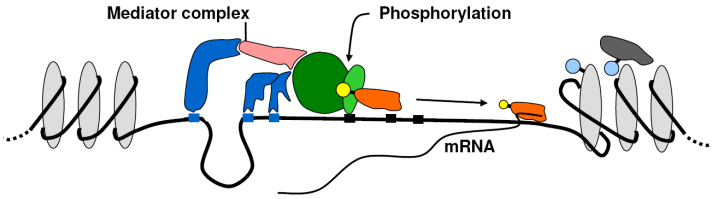
(A) Transcription factor recruits chromatin remodelling complex



(B) Preinitation complex formation at the promoter

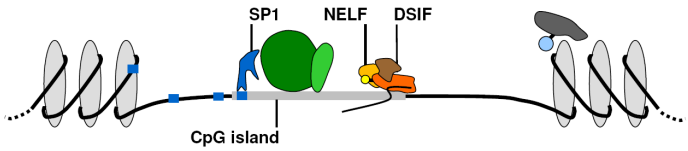


(C) Transcription: TFs induce the phosphorylation of RNAPII

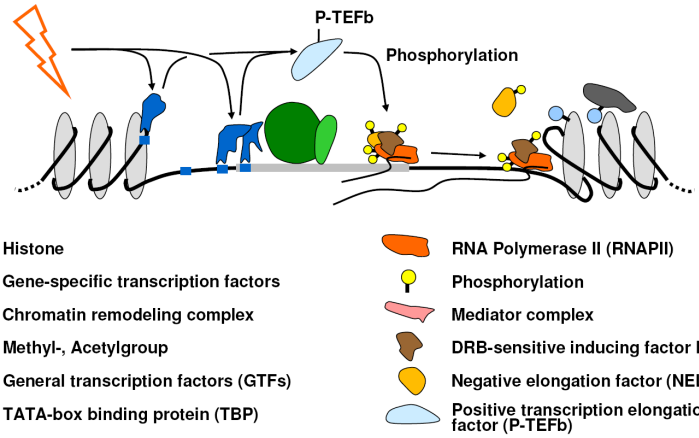


Primary response Genes

(D) Pol II stalling through association of NEFL and DSIF to RNAPII



(E) Transcription: phosphorylation of RNAPII, NELF and DSIF through recruitment of P-TEFb by DNA or chromatin bound activators



Overview figure on transcription regulation at primary and secondary response genes.

3.2. *Cis*-regulatory mutations and evolution

Differences in gene expression are a fundamental component of evolution. Such differences can arise from mutations in TFBSs that are referred to as *cis*-regulatory changes, affecting transcription initiation, transcription rates and transcript stability. On the other hand, such differences can arise from *trans*-regulatory changes that modify the activity of transcription factors interacting with *cis*-regulatory elements. It has been and still is under debate if *cis*- and *trans*-regulatory changes make qualitatively distinct contributions to phenotypic evolution [64]. However, several considerations and findings underline the importance of *cis*-regulatory changes.

The first argument was that the phenotypic impact of any gene results from two distinct components, which is not just the biochemical property of the encoded protein, but also the condition and location under which it is transcribed to fulfill it function [65]. Another argument is that *cis*-regulatory mutations affecting TFs can potentially cause a coordinated phenotypic response. TFs usually regulate several to many thousand functionally related genes and therefore, changes in their expression are more likely to produce functionally integrated phenotypic consequences [66]. Another observation was that many developmental regulatory genes are well conserved and widely spread throughout the animal kingdom. The question arose how orthologous regulatory proteins can control the development of very different organisms like flies and mice. Adaptations in the "targetome" of a TF might be the answer, stating that the battery of genes regulated by a TF, or its qualitative contribution, has changed through adaptations in *cis*-regulatory elements of target genes [67]. This argument is underlined by the finding that *cis*-regulatory mutations are often co-dominant, where natural selection acts very efficient, since heterozygote's can have immediate fitness consequences, rather than requiring genetic drift to raise allele frequencies up to the point at which homozygote's begin to appear [8].

There are numerous examples for *cis*-regulatory adaptations with phenotypic consequences for model organisms like *C. elegans*, fruit fly, mouse and others. But also in man, over 100 *cis*-regulatory mutations that segregate in human populations are known to affect diverse aspects of behavior, physiology and disease susceptibility [64, 68, 69].

Within promoter regions, *cis*-regulatory mutations in the form of single nucleotide polymorphisms (SNPs) are associated with susceptibility to diseases, such as schizophrenia [70], heart disease [71, 72], resistance to infection with malaria [73, 74] and leprosy [75].

Beyond that, promoter SNPs have been described to be involved in behavioral traits like anorexia nervosa and obsessive-compulsive disorder [76]. Also, aggressive behavior has been linked to a variable number tandem repeat polymorphism within the promoter of the monoamine oxidase A gene [77]. Another famous example is the persistence of lactase activity in most adult Europeans which is likely caused by a single genetic variant within the distal promoter that is strongly associated with lactase persistence [78]. However, *cis*-regulatory mutations contributing to human-specific traits in respect to our closest relative, the chimpanzees are few in numbers. Reported examples include cases of *cis*-regulatory mutations in promoters affecting genes involved in nutrition, immune response, neurological processes and development.

Elevated plasma levels of chimpanzee lipoprotein(a) are caused by three mutations located at -3, -2, and +8 bps relative to the TSS [79]. Similarly, a single nucleotide polymorphism in the human promoter of the multifunctional cytokine Interleukin 4 (IL4) that influences the balance of cytokine signaling in the immune system affects the binding of NFAT, a key transcriptional activator of IL4 in T cells [80]. Also, the Siglec-6 gene that is expressed in immune cells across hominids was found specifically expressed in human placental trophoblast, which could be linked to three nucleotide changes in the human promoter [81]. Furthermore, a cluster of human-specific substitutions within the promoter of prodynorphin has significant influence on gene expression [82]. Prodynorphin is the precursor molecule for a suite of endogenous opioids and neuropeptides with critical roles in regulating perception, behavior and memory, implying a functional relevance of human promoter mutations.

The most remarkable example is a human-specific gain of function in a developmental enhancer. Prabhakar et al. identified a highly conserved region among terrestrial vertebrates of 81 bps, where only humans accumulated a cluster of 13 substitutions [83]. In transgenic mice, this region induced strong limb expression including the presumptive anterior wrist and proximal thumb, while the orthologous chimpanzee region did not [84].

These examples underline the importance of identifying *cis*-regulatory mutations for the understanding of human disease and evolution.

3.3. Methods for analyzing transcription factor binding sites

The first challenge in TFBS characterization is their localization. Different approaches allowing for either “single site” or genome-wide identification of TFBSs were described in the fields of molecular and computational biology. However, the evaluation of TF binding to TFBSs on the one hand and the impact on transcription regulation on the other hand can only be addressed experimentally. Studies on TFBS influence on the expression of single genes mostly involve reporter gene assays. Approaches for genome-wide studies of TF targets and their regulatory impact involves several techniques, including ChIP-seq and TF knockdown by RNAi followed by expression profiling.

Bioinformatics approaches alone suffer mainly from false-positive predictions caused by inaccurate binding models and modest information content, since many binding sites are only 4-16 bps in length. But most importantly, TFBS function is strongly context-dependent, and we have too little knowledge on context-dependency for accurate predictions [64]. However, the combination of experimental and computational approaches can be very fruitful. An example is the recently emerged technique of ChIP-seq, were *in vivo*-occupied transcription factor binding regions are identified at high resolution. Here, within experimentally determined regions, bioinformatics tools are of great benefit, as they allow to precisely pinpointing the residing TFBSs.

3.3.1. Functional characterization of individual TFBSs

The most common assays to study individual TFBSs are DNase footprinting and mobility shift assays [9]. DNase footprinting [85] is a technique that detects DNA-protein interactions by exploiting the fact that DNA-binding proteins will often protect the bound regions from enzymatic cleavage by DNase. Subsequently, the bound region(s) can be identified by gel electrophoresis of digested and end-labeled DNA fragments, as opposed to a control sample of similarly treated DNA free of bound proteins. By comparing the two cleavage patterns of probe and control, blank regions (the 'footprints') observed in the probe lane indicate protein binding. Subsequently, binding affinities can be addressed by varying protein concentrations to find the concentration at which the footprint is observed [86].

Electrophoretic mobility shift assays (EMSA), also known as band-shift assays, are based on the principle that a protein-DNA complex migrates through a native gel more slowly than protein-free DNA fragments [87]. Purified proteins or crude cellular extracts are incubated with a radiolabeled DNA probe (other labeling methods exist) and subsequently, free DNA is separated from protein-bound DNA by gel electrophoresis.

Both techniques are widely in use and pivotal for localizing TF binding, and in addressing affinity and specificity of TFs towards specific DNA sequences. However, they cannot account for the impact of TFBSs on transcription. The only way to identify a binding site with a role in regulating transcription is to modify its sequence and assay transcription *in vivo* [8]. In this, a regulatory region is coupled to a reporter gene and assayed in embryos or cells, where it is exposed to the array of TFs that is encountered by the endogenous promoter. Common reporter genes are fluorescent proteins, such as GFP or YFP, that are non-toxic proteins and emit light when excited at certain wavelengths [88]. A great advantage of these reporters is that they can be monitored *in vivo* and at resolutions allowing detection of sub-cellular localizations.

However, for quantification of expression differences, more sensitive reporter gene systems are widely in use employing luciferases [89]. These enzymes, derived from bioluminescent organisms like Fireflies or Sea Pansy, oxidize substrates (luziferins) while emitting light at rates proportional to the enzyme concentration (as long as enough substrate is present). Normalization of reporter gene activity is necessary due to varying transfection efficiencies and is achieved by co-transfection of a second plasmid that stably expresses another luciferase converting another substrate. This system allows precise quantification of reporter gene activity with a sensitivity range spanning four orders of magnitude. Reporter-gene assays require amplification and cloning of candidate promoter regions. Here, amplification often represents the limiting step, as high GC-contents of CpG islands, that reside in more than 70% of all human promoters, hamper efficient amplification [90].

3.3.2. Bioinformatics approaches

Bioinformatics approaches include *de novo* identification of binding sites of unknown TFs, referred to as pattern detection, and searching for binding site occurrence of TFs with known binding preferences, referred to as pattern matching.

Examples for pattern detection approaches are overrepresentation studies and phylogenetic footprinting. Overrepresentation of a certain sequence motif within functionally related sequences is a hint towards functional relevance. For example, the core promoter element DPE was found by searching pools of well-defined promoter regions [91]. Another example is the recently identified "paused button" motif that was found by pooling promoters that showed high levels of RNAPII stalling [33]. For this approach, different algorithms have been developed and implemented, including GibbsSampler [92], Weeder [93] and MEME [94], which is currently widely in use.

Phylogenetic footprinting successfully introduces the filtering power of functional constraint, assuming that orthologous regions of high sequence conservation point towards functionally relevant sequence elements [95-97]. Nevertheless, this approach suffers from false positives and false negatives, as sequence conservation can occur by chance, and most TFBSs are degenerate as they can tolerate certain substitutions without losing functionality [8].

Pattern-matching approaches rely on prior knowledge of DNA sequences that are recognized by a specific TF. These sequences can be used to derive a consensus recognition sequence of the TF, which can be applied to *in silico* mapping on a genome-wide scale. Yet, a better way to represent the information content of different sequences bound by the same TF are position-specific weight matrices (PWMs) [98]. PWMs incorporate sequence variability by recording the frequencies of nucleotides at each position of the binding site.

Today, the largest database TRANSFAC (version 2009.1) contains 540 vertebrate PWMs corresponding to 336 human TFs [99, 100]. The majority of these PWMs are derived from *in vitro* SELEX assays (systematic evolution of ligands by exponential enrichment). SELEX is a technique for the specific enrichment of short oligonucleotides from random oligonucleotide pools using a TF or its DNA-binding domain as bait.

However, the problem here is that the enrichment procedure favors those oligonucleotides of highest affinity towards the bait protein under the chosen experimental conditions. Hence, the identified DNA fragments do not necessarily contain the sequences that are functional *in vivo* and can result in PWMs that do not reflect the true binding preferences of the assayed TF [101-103]. Therefore, also *in silico* mapping based on *in vitro*-derived PWMs are hampered by both false negatives and false positives [104]. However, there techniques have emerged that will

shine new light upon this issue. ChIP-seq together with pattern discovery and pattern matching can reveal TFBSs occupied *in vivo* and produce high quality PWMs [105], potentially lifting TFBS predictions to the next level.

3.3.3. Genome wide approaches, chromatin IP and 2nd-generation sequencing

Chromatin immunoprecipitation (ChIP) involves cross-linking of protein-DNA complexes in living cells [106]. The treatment of cells with formaldehyde covalently links genomic DNA with proteins of close proximity (~2 Å), thereby freezing the endogenous interactions [107]. Subsequently, chromatin is extracted and sheared into 150-300 bp fragments. Using a specific antibody allows the precipitation of the protein of interest together with the cross-linked DNA. After the cross-linking has been removed, free DNA fragments can be further analyzed by hybridization to DNA microarrays. DNA microarrays are slides spotted with tens of thousands of single strand DNA probes. Target DNA molecules, as derived from a ChIP, are fluorescently labeled and hybridized to the array of defined probes to be identified. This approach, called ChIP-chip, is well established and widely in use [108]. After whole-genome DNA microarrays have become available, limitations for *de novo* binding site identifications remained in the resolution, allowing for the identification of ~1kb regions, and in cross hybridization of DNA fragments to inappropriate probes. The recently introduced technique of massively parallel sequencing or 2nd generation sequencing opens a new era in genome- and transcriptiome-wide studies. This new sequencing approach generates several million short sequence reads (~35 bps) of accurate nucleotide sequence per experiment [109]. In this technique, common adaptors are ligated to fragmented DNA molecules. Subsequently, single molecules are immobilized onto a flat surface, were their amplification results in an array of millions of spatially immobilized PCR colonies. These colonies serve as templates for 35 rounds (or more) of sequencing by synthesis with fluorescent reversible terminator deoxyribonucleotides, to build up a contiguous sequencing read. The detection of fluorescent labels incorporated with each round of extension allows acquiring sequencing data on all features in parallel [109, 110]. Currently, the main applications for this technology are re-sequencing of genomes, transcriptome sequencing (RNA-seq) and most importantly in this context, sequencing following chromatin immunoprecipitation. Both techniques, ChIP-chip and ChIP-seq, permit the genome-wide identification of *in vivo* bound regions. However, ChIP-seq has several advantages over ChIP-chip. Most importantly, ChIP-seq enables more precise mapping of protein binding sites, has a higher dynamic range, is less prone to artifacts, such as cross-hybridization on microarrays, and produces disproportionally more data, including sequence information [111].

4. Aim of the project

What makes us human? In terms of genetics, this question is on topic since 1975, when King and Wilson postulated that protein differences of human and chimpanzee cannot account for the phenotypic differences between the two species. They postulated: "*A relatively small number of genetic changes in systems controlling the expression of genes may account for the major organismal differences between humans and chimpanzees*" [112]. However, more than three decades later, only a handful of potentially involved changes, including some experimental support, have been described [81, 82, 84, 113]. Hence, to develop a better understanding on the causes of the human uniqueness among primates, new and powerful approaches are needed for tracing DNA mutations potentially involved in phenotypic differences that are worthy for experimental testing. The work described here involved the design and implementation of bioinformatics approaches, the establishment of site-directed mutagenesis, cloning and reporter gene assays, as well as the development of an enhanced PCR protocol. This work aimed at the identification of functional *cis*-regulatory adaptations in the human lineage, including experimental validation in respect to the transcriptional impact of these adaptations.

5. Manuscript I

5.1. An efficient and economic enhancer mix for PCR

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 347 (2006) 747–751
BBRC
www.elsevier.com/locate/ybbrc

An efficient and economic enhancer mix for PCR ☆

Markus Ralser [1], Robert Querfurth [1], Hans-Jörg Warnatz, Hans Lehrach, Marie-Laure Yaspo, Sylvia Krobitsch *

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received 19 June 2006
Available online 5 July 2006

Abstract

Polymerase chain reaction (PCR) has become a fundamental technique in molecular biology. Nonetheless, further improvements of the existing protocols are required to broaden the applicability of PCR for routine diagnostic purposes, to enhance the specificity and the yield of PCRs as well as to reduce the costs for high-throughput applications. One known problem typically reported in PCR experiments is the poor amplification of GC-rich DNA sequences. Here we designed and tested a novel effective and low-cost PCR enhancer, a concentration-dependent combination of betaine, dithiothreitol, and dimethyl sulfoxide that broadly enhanced the quantitative and/or qualitative output of PCRs. Additionally, we showed that the performances of this enhancer mix are comparable to those of commercially available PCR additives and highly effective with different DNA polymerases. Thus, we propose the routine application of this PCR enhancer mix for low- and high-throughput experiments.

Keywords: Polymerase chain reaction; GC-rich sequence; Enhancer; Additive; Promoter PCR; Genomic PCR; PCR template; *Taq* DNA polymerase

The polymerase chain reaction (PCR) was developed in the 1980s by Kary Mullis and Fred Faloona [1,2]. Starting with the biotechnological application of thermostable DNA polymerases [3], PCR has become a fundamental technique in molecular biology. There are ever-increasing needs for further improvements of PCR protocols for low-cost and efficient high-throughput approaches in a wide range of applications ranging from quantitative analysis at the genome- or transcriptome levels to routine diagnostic purposes [4,5]. Large-scale PCR experiments require broadly applicable and reliable reaction conditions for establishing cost-effective production pipelines. *Taq* DNA polymerase, originally purified from the thermophilic bacterium *Thermus aquaticus* [6], is widely used, since it can be produced in every standard laboratory at low-cost [7–9]. One major factor limiting the output of PCR routines is that a number of DNA sequences are poorly or not amplifiable under standard reaction conditions, either because of their intrinsic properties to form secondary structures, and/or because of their high GC-content. Improvements of the PCR conditions can be achieved by modifying the classical reaction conditions, for example, by performing "touchdown" PCR, consisting of a stepwise reduction of the annealing temperature for each cycle [10], or by the use of modified DNA polymerases for carrying out "hot-start" reactions [11]. Typically, to overcome amplification problems of GC-rich DNA, the addition of substances that enhance the specificity and/or the yield of the PCR is necessary. The most prominent PCR enhancing additives that are currently used are either betaine [12], small sulfoxides like dimethyl sulfoxide (DMSO, [13]), small amides like formamide [14] or reducing compounds like β-mercaptoethanol or dithiothreitol (DTT, [10]). However, their capacity to significantly improve PCR yields mainly for

☆ *Abbreviations:* BSA, bovine serum albumin; CES, combinatorial PCR enhancer solution; DMSO, dimethyl sulfoxide; DTT, dithiothreitol; PCR, polymerase chain reaction; *Taq*, *Thermus aquaticus*.

* Corresponding author.
E-mail addresses: ralser@molgen.mpg.de (M. Ralser), krobitsc@molgen.mpg.de (S. Krobitsch).

[1] These authors contributed equally to this work.

0006-291X/S - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.06.151

high-throughput experiments is marginal. Commercial enhancers have led to better results but with two major drawbacks, their cost and the fact that their chemical composition is unknown.

Mammalian promoter sequences often contain highly GC-rich regions, which are difficult to amplify under standard reaction conditions [15]. In this study, we tested the efficacy of concentration-dependent combinations of different PCR additives for a reliable amplification of genomic DNA corresponding to a set of human promoter sequences and generated a novel, cheap, and flexible PCR enhancer.

Materials and methods

Primer design. PCR primers for the amplification of ~1000–1600 bp sized DNA fragments from human genomic DNA were designed using the "Primer 3.0" online service [16] on the basis of the human genome annotation build 35.1 (NCBI). The primer sequences, locus information, the overall GC-content, and the size of the expected amplicons are given in the supplementary material Table 1.

Purification of human genomic DNA. Human genomic DNA was prepared from oral mucosa. The mucosal smear was washed with water and dissolved in 400 μl lysis buffer (50 mM Tris–HCl, 10 mM EDTA, and 2% SDS, pH 8.8). The suspension was incubated for 5 min at 65 °C, then supplemented with 250 μl 4.5 M NaCl and cleared by centrifugation. Genomic DNA was recovered from the supernatant by isopropanol precipitation.

PCR conditions. PCRs were performed in a 30 μl volume in 96-well microtiter plates. Reaction buffer contained 65 mM Tris–HCl, 16.6 mM $(NH_4)_2SO_4$, 3.1 mM $MgCl_2$, and 0.01% (v/v) Tween 20, pH 8.0. 2.5 U *Taq* DNA polymerase purified from *Escherichia coli* according to the method of Engelke et al. [7], 0.6 μmol of each oligonucleotide, and 25 μmol dATP, dTTP, dCTP, and dGTP were added prior to the cycling reaction. The cycling reactions were performed in a PTC-200 Thermocycler (MJ Research) with an initial denaturation for 5 min at 96 °C followed by the thermal cycles as follows: denaturation step at 98 °C for 15 s, annealing step at 72 °C for 40 s, and an elongation step at 72 °C for 90 s. The annealing step was started at a temperature of 66 °C and declined in 0.5 °C steps for each cycle until a temperature of 56 °C was reached. Subsequently, 30 additional cycles were performed with a constant annealing temperature of 52 °C. The reaction was completed with a final elongation step at 72 °C for 2 min. PCR products were analyzed with agarose-gel electrophoresis and stained with ethidium bromide (Sigma).

Results and discussion

In the context of a systematic project aiming at the functional analysis of promoter elements, we set out to amplify 110 human promoter sequences from genomic DNA using classic touch-down PCR conditions (as described in Materials and methods). We observed that approx. 30% of the promoter regions could not be correctly amplified, either because the PCR products were unspecific or because of the poor yield of the amplicons. Most of these had an overall GC-content of 50–75% (62% in average).

In order to improve these results, we evaluated different PCR enhancing additives for their capacity to promote the amplification of three different gene promoter regions (SIM2, DIP2A, and SLC19A1, please refer to the supplementary material for detailed information) whose GC-content ranged from 71% to 75%. We designed three primer pairs for these promoters (named A for SIM2, B for DIP2A, and C for SLC19A1) and carried out touch-down PCR supplemented with different concentration ranges of the PCR additives betaine [12], dithiothreitol (DTT) [10], dimethyl sulfoxide (DMSO) [13], or formamide [14] as indicated in Fig. 1. We observed that betaine had the best PCR enhancing properties at a concentration of 0.8 M in all PCR samples for primer pairs A, B, and C, whereas DTT and DMSO were less effective since the PCR output was enhanced only for one gene out of three (3.2 mM DTT for primer pair A, or 3.2% DMSO for primer pair B, respectively) (Fig. 1). No PCR enhancing effects were observed by adding formamide to the respective PCRs at any of the indicated concentrations.

On this basis, we generated a 5-times concentrated preliminary combinatorial enhancer solution (preCES-I) composed of 4 M betaine, 16 mM DTT, and 16% DMSO. We included 83 μg/ml bovine serum albumin (BSA) in the solution, since BSA, which has no direct effect on the enzymatic reaction *per se*, can stabilize enzymes and neutralize inhibitory contaminants that may be present in the DNA

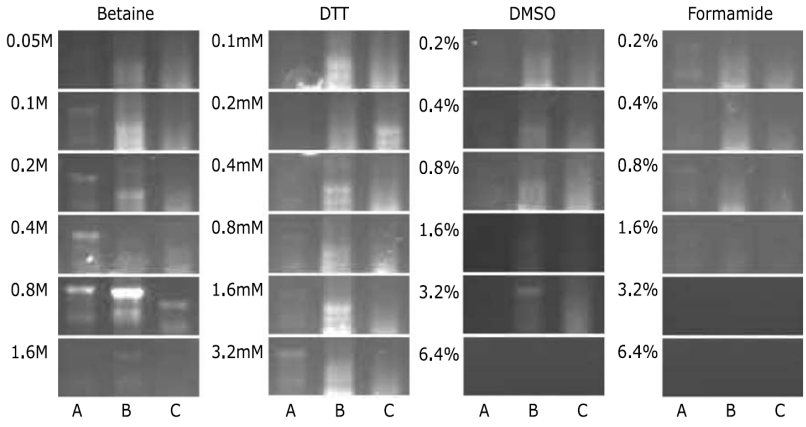


Fig. 1. Enhancing properties of known PCR additives. Betaine, DTT, DMSO, and formamide were applied to genomic PCRs at the indicated final concentrations. Letters represent the primer pairs used.

template preparation or in the reaction buffers [10,17]. Since high compound concentration could potentially inhibit the activity of the *Taq* DNA polymerase, we additionally tested two other preCESs containing lower concentrations of the respective additives, preCES-II (4 M betaine, 10 mM DTT, and 10% DMSO) and preCES-III, (2 M betaine, 5 mM DTT, and 5% DMSO). To analyze the efficiency of these preliminary enhancer solutions, we selected 12 (9 additional) primer pairs, of which 10 failed to produce adequate PCR products under standard conditions without additive. These primer pairs produced either non-specific products (H and I), prominent additional bands to the expected product (C, D, and L), very low yield (A, E, and K), or no product at all (B and G) (Fig. 2A). Subsequently, PCRs were repeated with these primers with or without preCESs I, II, or III. As demonstrated in Fig. 2A, the output of 10 out of 12 PCRs analyzed was enhanced by at least one of the three preCESs. For primer pairs B and G, which have not resulted in any detectable PCR product, the addition of the preCESs resulted in the amplification of a specific DNA fragment. For primer pairs C, D, H, L, or I, respectively, the preCESs enhanced the specificity of the PCRs, whereas for primer pairs A, E, and K the presence of at least one preCES resulted in a significantly improved product yield. However, the addition of preCESs did not improve the PCR performed with primer pair F, and in one case the addition of preCESs had a negative effect on the PCR yield (primer pair J).

Thus, these experiments clearly demonstrated that the addition of a preCES enhances the yield and/or the specificity of PCRs in virtually all cases, particularly for the amplification of highly GC-rich sequences up to 75%. Among the three tested enhancers, preCES-II containing the intermediate concentrated enhancer solution appeared to perform best, as visualized in Fig. 2B.

In the next step, we further optimized further the compound concentration of the preCES-II. Initially, the 30 μl PCR mixture was supplemented with incremental quantities of preCES-II in 2 μl steps (ranging from 0% to 40% of the final volume) and PCRs were performed with the various primer pairs as indicated (Fig. 2C). Best results in terms of specificity and yield were obtained with addition of 4 μl preCES-II to the 30 μl reaction volume, corresponding to final concentrations of 0.54 M betaine, 1.34 mM DTT, 1.34% DMSO, and 11 μg/ml BSA. Thus, we generated a 5-times concentrated combinatorial enhancer solution termed CES.

In a third step, we compared the efficiency of our CES with those of three commercial PCR enhancer solutions, namely Q-solution (Qiagen), PCR enhancer solution (Invitrogen), and Hi-Spec PCR additive (Bioline). For this comparative analysis, we selected 32 primer pairs designed for the amplification of DNA fragments with a GC-content ranging from 33% to 75% (Fig. 3A). PCRs were performed using the reaction buffer without additives or supplemented with the Q-solution, PCR enhancer solution, Hi-Spec PCR additive or our CES. As demonstrated in Fig. 3A, commercial PCR enhancers could improve 90% of the PCRs. The CES described in this study led to comparable performances in all tested PCRs. A major advantage of the CES is that it is much more economic for laboratory routine applications and that its composition is well described and can thus be tuned whenever necessary for more specific applications. Furthermore, Qiagen's Q-solution and Invitrogen's PCR enhancer mix can only be purchased conjoint with the suppliers *Taq*

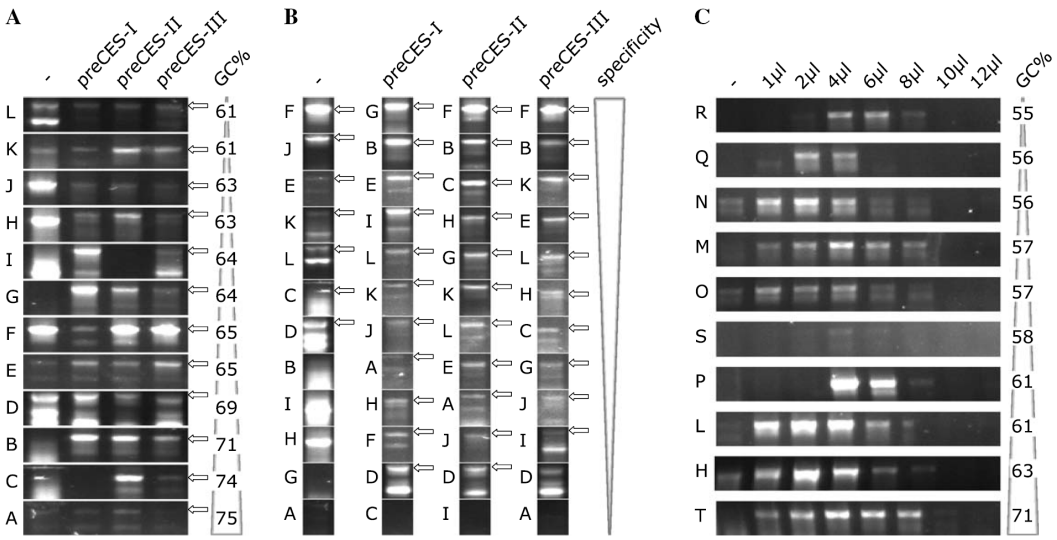


Fig. 2. Generation of a combinatorial enhancer solution (CES). (A) Comparative analysis of PCRs without enhancing additives (−) and PCRs supplied with preCES-I, preCES-II, or preCES-III, respectively. Primer pairs are sorted by ascending GC-content of the expected PCR product. Arrows highlight the specific DNA fragments. (B) Like (A), but sorted by descending product specificity and yield. (C) Concentration-dependent application of preCES-II to PCRs performed with primer pairs for the amplification of DNA sequences with varying GC-content.

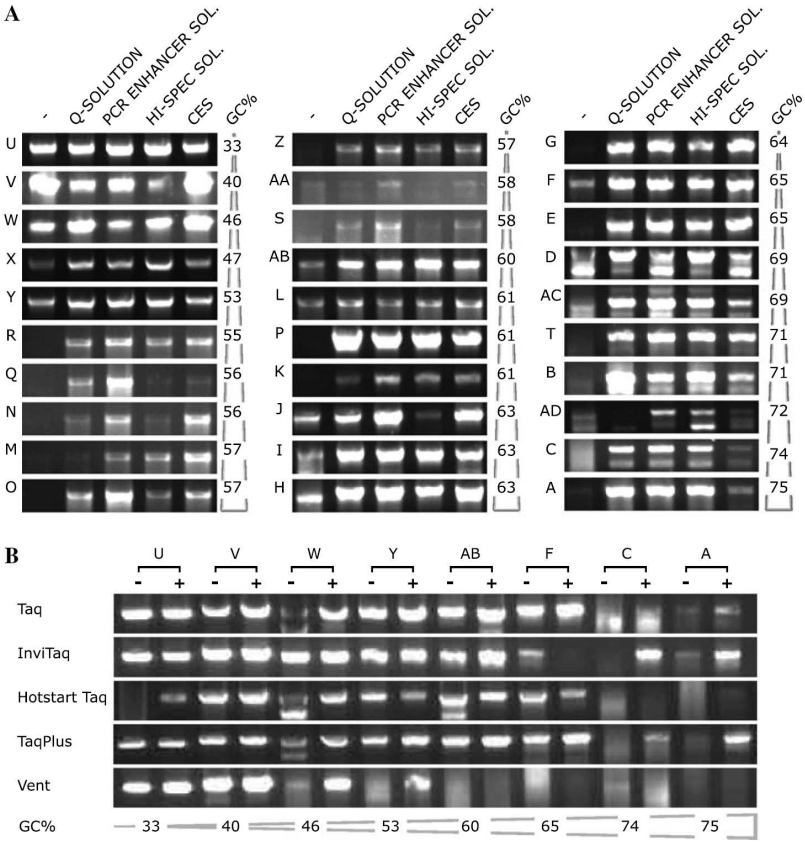


Fig. 3. Validation of the CES. (A) Comparison of PCRs without enhancing additives and PCRs supplied with Q-solution (Qiagen), PCR enhancer solution (Invitrogen), Hi-Spec PCR additive (Bioline), and the elaborated CES. Primer pairs are sorted by ascending GC-content of the expected DNA product. (B) Analysis of CES in combination with commercial DNA polymerases. PCRs were performed with (+) or without (−) CES using the commercial DNA polymerases as indicated.

polymerase (Qiagen) or with a proprietary reaction buffer (Invitrogen). To exclude that the PCR enhancing effects of CES are limited to PCRs performed with our home-made *Taq* DNA polymerase, we further analyzed the performance of CES with commercial polymerases. Using DNA polymerases like InviTaq (Invitek), HotStartTaq (Qiagen), TaqPlus (Stratagene) or even with Vent polymerase, originally purified from *Thermococcus litoralis* ([18], New England BioLabs), we amplified eight different genomic DNA fragments with or without CES (Fig. 3B). The majority of PCRs with inadequate products were enhanced and virtually no negative effects resulting from the addition of the CES were observed in all reactions. Thus, our enhancer solution can be used with any of the tested DNA polymerases. Interestingly, the extremely GC-rich DNA fragments resulting from primer pair A or C (75% and 74%, respectively) that were poorly amplified with the home-made *Taq* DNA polymerase even in the presence of CES were satisfactorily amplified using CES in combination with the InviTaq or TaqPlus enzyme, respectively (Fig. 3A and B). Finally, to support the broad applicability of this PCR enhancer, we also tested the CES on other types of template DNA like yeast genomic DNA, plasmids, and even glycerol stocks, and detected PCR enhancing effects of the CES (data not shown). The 5-times concentrated CES containing 2.7 M betaine, 6.7 mM DTT, 6.7% DMSO, and 55 μg/ml BSA was stable at −20 °C for at least 3 months. Since different *Taq* reaction buffers currently in use show diverse performances, we recommend to use a reaction buffer containing final concentrations of 65 mM Tris–Cl, 16.6 mM $(NH_4)_2SO_4$, 3.1 mM $MgCl_2$, and 0.01% (v/v) Tween 20 at a pH of 8.8 as described in the Materials and methods section.

In summary, we have demonstrated that the concentration-dependent combination of the known PCR additives betaine, DMSO, and DTT results in a cost-effective PCR enhancer solution showing equivalent performances compared with commercial enhancers, at least under the chosen experimental conditions. Since the CES is composed of low-cost components, the usage of this PCR enhancer solution is especially advantageous and attractive from the economic perspective for large-scale projects and routine applications requiring reliable PCR results.

Acknowledgments

We are grateful to Richard Reinhardt and Roman Pawlik (Max Planck Institute for Molecular Genetics, Berlin,

Germany) for providing *Taq* DNA polymerase and to our lab members for critical discussions. This work has been funded by the Max Planck Society and the Federal Ministry of Education and Research (BMBF) in the framework of the National Genome Research Network (NGFN) under project 01GR0414.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.06.151.

References

[1] K.B. Mullis, The unusual origin of the polymerase chain reaction, Sci. Am. 262 (1990) 56–61, 64–65.
[2] R.K. Saiki, S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, Science 230 (1985) 1350–1354.
[3] R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, H.A. Erlich, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, Science 239 (1988) 487–491.
[4] G. Csako, Present and future of rapid and/or high-throughput methods for nucleic acid testing, Clin. Chim. Acta 363 (2006) 6–31.
[5] C. Ding, C.R. Cantor, Quantitative analysis of nucleic acids—the last few years of progress, J. Biochem. Mol. Biol. 37 (2004) 1–10.
[6] F.C. Lawyer, S. Stoffel, R.K. Saiki, S.Y. Chang, P.A. Landre, R.D. Abramson, D.H. Gelfand, High-level expression, purification, and enzymatic characterization of full-length *Thermus aquaticus* DNA polymerase and a truncated form deficient in 5′ to 3′ exonuclease activity, PCR Methods Appl. 2 (1993) 275–287.
[7] D.R. Engelke, A. Krikos, M.E. Bruck, D. Ginsburg, Purification of *Thermus aquaticus* DNA polymerase expressed in *Escherichia coli*, Anal. Biochem. 191 (1990) 396–400.
[8] F.C. Lawyer, S. Stoffel, R.K. Saiki, K. Myambo, R. Drummond, D.H. Gelfand, Isolation, characterization, and expression in *Escherichia coli* of the DNA polymerase gene from *Thermus aquaticus*, J. Biol. Chem. 264 (1989) 6427–6437.
[9] F.G. Pluthero, Rapid purification of high-activity Taq DNA polymerase, Nucleic Acids Res. 21 (1993) 4850–4851.
[10] M. Nagai, A. Yoshida, N. Sato, Additive effects of bovine serum albumin, dithiothreitol, and glycerol on PCR, Biochem. Mol. Biol. Int. 44 (1998) 157–163.
[11] D.E. Kellogg, I. Rybalkin, S. Chen, N. Mukhamedova, T. Vlasik, P.D. Siebert, A. Chenchik, TaqStart Antibody: "hot start" PCR facilitated by a neutralizing monoclonal antibody directed against Taq DNA polymerase, Biotechniques 16 (1994) 1134–1137.
[12] W. Henke, K. Herdel, K. Jung, D. Schnorr, S.A. Loening, Betaine improves the PCR amplification of GC-rich DNA sequences, Nucleic Acids Res. 25 (1997) 3957–3958.
[13] R. Chakrabarti, C.E. Schutt, The enhancement of PCR amplification by low molecular-weight sulfones, Gene 274 (2001) 293–298.
[14] G. Sarkar, S. Kapelner, S.S. Sommer, Formamide can dramatically improve the specificity of PCR, Nucleic Acids Res. 18 (1990) 7465.
[15] F. Hube, P. Reverdiau, S. Iochmann, Y. Gruel, Improved PCR method for amplification of GC-rich DNA sequences, Mol. Biotechnol. 31 (2005) 81–84.
[16] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol. Biol. 132 (2000) 365–386.
[17] T. Maniatis, E.F. Fritsch, J. Sambrook, Molecular cloning : a laboratory manual, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1987.
[18] F.B. Perler, D.G. Comb, W.E. Jack, L.S. Moran, B. Qiang, R.B. Kucera, J. Benner, B.E. Slatko, D.O. Nwankwo, S.K. Hempstead, et al., Intervening sequences in an Archaea DNA polymerase gene, Proc. Natl. Acad. Sci. USA 89 (1992) 5577–5581.

5.2. Supplemental material

Markus Ralser, Robert Querfurth, Hans-Jörg Warnatz, Hans Lehrach, Marie-Laure Yaspo and Sylvia Krobitsch

Table S1: primer pairs used in this study

HGNC Gene symbol	Entrez GeneID	Primer pair Identifier	~Tm	GC% of amplicon	amplicon length	forward primer	reverse primer
SIM2	6493	A	60	75	1202bp	ccgttttcacgtgtgtgtgt	ttccttctccctcctggtct
APP	351	AA	66	58	1235bp	tgctacttcaggtcaagagcaggg	aggcggccagcaggagca
BTG3	10950	AB	71	60	1215bp	tcccaagcctagtgggcagtaaggaatc	gtgtcctggccgggaactgagg
PAFAH1B1	5048	AC	68	69	1256bp	gatacagttccaggcctttctttggg	gtctctcactcaacggccgtcg
LSS	4047	AD	62	72	1252bp	ctcactaggctggggcagtt	gtgccctccgtcattgct
DIP2A	23181	B	60	71	1185bp	aggggaaggaagcaggact	cagctcagccaggctctc
SLC19A1	6573	C	59	74	980bp	ccttctgttctgtgcagtgg	cggactccgggactacag
MGMT	4255	D	71	69	1222bp	ggatgaggggcccactaattgatgg	gtaaggcaggggctgccacg
PTTG1IP	754	E	72	65	1414bp	cccaaatcccaacctaaaatcaccacagg	accgaggcgcaacctccagtacg
CTNNB1	1499	F	59	65	1292bp	taatcgatagctttctctataaacatacttg	ttggctccgagaggaagc
PKNOX1	5316	G	72	64	1465bp	gttccaacacctattgacacttgcatctggatct	acactgacaagcggctgcagcaatc
MGMT	4255	H	72	63	1428bp	tggatcctgcaagtcacaaatacgaaggtatg	gttctaggggcgcgggctgtc
MCPH1	79648	I	67	63	1379bp	tttgaggctgcatatactcaaggcaat	ggttttggtgggacaggcagc
LHX1	3975	J	72	63	1876bp	aaccttccacaaggctcggtcttggactac	agaagcacttctcggtcaggttgcatttaca
CLIC6	54102	K	61	61	1137bp	acttggaggcagaagcactg	ctggctctccgggtctct
BACH1	571	L	72	61	1452bp	gcaagaacttcaatccttctttcatggtgtatttcc	gcgcggcccgactgactga
TMEM50B	757	M	72	57	1333bp	agctataggagaacattatccaggatggcatttttg	gaaggagactgctgcgccacaacc
CASP3	836	N	73	56	1427bp	tccctatagtcgaataggcgcaagtgttagaaacag	tctacaaccgcctcacaatagcacccatc
NRCAM	4897	O	81	57	1432bp	tatcatgctggttcaggaaccgagggaggctctgtg	ctggaccgccggtgtcctccgttctcgac
USP25	29761	P	69	61	1430bp	ccaagcttgctttccttgtcattgg	agcacgttctgctccacggtcat
PAFAH1B1	5048	Q	72	56	1660bp	gggtccaggatttacacctaagttgtctctttcg	tctcgctccctagactcccgtgtcg
PRDM15	63977	R	72	55	1448bp	aggcagagaaaccagccttcacagatcaa	ggaaactgcgcagcaccggaag
PFKL	5211	S	73	58	1647bp	cagcacatggatggactgcattgtgttc	acgcccgcagcttctccaggtc
SLC19A1	6573	T	71	71	1288bp	gccaacaaatccttttttaagttcctttgagatttg	gactccgggactacagcgcccac
LIPI	149998	U	65	33	1161bp	cttgacttactaaaatcagatgccctcaa	acataaataagaccctatcagaagttcactagctct
CLDN8	9073	V	63	40	1367bp	atttttctcagataaacatttatgcttagtatagcac	tccacagtccttcccagaagttttc
TMPRSS3	64699	W	69	46	1267bp	gctcaccttactgactaattaaagggaagccac	caaagccctttccattgctttttgcg
TCP10L	140290	X	72	47	801bp	acgtcccaagcaggctcaggtgag	tggggtcacggtcctcacagcc
CDK5RAP2	55755	Y	65	53	1273bp	agaggaaggagcaacactgagttgagg	ggtgacgtcctcttccaacacc
APP	351	Z	73	57	1237bp	tgtggcttggtaactaaatgctacttcaggtcaaga	cagtgccaaaccgggcagcatc

Abbreviations: GC%: overall GC-content of the expected amplicons; HGNC: Human Genome Association Nomenclature Consortium; Tm: primer melting temperature calculated by "Primer3"

5.3. Contributions

Markus Ralser: initial idea, and wrote the majority of the manuscript

Hans-Jörg Warnatz: was involved in discussion and writing of the manuscript

Hans Lehrach: involved in discussion

Marie-Laure Yaspo: was involved in writing of the manuscript

Sylvia Krobitsch: was involved in writing of the manuscript

6. Manuscript II

6.1. Analysis of activities, response patterns and cis-regulatory elements of human chromosome 21 gene promoters

Analysis of activities, response patterns and *cis*-regulatory elements of human chromosome 21 gene promoters

Hans-Jörg Warnatz[1], Robert Querfurth[1], Anna Guerasimova[1], Xi Cheng[1], Dominique Vanhecke[2], Andrew Hufton[1], Stefan Haas[1], Wilfried Nietfeld[1], Martin Vingron[1], Michal Janitz[3], Hans Lehrach[1], and Marie-Laure Yaspo[1]*

[1]Department for Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany, [2]Department Biomedicine, University Hospital Basel, Hebelstrasse 20, 4031 Basel, Switzerland and [3]School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney NSW 2052, Australia

*To whom correspondence should be addressed. Tel: +49 30 8413 1356; Email: yaspo@molgen.mpg.de

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

ABSTRACT

Given the inherent limitations of *in silico* studies relying solely on DNA sequence analysis, the functional characterization of mammalian promoters and associated elements requires experimental confirmation, demanding for cloning and analysis of putative promoter regions. Focusing on systematic promoter studies for human chromosome 21, we cloned 182 promoters of 2,500 base pairs length to conduct reporter gene assays on transfected-cell arrays for testing under different conditions. We found 56 promoters active in HEK293 cells, while another 49 promoters could be activated by treatment of cells with Trichostatin A or depletion of fetal calf serum. We observed high correlations between promoter activities and endogenous transcript levels, RNA polymerase II occupancy, presence of CpG islands and core promoter elements. We tested a subset of 62 truncated promoters (~500 bp) and found that truncation hardly resulted in loss of activity, but rather in loss of responses to external stimuli, supporting the presence of *cis*-regulatory response elements within distal promoter regions. In these regions, we found a strong enrichment of binding sites for transcription factors that integrate signals from the administered stimuli into gene expression. The identified promoter activity and response patterns represent a valuable resource for the elucidation of the complex mechanisms governing transcriptional regulation on the level of promoters under different conditions on a whole-chromosome scale.

INTRODUCTION

Gene expression in eukaryotic organisms requires coordinated regulation of thousands of genes. The challenge is to unravel the components and function of complex genetic networks and underlying regulatory processes. Although these processes are integrated at many different levels of the cellular machinery, the regulation of the initiation of transcription is essential and often the rate-limiting step (1) and involves mainly promoter regions located immediately upstream of the gene transcription start sites (TSSs). These regions can integrate various signals to control transcription rates of associated genes, such as spatial and temporal signals during development, hormonal, physiological and environmental signals (2). Promoter regions usually comprehend a core promoter within 50-100 base pairs surrounding the TSS (3), proximal response elements located up to 250 base pairs upstream of the TSS, and distal response elements, which can reside several kilo bases upstream of the TSS. The core promoter contains transcription factor binding sites (TFBSs) recognized by the general transcription factors and regulates basal transcription levels, while proximal and distal promoter regions are believed to harbor gene-specific TFBSs integrating additional signals for fine-tuning of transcription rates (4).

Mammalian promoter regions have been investigated mainly on a gene-by-gene basis using various reporter gene assays. Attempts to map promoter elements by computational analysis of DNA sequence widely made use of TFBS predictions and evolutionary conservation of short DNA stretches (5), but these approaches showed limitations due to the heterogenous nature of promoter regions and require experimental verification (2). To date, there have been few studies analyzing large numbers of promoters in parallel. A study describing a plasmid library-based approach designed to select human genomic fragments with promoter activity reported that 68% of the 130 tested fragments were active in transient transfection reporter assays (6). In the framework of the ENCODE project scanning 1% of the human genome, putative promoters were tested by high-throughput transient transfection reporter assays, which unveiled that 60% of 642 tested promoters of ~600 bp length showed activity in at least one of 16 cell lines (7). A follow-up study aiming at the validation of novel putative promoters predicted from the analysis of cDNAs found that 25% of the 163 tested fragments harbored promoter function (8). These and other studies converge in postulating that mammalian upstream regulatory regions represent a heterogeneous group with disparate structural features and cell-type specific activities (2,9,10). Additional systematic experimental analyses are necessary to gain broader insight into promoter structure and function, especially in the evaluation of the activities of large numbers of promoters in parallel.

We recently established a procedure using transfected-cell arrays for the functional characterization of promoters (11). Here, we expand this approach to the analysis of the promoters from the genes encoded on human chromosome 21 (HSA21), which is well annotated and serves as a model for pilot genomic studies due to its small size (48 Mega bases) and its medical relevance, in particular trisomy 21 or Down syndrome. For 231 annotated genes on HSA21, we cloned 182 promoter fragments of 2.5 kilo bases in size upstream of the presumed TSS, as well as a set of truncated fragments (500 bp upstream of the TSS) for 62 promoters. We used transfected-cell arrays (11-16) to carry out promoter reporter assays in HEK293 cells under normal growth conditions and after treatments known to alter gene expression. We correlated promoter activities with the presence of core promoter elements, with endogenous expression levels, with expression patterns derived from EST data for 45 different human tissues and with RNA polymerase II occupancy on promoters in HEK293 cells (17). We show that treatment of cells with different stimuli, in combination with analysis of promoter fragments of differing lengths, yields insights into the presence of functional *cis*-regulatory elements contained in the tested promoters. Taken together, we generated the first chromosome-scale reference data set on the structure, function and responses of human gene promoters.

MATERIALS AND METHODS

Promoter annotation, primer design and vector construction

Promoter annotation and primer design was based on human gene annotations from the Ensembl database v30. Promoter regions were defined relative to the most upstream TSS of all annotated transcripts. PCR primers for all 231 genes of chromosome 21 were designed using the software PRIDE (18) for amplification of ~2.5 kb fragments directly upstream of the respective TSS, including the TSS itself. In total, 223 primer pairs were obtained, to which 12 bases of adapter sequences were added for recombinatorial cloning after two-step PCR amplification of the fragments (Gateway technology, Invitrogen). The same approach was used for cloning of truncated promoter fragments of ~500 bp upstream of the TSS.

Promoter cloning

Touch-down PCR from genomic DNA was performed according to a protocol optimized for amplification of GC-rich promoter regions (19), using as templates genomic DNA as well as available genomic BAC and fosmid clones. Then, a secondary PCR was performed with Gateway adapter primers (Invitrogen), followed by PEG-8000 precipitation of PCR products. The modified reporter gene vector pZsGreen1-1 and the control plasmid pHcRed1-N1 were used as decribed before (11). PCR products were cloned into the pZsGreen vector using Gateway BP Clonase II Enzyme Mix (Invitrogen) and transformed into competent TOP10 cells following the manufacturer's recommendations. Resulting colonies were screened by colony PCR, plasmids were isolated from positive clones using a QIAprep Spin Miniprep Kit (Qiagen), and inserts were confirmed by 5' and 3' end sequencing.

Microarray spotting, cell culture and reverse transfection

Samples for array spotting were prepared as previously described (11). Spotting solutions containing 32 ng/µl of promoter construct and 7.5 ng/µl of reference plasmid were kept at 4°C until arraying. Automated spotting was performed with a high-speed non-contact dispensing system (instrumentONE, M2 Automation). Arrays were printed onto home-made poly-L-lysine (Sigma) coated microscope glass slides using a 500 µm outlet port solenoid valve, which delivered 20 nl of sample per spot. Average spot to spot center distance was 1.5 mm. Samples

were arrayed in triplicates. After arraying, slides were maintained in low humidity condition at 4°C. Human embryonic kidney cells (HEK293 from ATCC) were cultured in Dulbecco's modified Eagle's medium (DMEM, Gibco Invitrogen) supplemented with 10% (v/v) fetal calf serum (Biochrom) at 37°C in a humidified 6% CO_2 incubator. One day prior to transfection, cells were seeded in a 60 cm^2 culture plate in 10 ml of medium. On the day of transfection, cells were washed with PBS, detached with Accutase (PAA Laboratories) and seeded at $3.5x10^6$ per slide onto printed slides, which were placed into a QuadriPerm chamber (Greiner) for reverse transfection. For each treatment, two slides were used in parallel, so that for each construct, six replicate spots could be analyzed. For treatments after 24 hours of incubation at 37°C with 6% CO_2 in DMEM supplemented with 10% fetal calf serum, the medium was changed to DMEM with 200 nM Trichostatin A (Sigma) or fetal calf serum-free DMEM. After 48 hours of transfection, slides were washed with PBS, fixed in 3.7% formaldehyde with 4 M sucrose in PBS for 30 min, stained with DAPI and mounted with Fluoromount-G (Southern Biotech). The slides were kept in the dark at 4°C until analysis.

Image acquisition, object detection and scoring of promoter activity

Microscopy images were acquired and fluorescent objects were detected as previously described (11). The average total number of cells per image frame was ~500, as controlled by DAPI staining. In this area, the maximum number of cells that could theoretically be transfected, i.e. cells found in the area of the spotted DNA, was estimated to be ~300. For each scanning position, the number of red cells, green cells and co-transfected cells (red and green) was determined. The median number of transfected cells (positive for either fluorophore) was between 50 and 70 cells, resulting in about 20% transfection efficiency. To determine promoter reporter activities from numbers of fluorescent cells, two selection criteria were taken into account. First, the fraction of green-fluorescent cells among all red cells in a spot had to exceed 16% (transfection threshold). Second, the number of cells both green and red had to exceed the number of cells both green and red in the negative control spots (empty modified pZsGreen 1-1 spotted in 10 replicates) by three standard deviations (reporter activity threshold). A promoter was classified as active if both thresholds were exceeded in at least four out of six replicates on two different cell array slides. Thus, a binary promoter activity index (with 0 for inactive and 1 for active promoters) was generated for each promoter region under investigation.

Computational analyses

We used known position-weight matrices for TATA box, INR and DPE elements (20) together with the TransFac MATCH tool (21) for detection of common promoter motifs under default parameters. Genome-wide coordinates of CpG islands (22) were intersected with the coordinates of cloned promoters to identify CpG islands. In this, we required 500 bp immediately upstream of the transcription start site to overlap with at least 10% of the total sequence of a CpG island. Genome-wide coordinates of RNA polymerase IIA-bound regions in HEK293 (17) were intersected with the coordinates of cloned promoters to assess occupancy of the hypophosphorylated form of Pol IIA in promoter regions.

We retrieved associations of expressed sequence tag (EST) identifiers to UniGene cluster identifiers in 45 tissues generated for 5,799,931 human ESTs clustered into 116,190 UniGene clusters from the UniGene FTP site (Hs.profiles.gz for Homo sapiens build #207). EST expression profiles for these UniGene clusters were extracted from the 'Body Sites' category of the original file. The resulting EST set for 156 HSA21 genes consisted of 32,450 ESTs from 45 tissues. We then calculated for each gene with corresponding cloned promoter the number of different tissues where corresponding ESTs could be found.

For the set of promoter sequences that showed specific response patterns in our experiments, we searched for common transcription factors binding sites that might explain these responses. To score transcription factor binding, we used a physical affinity-based model described in previous publications (23,24), and matrices describing 610 vertebrate transcription factor binding preferences from TRANSFAC version 12.1 (21). For each binding matrix, we calculated the affinity of the matrix for each sequence, and then transformed these affinities into p-values as described before (23). These p-values represent the probability that the observed binding affinity is greater than would be expected from a random sequence from a human-promoter-based background model. The p-values for each sequence can then be combined using Fisher's method. Each binding matrix is then ranked according to its combined p-value, giving a natural ranking of the transcription factors that have the most enriched binding within the sequence set as a whole.

Enriched gene ontology (GO) terms were identified using the DAVID functional annotation tool (25). Entrez GeneIDs for 40 promoters activated by serum depletion and 28 promoters activated by Trichostatin A were compared to a background set of 126 inactive promoters within the GO category 'biological process'.

RESULTS

Cloning and reporter activity of HSA21 gene promoters in HEK293 cells

Promoter regions were defined relative to the most upstream annotated TSS of 231 genes on human chromosome 21 (see Methods for details). Primer pairs encompassing 2.5 kb of DNA sequence upstream of the TSS could be designed for 223 promoter fragments. Of these, 182 fragments were successfully amplified and cloned into a reporter vector upstream of the green fluorescent reporter gene FP506. Promoter coordinates and primer sequences are listed in Supplementary Table S1. HEK293 cells were co-transfected on cell arrays spotted with promoter reporter constructs and a normalization plasmid expressing red fluorescent protein HcRED. The transfection efficiency was estimated to be approximately 20% (see Methods). We measured the green and red fluorescence signals and used a cell number-based quantification approach for the determination of promoter activity. Reporter gene activities are listed in Supplementary Table S2. Stringent thresholds were set for co-transfection and the number of cells with reporter gene activity, ensuring a reliable scoring of activity. The mean number of co-transfected cells with reporter activity was 32.9±5.9 cells for active fragments and 6.2±4.8 cells for inactive fragments, whereas values for the negative controls were 1.7±2.0 cells. Figure 1A shows that active fragments could be clearly distinguished from inactive fragments by numbers of co-transfected cells with reporter gene activity on the transfected-cell arrays. Overall, 56 of 182 cloned 2.5 kb promoter regions were scored active in untreated HEK293 cells, whereas 126 remained silent.

Promoter reporter activities correlate with endogenous gene expression

We compared the promoter activities with the endogenous transcript levels in HEK293 cells reported from a transcriptome sequencing (RNA-seq) approach (17). Among 56 active promoter fragments, 50 corresponding genes were found expressed according to RNA-seq (Figure 1B and Figure 3A). We conclude that 89% of the active promoters score as true positives in the reporter assay. In contrast, only 37 of 126 inactive promoters are associated with expressed genes. The observed enrichment of expressed genes among active promoters is highly significant ($p=1.2\times10^{-14}$). To distinguish promoters of ubiquitously expressed genes from promoters of genes with a more restricted expression pattern, we made use of expression data compiled in UniGene EST clusters (26). We observed a strong correlation between promoter reporter activity in HEK293 cells and the number of tissues in which the associated gene is

transcribed (Figure 1C and Figure 3A). We found that 38 of 56 active promoters originate from genes with broad expression pattern *in vivo* (ESTs found in >25 different tissues). In contrast, 97 of the 126 inactive promoters belong to genes with a more restricted expression pattern. The enrichment of broadly expressed genes among active promoters is highly significant ($p=1.1\times10^{-8}$). Data on promoters and associated gene expression can be found in Supplementary Table S3.

Active promoters are enriched in RNA polymerase IIA-bound regions and activating core promoter elements

An overview of all data sets associated with the analyzed HSA21 promoters is shown in Figure 2. To investigate the influence of functional transcription start sites on promoter activity, we made use of previously published ChIP-seq data of hypophosphorylated RNA polymerase II polypeptide A (Pol IIA) used as a landmark of transcription initiation in HEK293 cells (17). The majority of active promoter fragments (35 of 56) contains or overlaps Pol IIA-bound regions (Figure 3A), whereas inactive promoter fragments are depleted in Pol IIA-bound regions (12 of 126 with Pol IIA occupancy). The observed enrichment of Pol IIA occupancy in active promoters is highly significant ($p=2.9\times10^{-13}$).

Core promoters are known to be associated with promoter-specific sequence elements controlling the initiation of transcription of downstream genes. For instance, CpG islands, the TATA box, initiator (INR) and downstream promoter elements (DPE) are functionally important, although their presence is not always required for promoter activity (9,10,27). We analyzed the occurrence of these four elements within the TSS near 500 bps of all 182 cloned HSA21 promoter fragments. Regarding CpG islands, we used a reference map with genome-wide coordinates of CpG islands (22) and found that 46% of the 182 cloned promoters (83 out of 182) overlap with a CpG island over a sequence length of at least 50 bps. In this, we observed a marked difference between active and silent promoters. Of the 56 active HSA21 promoters, 46 contain a GpC-island (Figure 3A). In contrast, only 37 of 126 silent promoters contain CpG islands. This enrichment of CpG islands in active fragments is highly significant ($p=2.3\times10^{-11}$). The TATA box, located 28-34 bp upstream of the TSS (28), is the best-known core promoter element. TATA boxes are often associated with strong tissue-specific promoters and result in clearly defined transcription start sites (10). We found TATA boxes in only 14 of all 182 cloned HSA21 fragments, which is less than the genome-wide occurrence in promoters reported before (29). TATA boxes were found slightly enriched in silent promoters (9% with TATA), as opposed to active promoters (5% with TATA). No trend was found for the INR element, which was present in 7% of active fragments (4 of 56) and in 6% of inactive

promoters (8 of 126). However, as shown in Figure 3A, DPE elements were found significantly enriched in half of the active promoters (28 of 56) as opposed to only in one-third of silent promoters (42 of 126; p=0.025). Coordinates of Pol IIA-bound regions, CpG islands and sequence elements in the cloned promoters can be found in Supplementary Table S4.

Regarding the co-occurrence of Pol IIA-bound regions and core promoter elements, we found three elements appearing together in a significant number of cases. Of the 47 promoters with Pol IIA occupancy, 45 contain a CpG island. Also, 24 of the Pol IIA-occupied promoters contain a DPE element. In line with this observation, we also noted that 40 of the 70 promoters with DPE element contain a CpG island, suggesting a functional connection between the three elements. Lastly, it is also notable that 53 inactive fragments do not overlap with any of the promoter elements or Pol IIA-bound regions analyzed here.

Different external stimuli modulate the activities of divergent sets of promoters

To assess the functionality of the cloned promoter fragments, we challenged the cells with external stimuli and monitored promoter activities after treatment with Trichostatin A (TSA) and after depletion of fetal calf serum (FCS) from the culture medium. The effects of TSA on cell function are complex, however, the expected effect of such a histone deacetylase inhibitor is the activation of transcription from repressed regions of the chromosomes (30). Indeed, TSA treatment activated 28 of the 126 previously inactive promoters, while only three of the 56 previously active promoters were silenced (see Figure 2). We analyzed the genes corresponding to these activated promoters regarding their expression patterns. As shown in Figure 3B, we found genes with broad expression (ESTs in >25 tissues) highly enriched among the TSA-activated promoters (15 of 28), while this fraction among the promoters that remained silent was much lower (14 of 98; $p=5\times10^{-5}$). Similarly, we observed significant enrichments of endogenously expressed genes and Pol IIA binding regions (Figure 3B), while no significant enrichment of TATA, INR and DPE elements among the activated promoters could be detected. Interestingly, the strongest observed enrichment concerned CpG islands, which were found present in 68% of TSA-activated promoters (Figure 3B). Regarding biological functions of the genes activated by TSA, no significant enrichment of any functional category could be detected (data not shown).

Serum depletion elicits stress responses and subsequent apoptosis through activation of several factors, such as NFκB and CREB (31-33). After depletion of serum, we found that no promoter was silenced. In contrast, 40 promoters were found activated (see Figure 2). Among the latter,

19 were also activated by TSA. A comparison of expression and sequence features of promoters activated by serum depletion (Figure 3C) to those of TSA-activated promoters (Figure 3B) revealed similarities as well as differences between both promoter sets. Similar to TSA treatment, but less pronounced, we found significant enrichment of broad expression patterns and CpG islands among promoters activated by serum depletion. In contrast, serum depletion-activated promoters are neither significantly enriched for genes with endogenous expression in HEK293 cells nor for RNA Pol IIA occupancy. Instead, a significant enrichment of DPE elements can be observed (Figure 3C). CpG and DPE elements appear together in 28% of promoters activated by serum depletion (11 of 40), but only in 8% of promoters that remained silent (7 of 86). Regarding biological functions of the genes activated by serum depletion, we found that 20% of the activated promoters correspond to genes associated with cellular responses to the environment (Supplementary Table S5). In contrast, only 8% of the promoters remaining inactive belong to this category.

Altogether, monitoring on transfected-cell arrays revealed that 56 promoters of 2,500 bp length drive reporter gene expression in HEK293 cells under normal growth conditions. Assays in the presence of different external stimuli showed that an additional 49 promoter fragments have the capacity to induce reporter gene expression. Regarding the remaining 77 silent fragments, only 17 are associated with genes expressed in HEK293 cells according to RNA-seq data. A closer inspection of these 17 inactive fragments, with integration of Pol IIA ChIP-seq and RNA-seq data, revealed that in four cases the core promoter was missed by 10-30 base pairs (promoters of C21orf19, C21orf90, HEMK2 and PFKL), and that in five cases an alternative TSS is employed for these genes in HEK293 cells (promoters of ABCG1, MRPS6, NCAM2, NRIP1 and PCBP3). Apart from these few examples, we can conclude that the majority of cloned HSA21 promoters recapitulate their function in living cells.

Truncation of promoters indicates the presence of distal regulatory elements

To investigate the influence of distal promoter regions on transcription, we cloned a subset of 62 truncated promoter fragments of ~500 base pairs, thus removing the distal ~2,000 bases. We found that 29 of the 62 short fragments were active in reporter assays under standard conditions (Figure 4). Interestingly, truncation of promoter length resulted in loss of activity for only three fragment (DSCR2, OLIG1 and SIM2). However, six promoters gained activity in their truncated form, while their longer version was inactive (MRPL39, RBM11, CHAF1B, HLCS, C21orf45 and SH3BGR), hinting at the presence of inhibitory regulatory regions in the distal ~2,000 bases.

Regarding the response of short fragments to treatments with TSA and depletion of serum, we found that 40 short promoters (66%) recapitulate, under all conditions, the activity patterns observed for the long fragments (Figure 4, lower part), while the other 21 behaved differently (Figure 4, upper part). For these, we observed two different possible results of truncation. First, fourteen truncated promoters could not be activated by treatments, while their longer counterparts were active or activatable, indicating the presence of activating *cis*-regulatory upstream elements. Second, seven truncated promoters showed to be active under more conditions than their longer counterparts (C21orf66, C21orf45, CHAF1B, MRPL39, HLCS, RBM11 and SH3BGR), hinting at inhibitory elements in the distal ~2,000 bp sequences.

Identification of enriched *cis*-regulatory elements among promoters responding to external stimuli

We were interested in the regulatory elements potentially contributing to the observed response patterns to external stimuli. We ranked affinities for 610 known vertebrate TF binding matrices (TRANSFAC database 12.1) in the entire sequences of the promoters activated by serum depletion or TSA treatment, and in the distal 2,000 bp of those promoters that lost their activation by stimuli after these regions had been removed by truncation. We found several TF binding matrices enriched in the sequences from each promoter category. For each enriched matrix, we analyzed if the corresponding TF is expressed in HEK293 cells according to the available RNA-seq data set. The top four enriched matrices of TFs expressed in HEK293 are listed in Table 1. All binding sites detected for these TFs in the analyzed promoter fragments are listed in Supplementary Results. We identified several connections between enriched TF binding matrices and the biological stimuli used here to modulate promoter activities (see references in Table 1). Serum responses has been reported before to influence the activities of USF1, NFκB, MYC and ETS1. Concerning TSA responses and associated histone deacetylase inhibition, we found reports describing sensitivity to TSA treatment for MAFG, AP1 (FOS/JUN), p53 and OCT1. Thus, four of seven TFs with enriched matrices in serum-sensitive promoters and four of eight TFs with enriched matrices in TSA-responsive promoters have been previously implicated in corresponding signal transduction pathways.

DISCUSSION

Using a transfected-cell array format, we were able to monitor the activities of 182 cloned promoters corresponding to ~80% of all human chromosome 21 genes in HEK293 cells. Compared to previous large-scale studies, where the length of promoters was restricted to a 1,000 or less base pairs of DNA (6-8), we aimed at more comprehensive coverage of upstream regulatory elements by cloning 2.5 kilo bases, so that additional potentially relevant regulatory elements could be covered by our analysis, and administered treatments with external stimuli to identify the regulatory nature of these elements. In this, the cell array format proved as reliable and cost-efficient alternative to conventional reporter gene assays in microtitre plates.

Promoter reporter activities recapitulate endogenous gene expression states

In order to assess promoter contribution to endogenous gene expression, we compared transcript levels for chromosome 21 genes in HEK293 cells with the corresponding promoter reporter activities. For the promoters active in the reporter assays, this comparison revealed a high level of overlap, with 89% of the corresponding genes expressed, which is in agreement with previous observations (7,11). Nevertheless, six promoters were found active without detection of corresponding transcripts, namely those of C21orf13, C21orf115, DSCR4, DSCR8, KRTAP21-2 and RSPH1. The discrepancy observed here might indicate the absence of inhibitory elements residing further upstream or downstream, which were not included in the promoter reporter constructs, or presence of inhibitory chromatin structures or DNA methylation in the genomic context of these genes. Regarding the promoters that were inactive under standard conditions, only 29% of the corresponding genes are expressed in HEK293 cells, and the majority of these promoters was activated by treatment of the cells with TSA or serum depletion. We conclude that the corresponding cloned fragments do not contain all regulatory elements, especially enhancers, which are necessary to reach the strength of the endogenous promoters in the context of the natural chromatin environment. The remaining 17 inactive promoters could not be activated by treatment conditions. We found that alternative transcription start sites are employed in HEK293 cells for five genes, while key elements required for transcription remained outside of the cloned fragments for four genes. The incorporation of RNA-seq and Pol IIA ChIP-seq data into the annotation process will significantly improve future promoter annotations.

Core promoter elements and Pol IIA occupancy strongly determine promoter activities

We have examined the correlation of various sequence and functional features involved in pre-initiation complex assembly with promoter activities. The finding of significant enrichments in CpG and DPE, but not TATA or INR elements in active promoters confirms previous observations (7,34). As in primary fibroblasts (35), CpG islands are present within more than 80% of the promoters active in HEK293 cells. Moreover, the strong correlation between gene expression levels and promoter activity in our reporter assays concerned mostly genes containing CpG islands, resulting in a wide tissue-representation of corresponding transcripts, indicating ubiquitous expression patterns. One the one hand, RNA polymerase IIA-binding indicates Pol II stalling at genes poised for activation (36,37), and on the other hand, active transcription start sites (17,38). As expected, we found the presence of Pol IIA-bound regions strongly associated with the activity of promoter fragments. The marked correlation of active promoters with both Pol IIA occupancy and CpG islands is not surprising, as CpG islands are known to be strongly enriched in regions with Pol II stalling (39). Conversely, a lack of Pol IIA occupancy was characteristic for inactive promoters of genes expressed in HEK293 and for inactive promoters of CpG-associated genes.

Promoters can be classified into subsets according to their responses to external stimuli

We have modulated promoter activities by treatment with TSA and by depletion of serum. The expected effect of a specific inhibitor of mammalian class I and II histone deacetylase enzymes (30), such as TSA, is activation of transcription from repressed chromosomal regions through chromatin remodeling (40). Even though transiently transfected plasmids, as used in this study, are not entirely subject to the same regulatory mechanisms that affect native chromatin, it has been shown that chromatin structures can be formed on plasmid DNA, although transfected DNA is generally more accessible than cellular chromatin (41). Subsequently, it should be possible to reverse histone deacetylase-dependent silencing mechanisms through activation by TSA (42,43). In fact, we found a considerable number of promoters activated upon treatment with TSA. Interestingly, we observed a striking enrichment of CpG islands, as 68% of TSA-activated promoters contain or overlap such a region, a finding that indicates another property of TSA, namely inhibition of DNA methyltransferase DNMT1 (44). Methylation of CpG islands is correlated with gene silencing, and inhibition of DNMT1 can enhance early expression of those genes that are silenced through CpG methylation, Subsequently, TSA-induced CpG demethylation can follow early transcription and fully activate gene expression of CpG-associated TSA-activated genes (45). On the basis of these findings, we assume that the

promoter-reporter constructs are sensible to endogenous histone deacetylation or, more probably, DNA methylation-mediated silencing (44,46,47).

Depletion of serum from the cell culture medium activates cell type-specific responses affecting cell cycle regulation, cell growth, differentiation and apoptosis (32,48,49). To address the question whether promoters activated by serum depletion belong to genes with functional similarities, we made use of gene ontology annotations (25). Of the 40 promoters activated by serum depletion, eight genes are annotated in the biological process category “response to stimulus” and related subcategories. The observed enrichment is not significant due to small sample size, however, it is notable as it underlines the finding that cloned promoter fragments can integrate endogenous signaling pathways into reporter gene expression. We conclude that the observed effect of serum depletion in our reporter assays has a biological correlate in terms of signal transduction *in vivo*.

Evidence for positive and negative response elements within distal promoter regions

Data obtained from assaying a set of promoters in long and truncated forms and under different conditions revealed that two thirds of truncated promoters reproduce the activity patters of long promoter fragments, while the other third behave differently. The finding that only three short fragment lost activity compared to its long counterpart implies that in general, ~500 base pair fragments, spanning core and proximal promoter region, are sufficient to drive gene expression. Six promoters were active in their short and inactive in their long version, with three of them highly expressed in HEK293 cells, suggesting the presence of inhibitory *cis*-regulatory elements within -500 to -2500 bps, but also the presence of strong genomic enhancers outside of the cloned regions. Regarding overall activity changes through external stimuli, 52 long fragments changed activity (29% of 182 tested), while only three short fragments were responsive to external stimuli (5% of 62 tested). This significant difference is evidence for the presence of inhibitory, but more importantly, for the presence of activating response elements within the cloned distal promoter regions. It has been reported that negative regulatory elements localize within the region 1,000 to 500 base pairs upstream of the TSS (7,50). We find that activating distal promoter elements outnumber negative elements in the chromosome-wide set of promoters analyzed here.

Identification of *cis*-regulatory elements with involvement in the responses to serum depletion and Trichostatin A

In order to identify *cis*-regulatory elements that can have an impact on the observed promoter response patterns, we have analyzed enriched transcription factor binding matrices in the sequences of the promoters activated by serum depletion and Trichostatin A. Similarly, we have scanned the upstream regions of those promoters that lose activation by these stimuli upon truncation.

The binding motif analyses identified seven candidate TFs with potential influence on promoter activity following serum depletion. Four of these TFs have been previously reported to be activated by this treatment, namely USF1, NFκB, MYC and ETS1. Serum starvation has been shown to enhance USF1 expression and the efficiency of USF1 binding to and upregulation of its target gene lipocalin-type PGD synthase in a human brain-derived cells (51). Somewhat similar, NFκB has been found potently activated upon serum starvation in HEK293 cells, leading to apoptosis (31). Responses to serum deprivation also involve the MYC protein. The signaling mechanism of MYC-induced apoptosis in human hepatoma cells under growth factor-deprived conditions was found dependent on FOS, with an ATF2-responsive element conferring the MYC-induced expression of FOS (52). For the promoter of ATF3, another transcription factor determining cell fate under stress conditions, it has been shown that the MYC complex plays a role in mediating the serum response of ATF3 gene expression in rat fibroblasts (53). Finally, also Ets domain-containing TFs, such as ELF2 and ETS1 identified here, are implicated in the response to serum. In a human endothelial cell line, transcriptional activation in response to serum was found to be regulated by a functional Ets motif in the promoter of CD13, where ETS2 and ETS1 can bind and regulate the CD13 promoter activity (54). The finding that we identified four TFs involved in serum response in our reporter assays confirms the reliability of the study concerning the integration of endogenous signaling pathways in promoter reporter gene activities.

Taking a closer look at the eight enriched TF binding matrices in the promoters activated by the histone deacetylase inhibitor Trichostatin A, we found that TSA treatment has been previously described to affect the functions of four of the associated TFs, namely MAFG, p53, OCT1 and AP1. TSA treatment has been reported to abolish MAFG-mediated repression of gene expression via Maf recognition elements in reporter gene assays performed in HEK293 cells (55), which is in line with our observations of the activating effect of TSA on promoters. The tumor suppressor p53 was described to follow a TSA-dependent mode of action, with TSA

causing p53 to induce apoptosis in a human colorectal cell line (56), while in a prostate cancer cell line, TSA stabilized the acetylation of p53, inducing cell cycle arrest, but not apoptosis (57). Alternatively, TSA can also induce gene expression via OCT1, another TF identified in our study, without the need for functional p53, as shown in a human osteosarcoma cell line (58). Lastly, concerning AP1 found enriched among TSA-induced promoters, it has been reported that upon TSA treatment, this activator complex binds to an AP1 recognition site in the osteopontin gene promoter and activates expression of this gene in a mouse mesenchymal cell line (59). AP1 is a variable complex composed of members of the JUN, FOS and CREB/ATF families. FOS is both a part of the TSA-responsive AP1 complex and a mediator of the MYC-induced apoptotic signaling following serum starvation, as described above (52). Here, we see a possible connection between promoter activation by serum depletion and activation by TSA treatment, which may explain the finding that 19 promoters were activated by both types of treatment in our reporter assays.

Taken together, our findings indicate that the observed promoter response patterns, depending on the presence or absence of the specific *cis*-regulatory elements, recapitulate the integration of endogenous signaling pathways into reporter gene expression. The list we provide here of chromosome 21 promoters with upstream positive and negative elements, along with possible involvements of various transcription factors in the promoter response patterns, constitutes a valuable resource for researches interested in the regulation of the corresponding genes.

CONCLUSION

Using reporter gene assays on transfected-cell arrays, we are able to draw a general picture of HSA21 promoter function in HEK293 cells, correlating promoter activities to the presence of core promoter elements, promoter occupancy by RNA polymerase II, and endogenous transcript levels. The identified correlations show that promoter studies greatly profit from incorporation of data sets generated by massively parallel sequencing technology. Proximal promoter regions were found generally sufficient to drive gene expression under standard cell culture conditions. However, extended promoter regions are more likely to integrate endogenous signaling pathways into reporter gene expression than proximal promoters. This finding is further underlined by the identification of genes involved in responses to stimuli found activated by serum depletion, and hints towards the presence of positive and negative *cis*-acting response elements in distal promoter regions. The analysis of promoter fragments of in different lengths allowed for identification of enrichment of binding sites for corresponding transcription factors that can integrate signals from administered stimuli into gene expression. The collection of cloned HSA21 promoters can be used in further studies on promoter activities in different cell lines and in combination with overexpression or knockdown of transcription factors, allowing to study transcriptional regulation in parallel and in more detail on the scale of a whole chromosome.

FUNDING

This work was supported by the Max Planck Society and the Federal Ministry of Education and Research (BMBF) in the framework of the National Genome Research Network (NGFN2, MP-DNA ‘TP Promotor-Ressourcen’, PDN-S02T14) [grant number 01GR0414].

ACKNOWLEDGEMENTS

We thank Sabine Thamm and Irina Girnus for assistance in preparation of plasmid constructs.

FIGURE LEGENDS

Figure 1. Active promoters can be reliably distinguished from inactive promoters using reporter gene assays on transfected-cell arrays. Promoter reporter activities strongly correlate with endogenous HEK293 gene expression levels and with breadth of expression according to EST data. (**A**) A promoter was scored active if a cell number threshold for reporter activity (three standard deviations over the mean of the negative controls) and a co-transfection threshold for the two plasmids (16%) were both exceeded (see Methods for details). The plot of the mean numbers and standard deviations of co-transfected cells with reporter activity shows the significant differences between active and inactive promoter fragments. (**B**) Active promoter fragments are mostly derived from genes with endogenous expression in HEK293, as detected by RNA-seq (17). In contrast, most inactive promoters correspond to genes without expression in HEK293. (**C**) The majority of inactive promoters is derived from genes with restricted expression patterns, whereas most active promoters correspond to genes with broader expression patterns (ESTs found in >25 tissues). For each gene, the number of different human tissues with corresponding ESTs according to the UniGene database was determined and plotted against active and silent promoter fractions.

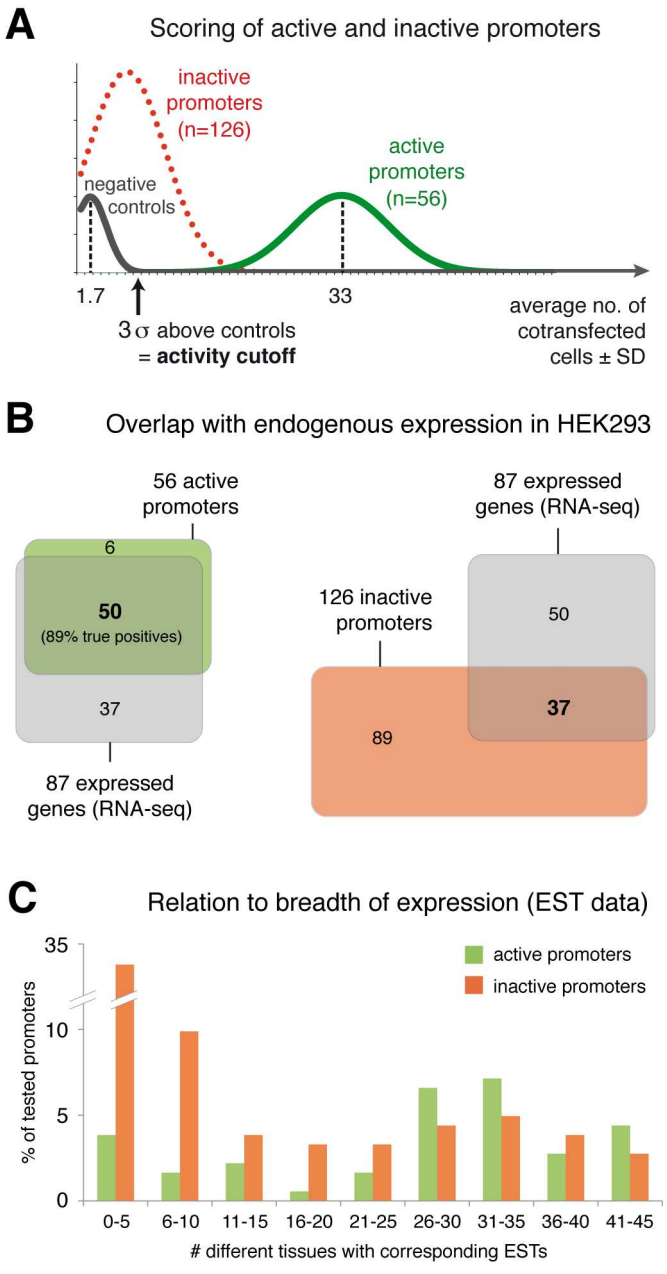
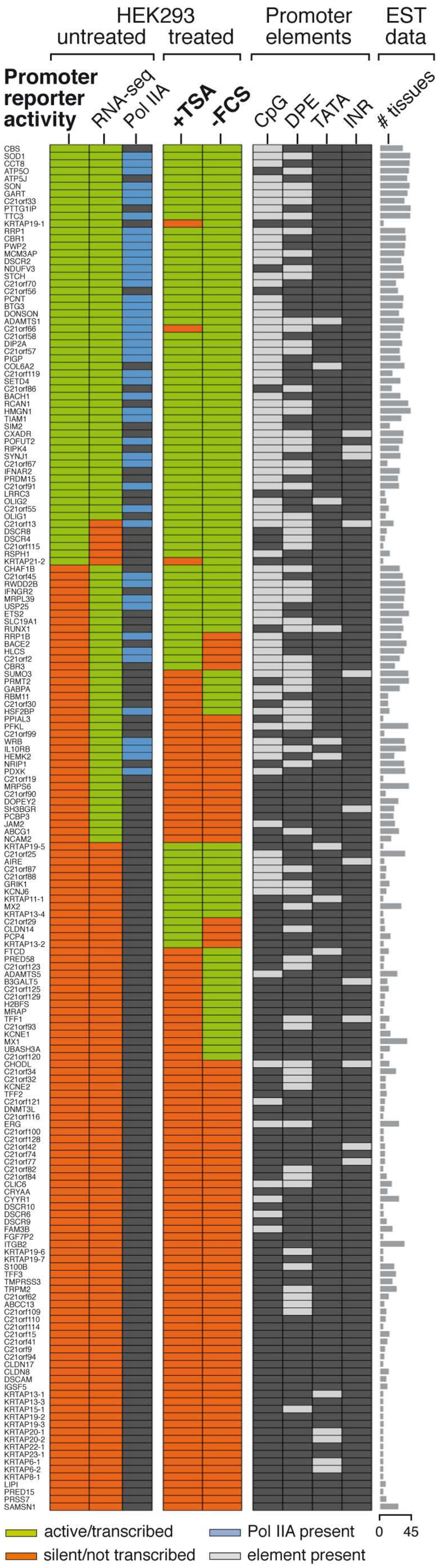

A
Scoring of active and inactive promoters
inactive promoters (n=126)
active promoters (n=56)
negative controls
1.7
33
3σ above controls = activity cutoff
average no. of cotransfected cells ± SD
B
Overlap with endogenous expression in HEK293
56 active promoters
6
50
(89% true positives)
37
87 expressed genes (RNA-seq)
87 expressed genes (RNA-seq)
126 inactive promoters
50
37
89
C
Relation to breadth of expression (EST data)
active promoters
inactive promoters
% of tested promoters
35
10
5
0
0-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
different tissues with corresponding ESTs

Figure 2. Overview of promoter reporter activities, gene expression data and core promoter elements for 182 tested 2.5 kb promoters. Each row in the panels represents one tested promoter, with gene symbols indicated on the left. The panel on the left show promoter reporter gene activities for untreated HEK293 cells measured on transfected-cell arrays, corresponding endogenous gene expression (RNA-seq) and RNA polymerase IIA occupancy (Pol IIA). The central panel show promoter reporter activities after treatment on transfected-cell arrays with Trichostatin A (+TSA) and after depletion of fetal calf serum (-FCS), The panel on the right show the presence or absence of core promoter elements residing in the cloned fragments. Promoter reporter activity: active promoters driving reporter gene expression are represented by green boxes, inactive promoters by red boxes. RNA-seq: transcriptome sequencing data from HEK293 cells (17); green – transcripts detected in HEK293; red – no transcripts detected or uncertain; Pol IIA: Blue boxes indicate RNA Polymerase IIA-bound regions in promoters identified by chromatin immunoprecipitation (ChIP-seq) from HEK293 cells (17). Promoter elements: Light gray boxes indicate the presence of CpG islands, downstream promoter elements (DPE), TATA boxes and initiator elements (INR). EST data: The number of different tissues (out of 45) in which corresponding ESTs are present in the UniGene database is represented here as the length of a horizontal bar. All rows are sorted (i) by promoter reporter activity in untreated conditions, (ii) by endogenous expression in HEK293 cells, and (iii) by promoter activities in treated conditions.



HEK293
untreated
treated
Promoter elements
EST data
Promoter reporter activity
RNA-seq
Pol IIA
+TSA
-FCS
CpG
DPE
TATA
INR
tissues
active/transcribed
silent/not transcribed
Pol IIA present
element present
0
45

Figure 3. Gene expression features and sequence elements associated with active and silent promoter fragments. (**A**) The 56 promoters active under untreated conditions are significantly enriched for genes with expression in >25 different tissues (broadly expressed), for genes endogenously expressed in HEK293 cells (HEK expressed), for occupancy of RNA polymerase IIA in the promoter fragment (Pol IIA binding), for presence of CpG islands and for downstream promoter elements (DPE). Dark grey bars represent the fraction of 56 active promoters associated with the indicated feature, light grey bars represent the fraction of 126 promoters inactive in untreated conditions. (**B**) Treatment of cells with Trichostatin A activates 28 promoters that are significantly enriched for broadly expressed genes, expression in HEK293, Pol IIA binding and CpG islands. (**C**) Depletion of serum from transfected cells activates 40 promoters that are significantly enriched for broadly expressed genes, CpG islands and DPE elements. Significance levels of enrichments were calculated by the hypergeometric test and are indicated as ***(p<0.001), **(p<0.01), *(p<0.05) and ns (not significant).

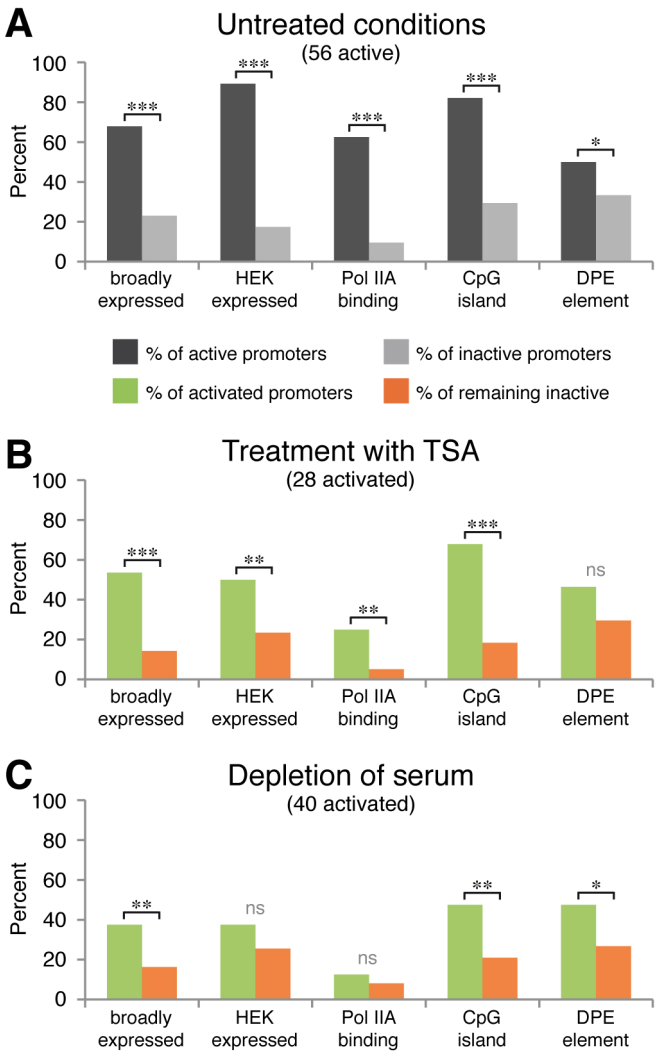


Figure 4. Truncation of promoter fragments can result in loss of responsiveness to external stimuli. The panels show an overview of reporter activities of 62 promoters that were assayed as both long (2.5 kb) and truncated fragments (~500 bp). Each row represents one tested pair of long and short promoter, with gene symbols indicated on the left. Active promoters are represented by green boxes, inactive promoters by red boxes. Reporter assays were carried out under standard growth conditions (untreated), after treatment of cells with Trichostatin A (+TSA) and after depletion of fetal calf serum (-FCS). Promoters are sorted by the result of truncation, which is either loss of response to external stimuli (21 promoters, upper parts) or no change in the response to stimuli (41 promoters, lower part). The presence of upstream *cis*-regulatory elements in distal promoter regions (-2,500 to -500 bp) can be inferred from the observed results of truncation, namely activating upstream elements (14 promoters) or inhibitory elements (7 promoters).

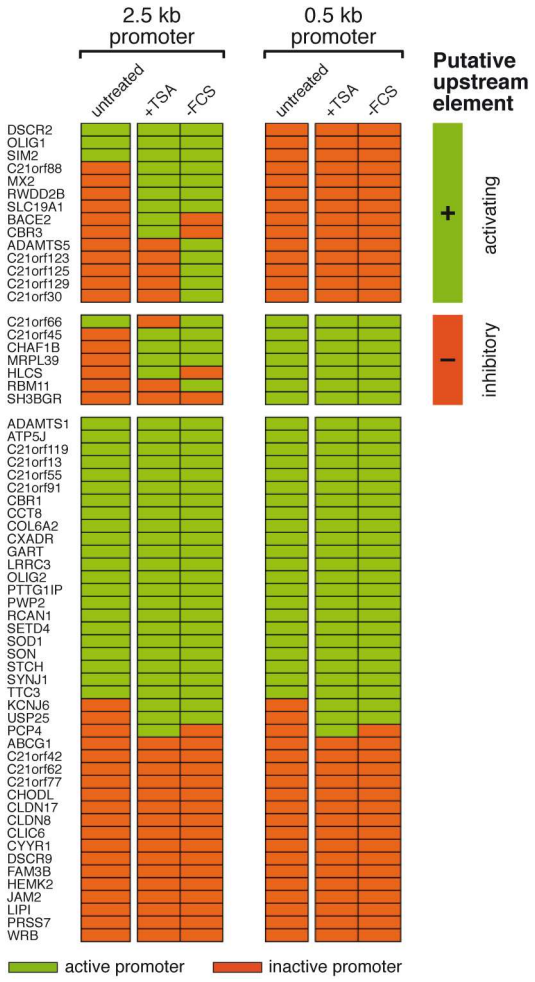


Table 1. Top-enriched transcription factor binding matrices among promoters responding to external stimuli

Promoter response category (no. of promoters; analyzed regions)	Enriched TF binding matrix	Combined p-value	Transcription factor	References to stimulus
Activation by serum depletion	V$USF_C	0.000104	USF1	(51)
(n=40; full 2.5 kb regions)	V$OSF2_Q6	0.000155	RUNX2	–
	V$NFKAPPAB_01	0.00279	NFκB	(31)
	V$MYC_Q2	0.00355	MYC	(52,53)
Loss of activation by serum depletion	V$NFKAPPAB_01	0.00287	NFκB	(31)
after truncation to 500 bp	V$NERF_Q2	0.007	ELF2	–
(n=9; 2 kb distal regions)	V$ETS1_B	0.00938	ETS1	(54)
	V$MAF_Q6	0.00979	MAF	–
Activation by Trichostatin A	V$POU3F2_01	0.00127	OCT7	–
(n=28; full 2.5 kb regions)	V$TCF11MAFG_01	0.00355	MAFG	(55)
	V$AP1_Q2	0.00372	FOS/JUN	(59)
	V$MEF2_01	0.00432	MYEF2	–
Loss of activation by Trichostatin A	V$P53_01	0.015	p53	(56,57)
after truncation to 500 bp	V$OCT1_01	0.0258	OCT1	(58)
(n=6; 2 kb distal regions)	V$MAF_Q6	0.0345	MAF	–
	V$TEF1_Q6	0.0396	TEAD1	–

For each promoter response category, the top four enriched non-redundant TRANSFAC binding matrices are listed for TFs with endogenous gene expression in HEK293 cells according to transcriptome sequencing data. P-values for all individual sequences in a set were combined by Fisher's method, allowing for detection of TF binding that is enriched across the entire sequence set.

Supplementary Table S5. Top group of enriched biological processes among promoters activated by serum depletion

Gene symbol	Gene name	Biological processes
MX1	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78	**S**, B, M, O, D, V
MX2	myxovirus (influenza virus) resistance 2	**S**, B, M, O, D, V
H2BFS	H2B histone family, member S	**S**, B, M, O, D
IFNGR2	interferon gamma receptor 2 (interferon gamma transducer 1)	**S**, B, M, O, V
TFF1	trefoil factor 1 (estrogen-inducible sequence expressed in breast cancer)	**S**, D
AIRE	autoimmune regulator	**S**
RUNX1	runt-related transcription factor 1	**S**
UBASH3A	ubiquitin associated and SH3 domain containing, A	**S**

Biological processes (gene ontology annotations) are abbreviated: S – response to stimulus (GO:0050896); B – response to biotic stimulus (GO:0009607); M – multi-organism process (GO:0051704); O – response to other organism (GO:0051707); D – defense response (GO:0006952); V – response to virus (GO:0009615).

REFERENCES

1. Asturias, F.J. (2004) Another piece in the transcription initiation puzzle. *Nat Struct Mol Biol*, **11**, 1031-1033.
2. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, **20**, 1377-1419.
3. Juven-Gershon, T., Hsu, J.Y., Theisen, J.W. and Kadonaga, J.T. (2008) The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol*, **20**, 253-259.
4. Novina, C.D. and Roy, A.L. (1996) Core promoters and transcriptional control. *Trends Genet*, **12**, 351-355.
5. Miller, W., Makova, K.D., Nekrutenko, A. and Hardison, R.C. (2004) Comparative genomics. *Annu Rev Genomics Hum Genet*, **5**, 15-56.
6. Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R.B., Batzoglou, S. and Myers, R.M. (2003) Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res*, **13**, 1765-1774.
7. Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L. and Myers, R.M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res*, **16**, 1-10.
8. Trinklein, N.D., Karaoz, U., Wu, J., Halees, A., Force Aldred, S., Collins, P.J., Zheng, D., Zhang, Z.D., Gerstein, M.B., Snyder, M. *et al.* (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res*, **17**, 720-731.
9. Muller, F., Demeny, M.A. and Tora, L. (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J Biol Chem*, **282**, 14685-14689.
10. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet*, **8**, 424-436.
11. Cheng, X., Guerasimova, A., Manke, T., Rosenstiel, P., Haas, S., Warnatz, H.J., Querfurth, R., Nietfeld, W., Vanhecke, D., Lehrach, H. *et al.* (2009) Screening of human gene promoter activities using transfected-cell arrays. *Gene*.

12. Baghdoyan, S., Roupioz, Y., Pitaval, A., Castel, D., Khomyakova, E., Papine, A., Soussaline, F. and Gidrol, X. (2004) Quantitative analysis of highly parallel transfection in cell microarrays. *Nucleic Acids Res*, **32**, e77.
13. Fiebitz, A., Nyarsik, L., Haendler, B., Hu, Y.H., Wagner, F., Thamm, S., Lehrach, H., Janitz, M. and Vanhecke, D. (2008) High-throughput mammalian two-hybrid screening for protein-protein interactions using transfected cell arrays. *BMC Genomics*, **9**, 68.
14. Hu, Y.H., Warnatz, H.J., Vanhecke, D., Wagner, F., Fiebitz, A., Thamm, S., Kahlem, P., Lehrach, H., Yaspo, M.L. and Janitz, M. (2006) Cell array-based intracellular localization screening reveals novel functional features of human chromosome 21 proteins. *BMC Genomics*, **7**, 155.
15. Vanhecke, D. and Janitz, M. (2005) Functional genomics using high-throughput RNA interference. *Drug Discov Today*, **10**, 205-212.
16. Ziauddin, J. and Sabatini, D.M. (2001) Microarrays of cells expressing defined cDNAs. *Nature*, **411**, 107-110.
17. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956-960.
18. Haas, S.A., Hild, M., Wright, A.P., Hain, T., Talibi, D. and Vingron, M. (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res*, **31**, 5576-5581.
19. Ralser, M., Querfurth, R., Warnatz, H.J., Lehrach, H., Yaspo, M.L. and Krobitsch, S. (2006) An efficient and economic enhancer mix for PCR. *Biochem Biophys Res Commun*, **347**, 747-751.
20. Jin, V.X., Singer, G.A., Agosto-Perez, F.J., Liyanarachchi, S. and Davuluri, R.V. (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics*, **7**, 114.
21. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, **29**, 281-283.
22. Bock, C., Walter, J., Paulsen, M. and Lengauer, T. (2007) CpG island mapping by epigenome prediction. *PLoS Comput Biol*, **3**, e110.
23. Manke, T., Roider, H.G. and Vingron, M. (2008) Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, **4**, e1000039.
24. Roider, H.G., Kanhere, A., Manke, T. and Vingron, M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134-141.

25. Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, **4**, P3.
26. Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, **31**, 28-33.
27. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J Mol Biol*, **196**, 261-282.
28. Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol*, **7**, R78.
29. Thomas, M.C. and Chiang, C.M. (2006) The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol*, **41**, 105-178.
30. Khan, N., Jeffers, M., Kumar, S., Hackett, C., Boldog, F., Khramtsov, N., Qian, X., Mills, E., Berghs, S.C., Carey, N. *et al.* (2008) Determination of the class and isoform selectivity of small-molecule histone deacetylase inhibitors. *Biochem J*, **409**, 581-589.
31. Grimm, S., Bauer, M.K., Baeuerle, P.A. and Schulze-Osthoff, K. (1996) Bcl-2 down-regulates the activity of transcription factor NF-kappaB induced upon apoptosis. *J Cell Biol*, **134**, 13-23.
32. Leicht, M., Briest, W., Holzl, A. and Zimmer, H.G. (2001) Serum depletion induces cell loss of rat cardiac fibroblasts and increased expression of extracellular matrix proteins in surviving cells. *Cardiovasc Res*, **52**, 429-437.
33. Li, G., Yang, Q., Krishnan, S., Alexander, E.A., Borkan, S.C. and Schwartz, J.H. (2006) A novel cellular survival factor--the B2 subunit of vacuolar H+-ATPase inhibits apoptosis. *Cell Death Differ*, **13**, 2109-2117.
34. Kim, T.H., Barrera, L.O., Qu, C., Van Calcar, S., Trinklein, N.D., Cooper, S.J., Luna, R.M., Glass, C.K., Rosenfeld, M.G., Myers, R.M. *et al.* (2005) Direct isolation and identification of promoters in the human genome. *Genome Res*, **15**, 830-839.
35. Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876-880.
36. Koch, F., Jourquin, F., Ferrier, P. and Andrau, J.C. (2008) Genome-wide RNA polymerase II: not genes only! *Trends Biochem Sci*, **33**, 265-273.
37. Wu, J.Q. and Snyder, M. (2008) RNA polymerase II stalling: loading at the start prepares genes for a sprint. *Genome Biol*, **9**, 220.

38. Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845-1848.
39. Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S. and Levine, M.S. (2008) Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A*, **105**, 7762-7767.
40. Taddei, A., Roche, D., Bickmore, W.A. and Almouzni, G. (2005) The effects of histone deacetylase inhibitors on heterochromatin: implications for anticancer therapy? *EMBO Rep*, **6**, 520-524.
41. Smith, C.L. and Hager, G.L. (1997) Transcriptional regulation of mammalian genes in vivo. A tale of two templates. *J Biol Chem*, **272**, 27493-27496.
42. Huang, W., Zhao, S., Ammanamanchi, S., Brattain, M., Venkatasubbarao, K. and Freeman, J.W. (2005) Trichostatin A induces transforming growth factor beta type II receptor promoter activity and acetylation of Sp1 by recruitment of PCAF/p300 to a Sp1.NF-Y complex. *J Biol Chem*, **280**, 10047-10054.
43. Yokota, T., Matsuzaki, Y., Miyazawa, K., Zindy, F., Roussel, M.F. and Sakai, T. (2004) Histone deacetylase inhibitors activate INK4d gene through Sp1 site in its promoter. *Oncogene*, **23**, 5340-5349.
44. Januchowski, R., Dabrowski, M., Ofori, H. and Jagodzinski, P.P. (2007) Trichostatin A down-regulate DNA methyltransferase 1 in Jurkat T cells. *Cancer Lett*, **246**, 313-317.
45. D'Alessio, A.C., Weaver, I.C. and Szyf, M. (2007) Acetylation-induced transcription is required for active DNA demethylation in methylation-silenced genes. *Mol Cell Biol*, **27**, 7462-7474.
46. Klan, N., Seuter, S., Schnur, N., Jung, M. and Steinhilber, D. (2003) Trichostatin A and structurally related histone deacetylase inhibitors induce 5-lipoxygenase promoter activity. *Biol Chem*, **384**, 777-785.
47. Lande-Diner, L., Zhang, J., Ben-Porath, I., Amariglio, N., Keshet, I., Hecht, M., Azuara, V., Fisher, A.G., Rechavi, G. and Cedar, H. (2007) Role of DNA methylation in stable gene repression. *J Biol Chem*, **282**, 12194-12200.
48. Cooper, S.J., Trinklein, N.D., Nguyen, L. and Myers, R.M. (2007) Serum response factor binding sites differ in three human cell types. *Genome Res*, **17**, 136-144.
49. Leicht, M., Marx, G., Karbach, D., Gekle, M., Kohler, T. and Zimmer, H.G. (2003) Mechanism of cell death of rat cardiac fibroblasts induced by serum depletion. *Mol Cell Biochem*, **251**, 119-126.
50. Negi, S., Singh, S.K., Pati, N., Handa, V., Chauhan, R. and Pati, U. (2004) A proximal tissue-specific module and a distal negative regulatory module control apolipoprotein(a) gene transcription. *Biochem J*, **379**, 151-159.

51. Fujimori, K., Aritake, K. and Urade, Y. (2008) Enhancement of prostaglandin D(2) production through cyclooxygenase-2 and lipocalin-type prostaglandin D synthase by upstream stimulatory factor 1 in human brain-derived TE671 cells under serum starvation. *Gene*, **426**, 72-80.

52. Kalra, N. and Kumar, V. (2004) c-Fos is a mediator of the c-myc-induced apoptotic signaling in serum-deprived hepatoma cells via the p38 mitogen-activated protein kinase pathway. *J Biol Chem*, **279**, 25313-25319.

53. Tamura, K., Hua, B., Adachi, S., Guney, I., Kawauchi, J., Morioka, M., Tamamori-Adachi, M., Tanaka, Y., Nakabeppu, Y., Sunamori, M. *et al.* (2005) Stress response gene ATF3 is a target of c-myc in serum-induced cell proliferation. *EMBO J*, **24**, 2590-2601.

54. Petrovic, N., Bhagwat, S.V., Ratzan, W.J., Ostrowski, M.C. and Shapiro, L.H. (2003) CD13/APN transcription is induced by RAS/MAPK-mediated phosphorylation of Ets-2 in activated endothelial cells. *J Biol Chem*, **278**, 49358-49368.

55. Motohashi, H., Katsuoka, F., Miyoshi, C., Uchimura, Y., Saitoh, H., Francastel, C., Engel, J.D. and Yamamoto, M. (2006) MafG sumoylation is required for active transcriptional repression. *Mol Cell Biol*, **26**, 4652-4663.

56. Habold, C., Poehlmann, A., Bajbouj, K., Hartig, R., Korkmaz, K.S., Roessner, A. and Schneider-Stock, R. (2008) Trichostatin A causes p53 to switch oxidative-damaged colorectal cancer cells from cell cycle arrest into apoptosis. *J Cell Mol Med*, **12**, 607-621.

57. Roy, S., Packman, K., Jeffrey, R. and Tenniswood, M. (2005) Histone deacetylase inhibitors differentially stabilize acetylated p53 and induce cell cycle arrest or apoptosis in prostate cancer cells. *Cell Death Differ*, **12**, 482-491.

58. Hirose, T., Sowa, Y., Takahashi, S., Saito, S., Yasuda, C., Shindo, N., Furuichi, K. and Sakai, T. (2003) p53-independent induction of Gadd45 by histone deacetylase inhibitor: coordinate regulation by transcription factors Oct-1 and NF-Y. *Oncogene*, **22**, 7762-7773.

59. Sakata, R., Minami, S., Sowa, Y., Yoshida, M. and Tamaki, T. (2004) Trichostatin A activates the osteopontin gene promoter through AP1 site. *Biochem Biophys Res Commun*, **315**, 959-963.

6.2. Supplemental material

OVERVIEW Chromosome 21 Promoter Cloning, 2.5 kb Fragments

UCSC Assembly hg18

Mean size: 2477
Min size: 2352
Max size: 2550

HGNC_Symbol	Cloned	Amplicon	Amplicon_Location	Strand	Ensembl_GeneID	Ensembl_TranscriptID	Primer_Sequence_fwd	Primer_Sequence_rev
ABCC13	yes	2477	chr21:14565550-14568026	+	ENSG00000155288	ENST00000318742	AAATACCACTTGTGCCTATGTTGA	GGCCTGGAGGAGGCAAGG
ABCG1	yes	2488	chr21:42506845-42509332	+	ENSG00000160179	ENST00000347800	AGCCCTGGGGGCCTAAGT	CGGAAGGTGAAAGGCATGTC
ADAMTS1	yes	2485	chr21:27139564-27142048	-	ENSG00000154734	ENST00000284984	TAGACTGAAATACATTACATATGTCTTT	ATAGCCCCTGGCTCGGAG
ADAMTS5	yes	2466	chr21:27260677-27263142	-	ENSG00000154736	ENST00000284987	AAAGCAAGGACCACAAAAGAGAA	CACACTTGCTTGCAGGATTGA
AIRE	yes	2423	chr21:44527749-44530171	+	ENSG00000160224	ENST00000291582	TTCCCAGGAGGCCATATCTCA	CGCGCCGCTGTTATACCC
ATP5J	yes	2488	chr21:26029811-26032298	-	ENSG00000154723	ENST00000284971	GAGAAAAGTATATCAAGCATACTCAT	CTCGTTCCGGGGCCTTTTC
ATP5O	yes	2465	chr21:34210007-34212471	-	ENSG00000159186	ENST00000290299	GCAGCCATTAAAATGATACTTAGAA	TGGCGTAACCGCTAGGTTC
B3GALT5	yes	2550	chr21:39951048-39953597	+	ENSG00000183778	ENST00000343118	GAGGAGATTGGCTGTGAGTC	TTCTGGGCTTGGTTTGTTTACC
BACE2	yes	2490	chr21:41459151-41461640	+	ENSG00000182240	ENST00000328735	AAGCCAAGGACCACGATAAAG	CGGGAAACTCGAGTCCCG
BACH1	yes	2474	chr21:29590654-29593127	+	ENSG00000156273	ENST00000286800	AAACAAAAAGTGCCACATCTTTCA	CGGCCCGACTGACTGAAG
BTG3	yes	2434	chr21:17907049-17909482	-	ENSG00000154640	ENST00000339775	TTTGGAGGTAGGGAAGTTAAAGAT	CGCGTTGAGAGGACTGGC
C21orf100	yes	2482	chr21:28831026-28833507	+	ENSG00000178446	ENST00000324961	ATAACCAGAGCACCCATTTATAATG	TTTTATACAGTCAGTCCCAATGTTTT
C21orf109	yes	2529	chr21:29485857-29488385	+	ENSG00000198966	ENST00000361202	TGGTTTGACTTCAACATCTATATTGTTTCC	CCAGAATTGCATCCTTCCTGTGTCATCC
C21orf110	yes	2493	chr21:13682137-13684629	-	ENSG00000175302	ENST00000309142	GCTGTGTTTTCATTTGAAGTTAGTT	AGGAGGAGATAGAAATAAAATATATTAGT
C21orf114	yes	2495	chr21:17997554-18000048	+	ENSG00000196736	ENST00000359578	AGATGCCAGCATCATGCTTC	ACTTCTTGTATCTTTTTTTTTTATTTCC
C21orf115	yes	2486	chr21:41480091-41482576	-	ENSG00000198308	ENST00000356319	TTCCCACCACCATGTTGACTTC	ACGAGGCAGCTGTACTATGG
C21orf116	yes	2468	chr21:15341962-15344429	+	ENSG00000197510	ENST00000355858	GGAGTAGCACTTTTAATGTCCTTTA	ATTGGTCTCAATAGCTTATCCTTG
C21orf119	yes	2482	chr21:32684864-32687345	+	ENSG00000186837	ENST00000331283	TTAGAATGGCGATCATTAAAAAGTC	CCCAAGACTCTGGATCCCTATC
C21orf120	yes	2486	chr21:33085192-33087677	-	ENSG00000198655	ENST00000354928	TAGGCTTGGTCTCTGCAACTA	TGGAGCTGAGCCACTGTAGA
C21orf121	yes	2477	chr21:42299928-42302404	+	ENSG00000182852	ENST00000358893	ATGGTCTCTAAGAGGTAATTAAGTTAA	GGAGGCGGGGTGTTTTTAGA
C21orf123	yes	2538	chr21:45669371-45671908	-	ENSG00000183535	ENST00000330082	GCAGAGGTGTGAGGTCAAGAGCAG	GCAGGAGGTTTCTCAGATGCCC
C21orf125	yes	2436	chr21:43691912-43694347	+	ENSG00000188660	ENST00000342757	GAACCCTGGGCACTCGACC	GGGGCAGAGGGCAATGGAC
C21orf128	yes	2488	chr21:42401581-42404068	-	ENSG00000184385	ENST00000329015	ACCAGGTCACCTGCCAAGAC	TGTTGTGGCAGCTGACGAG
C21orf129	yes	2480	chr21:42008964-42011443	-	ENSG00000185481	ENST00000331508	GGAGGATGGAAGTGGAGACTAC	TCCTCACCTCTGCGTTCCC
C21orf13	yes	2490	chr21:39739488-39741977	-	ENSG00000157578	ENST00000358268	TGATTAAATCATCCTGGATTTAGGG	TGAGGGGGACGACGGCAG
C21orf15	yes	2478	chr21:14138261-14140738	-	ENSG00000198401	ENST00000359319	CCACTAGCTTGCTGAGACATTA	CGTCCAGGGCCAGACCAG
C21orf19	yes	2450	chr21:36421720-36424169	+	ENSG00000179408	ENST00000319751	AGTAAACCACCAGAACCCACAA	CCGCCTATTTTTATTTTTATTCTTATTTA
C21orf2	yes	2452	chr21:44583683-44586134	-	ENSG00000160226	ENST00000339818	CCTGCCTTAACTGATAACATTGTC	CCGCTGACACGTTGGCTC
C21orf25	yes	2508	chr21:42247178-42249685	-	ENSG00000157617	ENST00000329623	GGGTGCTGGTTAGTTTACAGTTTG	GGTGGCCGGAGGAGGCTCT
C21orf29	yes	2488	chr21:44955928-44958415	-	ENSG00000175894	ENST00000323084	CCTCACAAACTGCTAGGATTTC	TCTGCGTCAGCCTTTCCAG
C21orf30	yes	2466	chr21:44701759-44704224	+	ENSG00000183637	ENST00000332734	CTGCTGTCACTTTTGTAGTAGGT	GGACAGGGCCTCCGGTGT
C21orf33	yes	2461	chr21:44375499-44377959	+	ENSG00000160221	ENST00000291577	GTCTTGTGGGGGCTCAAGAC	GCACCCACCCCTCTACAAAC
C21orf32	yes	2449	chr21:44418039-44420487	-	ENSG00000185225	ENST00000327663	GAGACACCAGGGCCACAAATC	CTTCTTGGCGTGTCTTTAAATGT
C21orf34	yes	2496	chr21:16362927-16365422	+	ENSG00000174496	ENST00000308787	ATTCACTTAGAGAAAAAAGGATTTTAC	AGTAGTGGCACCCTAAGAGAG
C21orf41	yes	2480	chr21:29888048-29890527	+	ENSG00000183653	ENST00000333765	ACCCCACAGCTTGATGTCTG	AACAGGTTTGCAGTGGATTAAATAA
C21orf42	yes	2523	chr21:25725452-25727974	-	ENSG00000185433	ENST00000332587	TGCACTGTAGCAAGGCTGGCATTT	GCATGAAGGAGATGAAGATATCTGTTCTCA
C21orf45	yes	2453	chr21:32573285-32575737	-	ENSG00000159055	ENST00000290130	GTCATCTTTTTCAAGCATCTCATTT	AGAAGGGGCAAAAGCCTGAA
C21orf55	yes	2370	chr21:33785722-33788091	-	ENSG00000177692	ENST00000314399	TGTTTTCTGCGTGAAGTGTTG	TAGACCATTTGCCACTTTACACT
C21orf56	yes	2472	chr21:46428691-46431162	-	ENSG00000160284	ENST00000330205	GCATTCCCACCAACAATGAGTT	AAATCACCTCAGTGTTGTTTCTTG
C21orf57	yes	2492	chr21:46528277-46530768	+	ENSG00000182362	ENST00000339195	CCTGCTCCGATGACATGTTAG	CCCGGACGTACCGGTGTC
C21orf58	yes	2466	chr21:46568194-46570659	-	ENSG00000160298	ENST00000310126	CCACAAAGTCAACAGACTGGTG	CCAGCAGCGCGAACTTTT
C21orf62	yes	2494	chr21:33107818-33110311	-	ENSG00000159100	ENST00000290186	AATAGAGGTTGAGTACGCCAAAC	TCTGCCTGGCCACGGATC
C21orf66	yes	2478	chr21:33066005-33068482	-	ENSG00000159086	ENST00000331923	GGATGCTCTTAGCTCCAAAACT	CCCACGCGCTCTCAATACT
C21orf67	yes	2494	chr21:45184211-45186704	-	ENSG00000183250	ENST00000332379	GTACATGGGAAAATATCAGACTCAC	AGGTGCGTCGGTGAAAGG
C21orf70	yes	2480	chr21:45181880-45184359	+	ENSG00000160256	ENST00000291634	ACCTGATTAACACAGTTCCTTATTG	AAGCTGGTCCTGCGTTCTC
C21orf74	yes	2527	chr21:22401164-22403690	+	ENSG00000184856	ENST00000333482	TTCCTGTTGTCCATTTGTATGTTTTCTTTT	TCCCAAACAGCTGAGATTAAAGACATAAGC
C21orf77	yes	2466	chr21:32870050-32872515	-	ENSG00000186842	ENST00000334165	CCGTGCTCCCTGCTTCTG	CAAATGGAGAAAGGGGAACAAAG
C21orf82	yes	2452	chr21:34474400-34476851	+	ENSG00000198773	ENST00000361850	CACTTCATCACTCTGGTATATTATTAT	TTTATAGACTCCCACCCCCTTT
C21orf84	yes	2492	chr21:43722444-43724935	-	ENSG00000185186	ENST00000332440	TTCCGTGGCCGTGTACACA	CACCTTTGCCCTGGCTATGTA
C21orf86	yes	2477	chr21:45538991-45541467	-	ENSG00000184836	ENST00000333529	GTCCTAATTTTAAAGCACACACAAC	TGGGGACTGTGAGAATCTGTC
C21orf87	yes	2465	chr21:39608216-39610680	-	ENSG00000183582	ENST00000327808	TCACTGCAGCGTCAATCCC	ACAGGAAAACGGGACTGTGG
C21orf88	yes	2536	chr21:39906588-39909123	-	ENSG00000184809	ENST00000359876	TCACCTTTCATCTTCAGACACAGG	TCCCGTCGCCTCCCGGAGC
C21orf9	yes	2490	chr21:30040637-30043126	+	ENSG00000174680	ENST00000309331	TGTGGGAATAAAATACAGAGGATAAC	AACTCCAGATCGAATTAGGATCA
C21orf90	yes	2482	chr21:44759035-44761516	+	ENSG00000182912	ENST00000330490	GGGAGGCCCAGTCATTGGA	AACTGGTATGTGTCCTAGAGAGA
C21orf91	yes	2522	chr21:18113524-18116045	-	ENSG00000154642	ENST00000284881	CACAATCTCAGCTCACACAACATCCACCTC	TGCGGCTCGGGTTCCTCCAC
C21orf93	yes	2403	chr21:45549612-45552014	-	ENSG00000184274	ENST00000328338	ATTTGCACATGCGTTCATGTTC	GGTGCTTCTCGCTTTTAATTTTAC
C21orf94	yes	2480	chr21:28305075-28307554	+	ENSG00000178457	ENST00000324988	TCTAATTTTAGTCCCCATTGTTGG	TGTTCTTTGCTTTCTTCATTGATTT
C21orf99	yes	2476	chr21:13329918-13332393	+	ENSG00000155258	ENST00000285609	AACTCATGTGATCTGGGTGGTG	AGCCACCTGCTAGAGAGAG
CBR1	yes	2517	chr21:36361659-36364175	+	ENSG00000159228	ENST00000290349	CCCTCTTAATAAATTTAATCCATCCACTCT	CACCCGTGGGTTACACGCCC
CBR3	yes	2496	chr21:36426826-36429321	+	ENSG00000159231	ENST00000290354	GTCTGTTCGAAGGCATTGCC	ACCGGGCTTCGGACCACC
CBS	yes	2477	chr21:43369513-43371989	-	ENSG00000160200	ENST00000359624	TGATATAACTAAGTGAATAAATAAATGAG	CGTCCTCGAGGACCGACC
CCT8	yes	2492	chr21:29367838-29370329	-	ENSG00000156261	ENST00000286788	CCCAAAAGGCCCTACGTAATAC	AGGAGCACGCACAGCCTT
CHAF1B	yes	2470	chr21:36677130-36679599	+	ENSG00000159259	ENST00000314103	CCTCCCAGATGCTGCAATTAC	AGAGGCACCGTCACCTCC
CHODL	yes	2471	chr21:18536580-18539050	+	ENSG00000154645	ENST00000299295	ATTTTAATTCATAGCCCCTACTTTTTT	TGCGCCCTGGTCCTGGAT
CLDN14	yes	2484	chr21:36870702-36873185	-	ENSG00000159261	ENST00000342108	GCCTCCAGAACGCTAAAACA	CCTGCCTCTCTCAGCTTTTC
CLDN17	yes	2486	chr21:30460811-30463296	-	ENSG00000156282	ENST00000286808	GAGGGAGAAGGTGTCAGTACAAT	TTTACTTTGAAGACTGGTTCAATTC
CLDN8	yes	2477	chr21:30510182-30512658	-	ENSG00000156284	ENST00000286809	TTCTGTCTCATGCGTTCCTGT	TCTGTGCCACAGAAGAGAACA
CLIC6	yes	2476	chr21:34961076-34963551	+	ENSG00000159212	ENST00000360731	ACACCCCCCCCCAATAGCC	AGATCCCTGTCCTTGATCGG
COL6A2	yes	2529	chr21:46339940-46342468	+	ENSG00000142173	ENST00000300527	CCAGAATCATCTTTGACCCACGACCT	GAGTAAGCAGCAGGGAGGGAA
CRYAA	yes	2421	chr21:43459831-43462251	+	ENSG00000160202	ENST00000291554	TAAATCAAAATGAATGCAAAAAATTTGT	GAGGCTGGCAGGAGTCAG
CXADR	yes	2417	chr21:17804755-17807171	+	ENSG00000154639	ENST00000284878	ACAAAACTAATATTGGATAATATTTGTCC	AGGTGAACTCGCGTCACTTC
CYYR1	yes	2462	chr21:26867490-26869951	-	ENSG00000166265	ENST00000299340	GCCTTTCCATGTCAATTGAGTTT	GGGTGGCCCCGACTCCTC
DIP2A	yes	2512	chr21:46700731-46703242	+	ENSG00000160305	ENST00000339331	CCCTAATCAGTATGACTGGTGTCCTTCTA	CGGCCCTACACGGGAGCG
DNMT3L	yes	2468	chr21:44506479-44508946	-	ENSG00000142182	ENST00000270172	ATAGAGGGAGACTCCTCAAAAAAC	TTTGGGGACCTCAGTGGGGA
DONSON	yes	2498	chr21:33936516-33939013	-	ENSG00000159147	ENST00000290244	AAAAAAAAAAGGAAAGGAATTTACCAG	GTAGGGGGTGGTGAGAAGGTG
DOPEY2	yes	2443	chr21:36456275-36458717	+	ENSG00000142197	ENST00000270190	CAAGGTGCTGGCAGATTCG	ACTGTGGGAGATGAAAGATAAATAG
DSCAM	yes	2492	chr21:41002576-41005067	-	ENSG00000171587	ENST00000304537	GCAGGAGCCTGTAACCCCT	AGACAACACAGGGATCCATAGG
DSCR10	yes	2473	chr21:38497679-38500151	+	ENSG00000181425	ENST00000348862	TAAGGTCAGAATCTTCTCCCATAAT	CCCAGCACAACTGATCTTCTAT
DSCR2	yes	2490	chr21:39477262-39479751	-	ENSG00000183527	ENST00000331573	CAATAGGATTTAATTTGTTGTTAATATGT	GCGCGGGCAATAAGTCCC
DSCR4	yes	2488	chr21:38415231-38417718	-	ENSG00000184029	ENST00000328264	GGACTGTAAGGAGAGAGGTCAC	GGTGAGTTAAGAAGCGGAAAGT
DSCR6	yes	2481	chr21:37297882-37300362	+	ENSG00000183145	ENST00000357131	ACATATTTTTATTAACTGCATACTATTCT	CGGCTTCTCCGTCCCTCT
DSCR8	yes	2482	chr21:38413800-38416281	+	ENSG00000198054	ENST00000357704	TTTAAGGAAGGACAAAAAAAATCATAG	ATGAAGCTAAGGAAACAGGAAATAA
DSCR9	yes	2495	chr21:37500228-37502722	+	ENSG00000182342	ENST00000333494	ACAAGATCTCCATATGAAAATAAATGA	GGCTCACCCCAGAGATAAAG
ERG	yes	2482	chr21:38955504-38957985	-	ENSG00000157554	ENST00000357391	CCAGGATTTTATAACTCATAACACCA	CGGAGCGATCTCTAGGAAC
ETS2	yes	2476	chr21:39096606-39099081	+	ENSG00000157557	ENST00000288325	ATTTGTCTTCTCAGCATCTGACT	TGAGGCGCTTCCTGTAGAAA
FAM3B	yes	2439	chr21:41608107-41610545	+	ENSG00000183844	ENST00000357985	GAGAAATAAATGTGTGCTGTTTAAGT	GGCCAGTCGGGCAAATGG
FGF7P2	yes	2460	chr21:13640996-13643455	+	ENSG00000185390	ENST00000332473	ACCATGTTGGCCAGGTTAGTATT	TGCAAGGAATCACAGAAAGACAT
FTCD	yes	2475	chr21:46399933-46402407	-	ENSG00000160282	ENST00000355384	GCTTAAACCACAGAAATTTATTCTATC	GCCGCCAGGGCCTTTATAC
GABPA	yes	2452	chr21:26026308-26028759	+	ENSG00000154727	ENST00000354828	AAGTGAGTTTTTTTCTCTCCAAGG	GCTGGGGCAACCCAAGGTC
GART	yes	2492	chr21:33837070-33839561	-	ENSG00000159131	ENST00000290230	CTGATAAAAGTCTGAATGGTCAAAAA	CCCCAGCCGCTCAGAAAC
GRIK1	yes	2482	chr21:30233700-30236181	-	ENSG00000171189	ENST00000327783	TTCAAACTTTAGCCACTCACTATC	TCTTAATTCATGCCGAGATACAG
H2BFS	yes	2470	chr21:43807052-43809521	+	ENSG00000197597	ENST00000361105	TCACCTCATCAAGTACTTTGTACT	CGAGCAGCAGATCGTGAAA
HLCS	yes	2522	chr21:37284363-37286884	-	ENSG00000159267	ENST00000336648	TTTTGGAATTTCTCTGGAAATTTCC	TGGTCGCCCGCGAATCCCTG
HMGN1	yes	2487	chr21:39642871-39645357	-	ENSG00000157576	ENST00000288344	AGCCAGCCCCTCAAATTTTTC	TGGGAGAACCGGATGGAACC
HSF2BP	yes	2417	chr21:43903829-43906245	-	ENSG00000160207	ENST00000291560	GAGGCTTTTTGAAGGAAGTGATAC	TTTGTGACGCCATCAGCCC
IFNAR2	yes	2477	chr21:33521655-33524131	+	ENSG00000159110	ENST00000318433	AAGTTTCTTAACTTTTGCTTTTTTCC	GGCCTACGAGGAGCGATAC

IFNGR2	yes	2515	chr21:33695070-33697584	+	ENSG00000159128	ENST00000290219	CAAAGCGCTGGGATTAGAGG	GGGCGTCCCGGTGGAGGG
IGSF5	yes	2493	chr21:40056834-40059326	+	ENSG00000183067	ENST00000328060	TCTTCTTTACAGTTTTCTCCTATTCT	AGGTAGCCAATTTTTCAAGTTGTT
IL10RB	yes	2415	chr21:33558145-33560559	+	ENSG00000159113	ENST00000290200	TCAGCCTCCCTCGCAAGTA	CCGGGAACCAGCGGAGAT
ITGB2	yes	2455	chr21:45173136-45175590	-	ENSG00000160255	ENST00000355153	TGTGCTGGGGTGGGTAGGA	GGATGTGTCTCCAAGAAAAGTGT
JAM2	yes	2492	chr21:25931012-25933503	+	ENSG00000154721	ENST00000312957	ACGGTGACCATTTACTTCACAAC	GGGAAACGCCTCCTGTTTTG
KCNE1	yes	2488	chr21:34806397-34808884	-	ENSG00000180509	ENST00000337385	TTCTCACTCCGTGACTCACC	AGGCTTATTTCCAGAAAGTACTCT
KCNE2	yes	2464	chr21:34655744-34658207	+	ENSG00000159197	ENST00000290310	GTGATTTCTGTTTAAAAGTTTTTTCTTAA	GCACCTTCACCTTACAACCATC
KCNJ6	yes	2450	chr21:38210610-38213059	-	ENSG00000157542	ENST00000288309	CCCAAGGGTGGAGACCAAAAA	ACGCGGCTCCGAGATAAAA
KRTAP11-1	yes	2492	chr21:31175722-31178213	-	ENSG00000182591	ENST00000332378	GTTCAGAGACTCTGTTTTAGTTGA	GACAGAGGGCTGCAGGTAG
KRTAP13-1	yes	2481	chr21:30687730-30690210	+	ENSG00000182816	ENST00000328750	CAGTATCTGTTGTATAATTACCTTTTTAT	GGACAGGACCATTTATATACGTTC
KRTAP13-2	yes	2483	chr21:30666415-30668897	-	ENSG00000198390	ENST00000355459	GCAATAGAAGATTGTCAGTTTAAAG	TGTGAGTTCAGCTGAGTTATAGT
KRTAP13-3	yes	2463	chr21:30720137-30722599	-	ENSG00000186975	ENST00000334147	CTCTCTGCCTCTAAATCTCTAAA	TGGGAAGATTCTGAGTTTTAACG
KRTAP13-4	yes	2459	chr21:30721966-30724424	+	ENSG00000186971	ENST00000334068	AGTTGATGTCTTTTATCAAGATTGGT	GATTCTGAGTTTTAATGTCGCATC
KRTAP15-1	yes	2485	chr21:30732023-30734507	+	ENSG00000186970	ENST00000334067	TTAAAGAAAGTTTATAAAGTAACTGCTT	AGTTCTGAGCTCAGGTGATTTG
KRTAP19-1	yes	2489	chr21:30774516-30777004	-	ENSG00000184351	ENST00000331764	AGTTTACACGGCTAGTGGACAA	GAGTGGTGAGTTGGTTTGTTGTT
KRTAP19-2	yes	2434	chr21:30781593-30784026	-	ENSG00000186977	ENST00000334151	AACAAAAAATACTCTTGGCAAAATGT	GCTCACGGTGTCAGAGATAGT
KRTAP19-3	yes	2527	chr21:30786097-30788623	-	ENSG00000186967	ENST00000334063	CCTGGCCAACAGCTTTCTTTATTC	CCTCCACAGCCATAGCCCAG
KRTAP19-5	yes	2491	chr21:30796288-30798778	-	ENSG00000186965	ENST00000334055	TTGAGGTTTTCCATTTATATCTTGTTT	GAGTGGTGAGTTGGTATGCTG
KRTAP19-6	yes	2492	chr21:30836027-30838518	-	ENSG00000186964	ENST00000334046	TTCATACTACCCAAAGTGCTCTAA	GTCAGGGACTAGAGATTCAGTTC
KRTAP19-7	yes	2491	chr21:30855484-30857974	-	ENSG00000186925	ENST00000334849	GAATCATCCATAAAAATGTGTTTTTTTC	TCAGAAGTGGTTAGCTGTTGTAG
KRTAP20-1	yes	2454	chr21:30908189-30910642	+	ENSG00000186923	ENST00000334664	TGTGGGTGGGAACCATACACAA	TGTCAGAAATGGATGGTTCAGGT
KRTAP20-2	yes	2491	chr21:30926962-30929452	+	ENSG00000184032	ENST00000330798	AGTCTTCTCCAAATATCAGTGTTTC	GTTTCAAGAGTGGAGGATTTGGT
KRTAP21-2	yes	2519	chr21:31041363-31043881	-	ENSG00000187026	ENST00000333892	TCCCTGATACCTCAATAAAGCTGGAAAA	CAGCAGTTTCTGTAGTAGTTGCAACACA
KRTAP22-1	yes	2481	chr21:30892821-30895301	+	ENSG00000186924	ENST00000334680	CAGAATTGGTGAATCCCGATGT	AAAGGTAGGCTCAGTTGTATAGC
KRTAP23-1	yes	2458	chr21:30642830-30645287	-	ENSG00000186980	ENST00000334160	CTATCTTAACAGGCCTTGGTTTAC	TGAGAAGATGCCCTGATGTCC
KRTAP6-1	yes	2516	chr21:30908072-30910587	-	ENSG00000184724	ENST00000329122	CCATCTTCACAGGTCTTTTATACTTCCTC	TTTCCGTAGTAGCTGCCACACA
KRTAP6-2	yes	2454	chr21:30893110-30895563	-	ENSG00000186930	ENST00000334897	CAATGTGGGCTAAATTTTAGTTGTC	GACTTCAGTATGGATACTTCCTGTA
KRTAP8-1	yes	2492	chr21:31107417-31109908	-	ENSG00000183640	ENST00000329621	GAAAACCTTGCTCTTTCTTTGATA	GGAGTGGAGGGCTTAGGTAG
LIPI	yes	2489	chr21:14501128-14503616	-	ENSG00000188992	ENST00000344577	AGTGTGTTTTATTTATTACACATGCTC	ATCAGAAGTTCACTAGCTCTTTG
LRRC3	yes	2393	chr21:44697380-44699772	+	ENSG00000160233	ENST00000291592	CGAGATTGCCCCCACTGAAC	GGCGAAAGGCGGAGACAG
MCM3AP	yes	2387	chr21:46529668-46532054	-	ENSG00000160294	ENST00000291688	GGCTGTCACCTTTCTTTTTCCT	GTACAAAATTAATTGGCTCTTCCTGAAA
MRAP	yes	2457	chr21:32590613-32593069	+	ENSG00000170262	ENST00000303645	GGGGGATGTCTTTGTGATATTTTG	TGATTTCTGCAGGAATCGTCC
MRPL39	yes	2532	chr21:25901652-25904183	-	ENSG00000154719	ENST00000352957	TCTTTAGTCTCCTTTAATTTAGGACAGTTT	CCTCCATAGCAGCGGTGAGAA
MRPS6	yes	2473	chr21:34416907-34419379	+	ENSG00000198743	ENST00000362077	AGCTGAGACCCGACTCTTTAT	AAGCTGGTTTTGGTTAACTTAAAATA
MX1	yes	2495	chr21:41723064-41725558	+	ENSG00000157601	ENST00000288383	TCAGGGAAAGGTACAAACGTTTAT	GGGCTCACGCCTGTAATCT
MX2	yes	2459	chr21:41653349-41655807	+	ENSG00000183486	ENST00000330714	AGGGGGCTTGCCACATCAG	CTTCACTTTTATGTGGCATATGAAC
HEMK2	yes	2465	chr21:29179593-29182057	-	ENSG00000156239	ENST00000303775	TTGGATGTCACAACTAGTGAATATG	TGCCCTCACCCTCATGCG
NCAM2	yes	2492	chr21:21572318-21574809	+	ENSG00000154654	ENST00000284894	AAAGAAAAGAGAAGACACATAATTATAA	AAGCTCTACTTTGCTAAGTGAAAT
NDUFV3	yes	2431	chr21:43184006-43186436	+	ENSG00000160194	ENST00000340344	TTTTATTTTAATGCAGCACCTTATATTT	CCGACTGCGCCAATTACAG
NRIP1	yes	2498	chr21:15262386-15264883	-	ENSG00000180530	ENST00000318948	TGAGTCTACCCAAGAGCAAAAA	TTCAATAGAAGTGTTCACAAGGG
OLIG1	yes	2421	chr21:33361876-33364296	+	ENSG00000184221	ENST00000333063	CCTACGGGCCCCACCTAG	ACCCCTTTAAACCCGGCTTG
OLIG2	yes	2483	chr21:33317619-33320101	+	ENSG00000185569	ENST00000333337	GGCACCAGGGCCAAGGGA	GGCTCGGCCGGTTTTTATAG
PCBP3	yes	2472	chr21:46138145-46140616	+	ENSG00000183570	ENST00000330561	CGGGCTCCATTCTGAGAAAC	TGAGGGTGGTGGCTTAAGGT
PCNT	yes	2481	chr21:46566034-46568514	+	ENSG00000160299	ENST00000359568	ACTGTCACCTCGACTATAACTG	TTCGCTCTATTTACACTCCCTC
PCP4	yes	2484	chr21:40158769-40161252	+	ENSG00000183036	ENST00000328619	TGAGAAAAATCCCCCTTAAGGTAA	ACACTGCTCAGGGGTCGC
PDXK	yes	2526	chr21:43960899-43963424	+	ENSG00000160209	ENST00000291565	CCCTCCAGGGCTCTCTCCCT	TCCGAACCCGCGAGTTCCGC
PFKL	yes	2506	chr21:44541836-44544341	+	ENSG00000141959	ENST00000312352	CCTCGGGCCTCAGTTTCCC	CGCGCCTTCCGCCCGCCC
PIGP	yes	2525	chr21:37367166-37369690	-	ENSG00000185808	ENST00000360525	GACAGTGAAAATTGGTACAACCACTTC	CCACCTCAGCGCTGAGTTCTCC
POFUT2	yes	2465	chr21:45532267-45534731	-	ENSG00000186866	ENST00000349485	GCCTGAGGTTAAAAGATTACCCC	CCCCGCCCCCGTGAAATG
PPIAL3	yes	2464	chr21:19149473-19151936	+	ENSG00000198618	ENST00000358455	CATTATATTTAGAGGTGCAGAATACA	GCGTCTGCAAAGCCAAGTAA
PRDM15	yes	2528	chr21:42172481-42175008	-	ENSG00000141956	ENST00000269844	CTGATAGCCATGGAAGCTCTTGTCTGTT	CGGACTCCGGGTTGCGATCC
PRED15	yes	2489	chr21:21935609-21938097	+	ENSG00000196190	ENST00000358903	TGTTGGTCTTGCAGTCAATGTTA	GAGTTTAATTGTCCTTACAGTTTTG
PRED58	yes	2502	chr21:45359281-45361782	-	ENSG00000184941	ENST00000328881	CCCATATACATAACAGAAACTCCTCATC	CTCGCACTGACCTGTAGAGA
PRMT2	yes	2475	chr21:46877008-46879482	+	ENSG00000160310	ENST00000355680	TAGCCTCTGTTGCAGTAACTATT	TTGTGGGGAGAGGAAGTTTGG
PRSS7	yes	2515	chr21:18697811-18700325	-	ENSG00000154646	ENST00000284885	CTTTATCACCCTCCATCTCTTCACTTCCC	TTTTGGTTTTGAAGGCTTGCTAATTTAAG
PTTG1IP	yes	2423	chr21:45118173-45120595	-	ENSG00000183255	ENST00000330938	AGGCTGAGGCAGAGGAGTC	GCCTGCGCGATAGAGGAA
PWP2	yes	2481	chr21:44349197-44351677	+	ENSG00000160220	ENST00000291576	GAGAAGGAGCACCGGGTTC	AGACACTTCCCGGGGGTCC
RBM11	yes	2537	chr21:14507854-14510390	+	ENSG00000185272	ENST00000331453	AAAAGAGAGTAAGAGTTGGCTGCACA	AGCAGGGAACATCCTCCGGTCTC
RCAN1	yes	2479	chr21:34909272-34911750	-	ENSG00000159200	ENST00000313806	TACCTTCTCTCTCTGTGGATTTG	GCTCCGCCGTTAACCCCC
RIPK4	yes	2453	chr21:42060321-42062773	-	ENSG00000183421	ENST00000332512	AGGAAACTGGCTTGGAATTCTT	GCTCCGCCCTGTTACGTC
RRP1	yes	2454	chr21:44031450-44033903	+	ENSG00000160214	ENST00000291569	ACTGTGCATCTGACACGCC	CCTGGTACCCAGCACACTGA
RRP1B	yes	2352	chr21:43901496-43903847	+	ENSG00000160208	ENST00000340648	ATGGTGAAACTCCATCTCACTAAA	GGGCTGATGGCGTCACAAA
RSPH1	yes	2429	chr21:42789539-42791967	-	ENSG00000160188	ENST00000291536	GTCACTGCCAGAAAACTATAGTTAT	AGCCTTTGCGTACTGTCATAG
RUNX1	yes	2492	chr21:35343467-35345958	-	ENSG00000159216	ENST00000300305	CAGACTCTTGGGTACAATCCCC	AGTTTTCACACAACCCAAATTACAA
RWDD2B	yes	2449	chr21:29313588-29316036	-	ENSG00000156253	ENST00000286777	GCTATCATACCCAGCCATAATTTT	GACCGGAAGCTGGGGCTC
S100B	yes	2453	chr21:46849396-46851848	-	ENSG00000160307	ENST00000291700	CTTGGCTTCTCCCTTCAATCA	GTCGTCTCTGCTGCGGGT
SAMSN1	yes	2498	chr21:14877546-14880043	-	ENSG00000155307	ENST00000285670	AATAATTTGCTACTAGAAAGTAATGAG	TTGTTTTCTATGGCCAAGGAATG
SETD4	yes	2443	chr21:36354560-36357002	-	ENSG00000185917	ENST00000344232	TTTTGTTTTGTTTTGTTTTGTTTTAGC	GCCCAAGGCACGTGGAAATA
SH3BGR	yes	2468	chr21:39743158-39745625	+	ENSG00000185437	ENST00000333634	CTGGTCTTTGTGAACAGACATAAA	CACTGGACCCTTTCCCTATCAC
SIM2	yes	2522	chr21:36990821-36993342	+	ENSG00000159263	ENST00000290399	TTGGGTTTCTTCCTTCACAATTCCAC	TGTGTGGCTAGGCGCAGCAC
SLC19A1	yes	2520	chr21:45786768-45789287	-	ENSG00000173638	ENST00000311124	TCTCTCCAGCCAAGGCAGAGGA	CTCCGGGACTACAGCGCCCA
SOD1	yes	2463	chr21:31951461-31953923	+	ENSG00000142168	ENST00000270142	GTGGCCTGTTCAGGATCCTATA	CCGAGGACTGCAACGGAAA
SON	yes	2495	chr21:33834772-33837266	+	ENSG00000159140	ENST00000290239	AGCAGTGTAACCTAGGAGATTAAC	CGTCCGCTCCGTTCTCTC
STCH	yes	2478	chr21:14677339-14679816	-	ENSG00000155304	ENST00000285667	TACTTCCCAACCCCAGTGCC	GACTGTACCAGACGTGAGGC
SUMO3	yes	2388	chr21:45062530-45064917	-	ENSG00000184900	ENST00000332859	CGTAGCCCTTCCTGTACTGA	GCTCTGGGCGGTTGGAGA
SYNJ1	yes	2440	chr21:33022115-33024554	-	ENSG00000159082	ENST00000357345	CTGTGTGCGTGTTTAATTTTCCT	CCGGCTTGCTCACCTCTTC
TFF1	yes	2482	chr21:42659723-42662204	-	ENSG00000160182	ENST00000291527	TAAAATCCAGCTGGGCACAGT	CCCGAGCCCCGGATTTTAT
TFF2	yes	2470	chr21:42644180-42646649	-	ENSG00000160181	ENST00000291526	GAAGACGGTGGGTGAAGTGA	GTCCACTGCCGGGTATGTTTT
TFF3	yes	2541	chr21:42608540-42611080	-	ENSG00000160180	ENST00000291525	CACCACTTCATGCTGTCCTGTG	GCTCTGGCAGCCATGACCACC
TIAM1	yes	2485	chr21:31853117-31855601	-	ENSG00000156299	ENST00000286827	CCTGGCCAGCACATTGTTAATA	CGGTCGTCCGGGTATCTCC
TMPRSS3	yes	2461	chr21:42689307-42691767	-	ENSG00000160183	ENST00000291532	CATCTGACACACGTGTTAACATG	CCAACGGCTTGCATCAAAATAA
TRPM2	yes	2468	chr21:44595091-44597558	+	ENSG00000142185	ENST00000300482	GAGCCGGGTGCGGTCTGT	TTCCAGACACCAGTGACTACTTT
TTC3	yes	2424	chr21:37364992-37367415	+	ENSG00000182670	ENST00000354749	GTGCAATCTCAGCTCACGG	TGAAGCGAGGTGCGTCAC
UBASH3A	yes	2541	chr21:42694574-42697114	+	ENSG00000160185	ENST00000319294	GTAGCTGGAACTAATGCCACCATGTCTTTT	CCCTTAGGCCAAGAAGCTCGC
USP25	yes	2484	chr21:16021771-16024254	+	ENSG00000155313	ENST00000285681	TGAAGGCCTTCCTCCTGTATC	GCGTTCGCACGTTCCTTTG
WRB	yes	2444	chr21:39671650-39674093	+	ENSG00000182093	ENST00000333781	TGGGGTGTGGTAGTGGACTTC	CAAGCGCCCCAGAACAGG
ADARB1	–	2493	chr21:45316499-45318991	+	ENSG00000197381	ENST00000354687	CAGGGAGAGCGTCACAATAG	GCCAACCTGGCCGCTTGG
ADARB1	–	2539	chr21:45316398-45318936	+	ENSG00000197381	ENST00000354687	AACAGCTGGGAACGCCATGTC	CGCCTCAGCGCCTCAGCCC
AGPAT3	–	2281	chr21:44107169-44109449	+	ENSG00000160216	ENST00000327505	TCTGAGTAACTAGGATTACAAGAGA	GCGGCGATCCCACAGTTC
AGPAT3	–	2507	chr21:44106956-44109462	+	ENSG00000160216	ENST00000327505	TTTCTTTACCCTTGCCCATTCACTATC	CCACTTCCTGTGCGCGGCGA
ANKRD21	–	2487	chr21:13901921-13904407	+	ENSG00000166351	ENST00000360495	AGGGATAAGAGCAGAAATAAATGAA	CCAGGGGAGGCCGGTAGT
ANKRD21	–	2530	chr21:13894847-13897376	+	ENSG00000166351	ENST00000360495	GCAAATCTGAGGACAGGAAGAAATTAGGGA	GACGGAAATTTGAACCAAGAGAAATGGG
APP	–	2463	chr21:26464841-26467303	-	ENSG00000142192	ENST00000348990	GCCTCAGCCTCCCCAGTA	CTGTGCGAGTGGGATCCG
B3GALT5	–	2491	chr21:39951101-39953591	+	ENSG00000183778	ENST00000343118	TGGGCAGGCACCATCCATT	GCTTGGTTTGTTTACCAAAAATGG
BRWD1	–	2461	chr21:39607465-39609925	-	ENSG00000185658	ENST00000333229	TTAACTTTCTCAAATCCATCCTAAAATA	CGCGGGAGACGAAAAGTAG
C21orf104	–	2489	chr21:39317109-39319597	-	ENSG00000183412	ENST00000329367	GCCTTGCTCTGTAATAACATGAA	ACCCCCGCCCACAGTGTC
C21orf104	–	2507	chr21:39322896-39325402	-	ENSG00000183412	ENST00000329367	GGGAAGTGGCCAGGCATGAGAG	GAGCAATCTGATGACTCCGTTTCTGTTTC
C21orf105	–	2473	chr21:43218747-43221219	-	ENSG00000196847	ENST00000356939	TGAGGCTCCCCTGAGACAAA	AGCAACACACTGTTTTAATAAGCA
C21orf105	–	2528	chr21:43218757-43221284	-	ENSG00000196847	ENST00000356939	TCCCTCACAATTTGCTCACAGGAAAATTCC	TGTTTTAATAAGCACCTGGATGCAGACAGA
C21orf109	–	2496	chr21:29485849-29488344	+	ENSG00000198966	ENST00000361202	TTCAAATCTGGTTTGACTTCAACA	CAAAGAATCAGTCAATCAAAGATGG

Tabel continued

IFNGR2	yes	2515	chr21:33695070-33697584	+	ENSG00000159128	ENST00000290219	CAAAGCGCTGGGATTAGAGG	GGGCGTCCCGGTGGAGGG
IGSF5	yes	2493	chr21:40056834-40059326	+	ENSG00000183067	ENST00000328060	TCTTCTTTACAGTTTTCTCCTATTCT	AGGTAGCCAATTTTTCAAGTTGTT
IL10RB	yes	2415	chr21:33558145-33560559	+	ENSG00000159113	ENST00000290200	TCAGCCTCCCTCGCAAGTA	CCGGGAACCAGCGGAGAT
ITGB2	yes	2455	chr21:45173136-45175590	-	ENSG00000160255	ENST00000355153	TGTGCTGGGGTGGGTAGGA	GGATGTGTCTCCAAGAAAAGTGT
JAM2	yes	2492	chr21:25931012-25933503	+	ENSG00000154721	ENST00000312957	ACGGTGACCATTTACTTCACAAC	GGGAAACGCCTCCTGTTTTG
KCNE1	yes	2488	chr21:34806397-34808884	-	ENSG00000180509	ENST00000337385	TTCTCACTCCGTGACTCACC	AGGCTTATTTCCAGAAAGTACTCT
KCNE2	yes	2464	chr21:34655744-34658207	+	ENSG00000159197	ENST00000290310	GTGATTTCTGTTTAAAAGTTTTTTCTTTAA	GCACCTTCACCTTACAACCATC
KCNJ6	yes	2450	chr21:38210610-38213059	-	ENSG00000157542	ENST00000288309	CCCAAGGGTGGAGACCAAAAA	ACGCGGCTCCGAGATAAAA
KRTAP11-1	yes	2492	chr21:31175722-31178213	-	ENSG00000182591	ENST00000332378	GTTCAGAGACTCTGTTTTAGTTGA	GACAGAGGGCTGCAGGTAG
KRTAP13-1	yes	2481	chr21:30687730-30690210	+	ENSG00000182816	ENST00000328750	CAGTATCTGTTGTATAATTACCTTTTTAT	GGACAGGACCATTTATATACGTTC
KRTAP13-2	yes	2483	chr21:30666415-30668897	-	ENSG00000198390	ENST00000355459	GCAATAGAAGATTGTCAGTTTAAAG	TGTGAGTTCAGCTGAGTTATAGT
KRTAP13-3	yes	2463	chr21:30720137-30722599	-	ENSG00000186975	ENST00000334147	CTCTCTGCCTCTAAATCTCTAAA	TGGGAAGATTCTGAGTTTTAACG
KRTAP13-4	yes	2459	chr21:30721966-30724424	+	ENSG00000186971	ENST00000334068	AGTTGATGTCTTTTATCAAGATTGGT	GATTCTGAGTTTTAATGTCGCATC
KRTAP15-1	yes	2485	chr21:30732023-30734507	+	ENSG00000186970	ENST00000334067	TTAAAGAAAGTTTATAAAGTAACTGCTT	AGTTCTGAGCTCAGGTGATTTG
KRTAP19-1	yes	2489	chr21:30774516-30777004	-	ENSG00000184351	ENST00000331764	AGTTTACACGGCTAGTGGACAA	GAGTGGTGAGTTGGTTTGTTGTT
KRTAP19-2	yes	2434	chr21:30781593-30784026	-	ENSG00000186977	ENST00000334151	AACAAAAAATACTCTTGGCAAAATGT	GCTCACGGTGTCAGAGATAGT
KRTAP19-3	yes	2527	chr21:30786097-30788623	-	ENSG00000186967	ENST00000334063	CCTGGCCAACAGCTTTCTTTATTC	CCTCCACAGCCATAGCCCAG
KRTAP19-5	yes	2491	chr21:30796288-30798778	-	ENSG00000186965	ENST00000334055	TTGAGGTTTTCCATTTATATCTTGTTT	GAGTGGTGAGTTGGTATGCTG
KRTAP19-6	yes	2492	chr21:30836027-30838518	-	ENSG00000186964	ENST00000334046	TTCATACTACCCAAAGTGCTCTAA	GTCAGGGACTAGAGATTCAGTTC
KRTAP19-7	yes	2491	chr21:30855484-30857974	-	ENSG00000186925	ENST00000334849	GAATCATCCATAAAAATGTGTTTTTTTC	TCAGAAGTGGTTAGCTGTTGTAG
KRTAP20-1	yes	2454	chr21:30908189-30910642	+	ENSG00000186923	ENST00000334664	TGTGGGTGGGAACCATACACAA	TGTCAGAAATGGATGGTTCAGGT
KRTAP20-2	yes	2491	chr21:30926962-30929452	+	ENSG00000184032	ENST00000330798	AGTCTTCTCCAAATATCAGTGTTTC	GTTTCAAGAGTGGAGGATTTGGT
KRTAP21-2	yes	2519	chr21:31041363-31043881	-	ENSG00000187026	ENST00000333892	TCCCTGATACCTCAATAAAGCTGGAAAA	CAGCAGTTTCTGTAGTAGTTGCAACACA
KRTAP22-1	yes	2481	chr21:30892821-30895301	+	ENSG00000186924	ENST00000334680	CAGAATTGGTGAATCCCGATGT	AAAGGTAGGCTCAGTTGTATAGC
KRTAP23-1	yes	2458	chr21:30642830-30645287	-	ENSG00000186980	ENST00000334160	CTATCTTAACAGGCCTTGGTTTAC	TGAGAAGATGCCCTGATGTCC
KRTAP6-1	yes	2516	chr21:30908072-30910587	-	ENSG00000184724	ENST00000329122	CCATCTTCACAGGTCTTTTATACTTCCTC	TTTCCGTAGTAGCTGCCACACA
KRTAP6-2	yes	2454	chr21:30893110-30895563	-	ENSG00000186930	ENST00000334897	CAATGTGGGCTAAATTTTAGTTGTC	GACTTCAGTATGGATACTTCCTGTA
KRTAP8-1	yes	2492	chr21:31107417-31109908	-	ENSG00000183640	ENST00000329621	GAAAACCTTGCTCTTTCTTTGATA	GGAGTGGAGGGCTTAGGTAG
LIPI	yes	2489	chr21:14501128-14503616	-	ENSG00000188992	ENST00000344577	AGTGTGTTTTATTTATTACACATGCTC	ATCAGAAGTTCACTAGCTCTTTG
LRRC3	yes	2393	chr21:44697380-44699772	+	ENSG00000160233	ENST00000291592	CGAGATTGCCCCCACTGAAC	GGCGAAAGGCGGAGACAG
MCM3AP	yes	2387	chr21:46529668-46532054	-	ENSG00000160294	ENST00000291688	GGCTGTCACCTTTCTTTTTCCT	GTACAAAATTAATTGGCTTCCTGAAA
MRAP	yes	2457	chr21:32590613-32593069	+	ENSG00000170262	ENST00000303645	GGGGGATGTCTTTGTGATATTTTG	TGATTTCTGCAGGAATCGTCC
MRPL39	yes	2532	chr21:25901652-25904183	-	ENSG00000154719	ENST00000352957	TCTTTAGTCTCCTTTAATTTAGGACAGTTT	CCTCCATAGCAGCGGTGAGAA
MRPS6	yes	2473	chr21:34416907-34419379	+	ENSG00000198743	ENST00000362077	AGCTGAGACCCGACTCTTTAT	AAGCTGGTTTTGGTTAACTTAAAATA
MX1	yes	2495	chr21:41723064-41725558	+	ENSG00000157601	ENST00000288383	TCAGGGAAAGGTACAAACGTTTAT	GGGCTCACGCCTGTAATCT
MX2	yes	2459	chr21:41653349-41655807	+	ENSG00000183486	ENST00000330714	AGGGGGCTTGCCACATCAG	CTTCACTTTTATGTGGCATATGAAC
HEMK2	yes	2465	chr21:29179593-29182057	-	ENSG00000156239	ENST00000303775	TTGGATGTCACAACTAGTGAATATG	TGCCCTCACCCTCATGCG
NCAM2	yes	2492	chr21:21572318-21574809	+	ENSG00000154654	ENST00000284894	AAAGAAAAGAGAAGACACATAATTATAA	AAGCTCTACTTTGCTAAGTGAAAT
NDUFV3	yes	2431	chr21:43184006-43186436	+	ENSG00000160194	ENST00000340344	TTTTATTTTAATGCAGCACCTTATATTT	CCGACTGCGCCAATTACAG
NRIP1	yes	2498	chr21:15262386-15264883	-	ENSG00000180530	ENST00000318948	TGAGTCTACCCAAGAGCAAAAA	TTCAATAGAAGTGTTCACAAGGG
OLIG1	yes	2421	chr21:33361876-33364296	+	ENSG00000184221	ENST00000333063	CCTACGGGCCCCACCTAG	ACCCCTTTAAACCCGGCTTG
OLIG2	yes	2483	chr21:33317619-33320101	+	ENSG00000185569	ENST00000333337	GGCACCAGGGCCAAGGGA	GGCTCGGCCGTTTTTATAG
PCBP3	yes	2472	chr21:46138145-46140616	+	ENSG00000183570	ENST00000330561	CGGGCTCCATTCTGAGAAAC	TGAGGGTGGTGGCTTAAGGT
PCNT	yes	2481	chr21:46566034-46568514	+	ENSG00000160299	ENST00000359568	ACTGTCACCTCGACTATAACTG	TTCGCTCTATTTACACTCCCTC
PCP4	yes	2484	chr21:40158769-40161252	+	ENSG00000183036	ENST00000328619	TGAGAAAAATCCCCCTTAAGGTAA	ACACTGCTCAGGGGTCGC
PDXK	yes	2526	chr21:43960899-43963424	+	ENSG00000160209	ENST00000291565	CCCTCCAGGGCTCTCTCCCT	TCCGAACCCGCGAGTTCCGC
PFKL	yes	2506	chr21:44541836-44544341	+	ENSG00000141959	ENST00000312352	CCTCGGGCCTCAGTTTCCC	CGCGCCTTCCGCCCGCCC
PIGP	yes	2525	chr21:37367166-37369690	-	ENSG00000185808	ENST00000360525	GACAGTGAAAATTGGTACAACCACTTC	CCACCTCAGCGCTGAGTTCTCC
POFUT2	yes	2465	chr21:45532267-45534731	-	ENSG00000186866	ENST00000349485	GCCTGAGGTTAAAAGATTACCCC	CCCCGCCCCCGTGAAATG
PPIAL3	yes	2464	chr21:19149473-19151936	+	ENSG00000198618	ENST00000358455	CATTATATTTAGAGGTGCAGAATACA	GCGTCTGCAAAGCCAAGTAA
PRDM15	yes	2528	chr21:42172481-42175008	-	ENSG00000141956	ENST00000269844	CTGATAGCCATGGAAGCTCTTGTCTGTT	CGGACTCCGGGTTGCGATCC
PRED15	yes	2489	chr21:21935609-21938097	+	ENSG00000196190	ENST00000358903	TGTTGGTCTTGCAGTCAATGTTA	GAGTTTAATTGTCCTTACAGTTTTG
PRED58	yes	2502	chr21:45359281-45361782	-	ENSG00000184941	ENST00000328881	CCCATATACATAACAGAAACTCCTCATC	CTCGCACTGACCTGTAGAGA
PRMT2	yes	2475	chr21:46877008-46879482	+	ENSG00000160310	ENST00000355680	TAGCCTCTGTTGCAGTAACTATT	TTGTGGGGAGAGGAAGTTTGG
PRSS7	yes	2515	chr21:18697811-18700325	-	ENSG00000154646	ENST00000284885	CTTTATCACCCTCCATCTCTTCACTTCCC	TTTTGGTTTTGAAGGCTTGCTAATTTAAG
PTTG1IP	yes	2423	chr21:45118173-45120595	-	ENSG00000183255	ENST00000330938	AGGCTGAGGCAGAGGAGTC	GCCTGCGCGATAGAGGAA
PWP2	yes	2481	chr21:44349197-44351677	+	ENSG00000160220	ENST00000291576	GAGAAGGAGCACCGGGTTC	AGACACTTCCCGGGGGTCC
RBM11	yes	2537	chr21:14507854-14510390	+	ENSG00000185272	ENST00000331453	AAAAGAGAGTAAGAGTTGGCTGCACA	AGCAGGGAACATCCTCCGGTCTC
RCAN1	yes	2479	chr21:34909272-34911750	-	ENSG00000159200	ENST00000313806	TACCTTCTCTCTCTGTGGATTTG	GCTCCGCCGTTAACCCCC
RIPK4	yes	2453	chr21:42060321-42062773	-	ENSG00000183421	ENST00000332512	AGGAAACTGGCTTGGAATTCTT	GCTCCGCCCTGTTACGTC
RRP1	yes	2454	chr21:44031450-44033903	+	ENSG00000160214	ENST00000291569	ACTGTGCATCTGACACGCC	CCTGGTACCCAGCACACTGA
RRP1B	yes	2352	chr21:43901496-43903847	+	ENSG00000160208	ENST00000340648	ATGGTGAAACTCCATCTCACTAAA	GGGCTGATGGCGTCACAAA
RSPH1	yes	2429	chr21:42789539-42791967	-	ENSG00000160188	ENST00000291536	GTCACTGCCAGAAAACTATAGTTAT	AGCCTTTGCGTACTGTCATAG
RUNX1	yes	2492	chr21:35343467-35345958	-	ENSG00000159216	ENST00000300305	CAGACTCTTGGGTACAATCCCC	AGTTTTCACACAACCCAAATTACAA
RWDD2B	yes	2449	chr21:29313588-29316036	-	ENSG00000156253	ENST00000286777	GCTATCATACCCAGCCATAATTTT	GACCGGAAGCTGGGGCTC
S100B	yes	2453	chr21:46849396-46851848	-	ENSG00000160307	ENST00000291700	CTTGGCTTCTCCCTTCAATCA	GTCGTCTCTGCTGCGGGT
SAMSN1	yes	2498	chr21:14877546-14880043	-	ENSG00000155307	ENST00000285670	AATAATTTGCTACTAGAAAGTAATGAG	TTGTTTTCTATGGCCAAGGAATG
SETD4	yes	2443	chr21:36354560-36357002	-	ENSG00000185917	ENST00000344232	TTTTGTTTTGTTTTGTTTTGTTTTAGC	GCCCAAGGCACGTGGAAATA
SH3BGR	yes	2468	chr21:39743158-39745625	+	ENSG00000185437	ENST00000333634	CTGGTCTTTGTGAACAGACATAAA	CACTGGACCCTTTCCCTATCAC
SIM2	yes	2522	chr21:36990821-36993342	+	ENSG00000159263	ENST00000290399	TTGGGTTTCTTCCTTCACAATTCCAC	TGTGTGGCTAGGCGCAGCAC
SLC19A1	yes	2520	chr21:45786768-45789287	-	ENSG00000173638	ENST00000311124	TCTCTCCAGCCAAGGCAGAGGA	CTCCGGGACTACAGCGCCCA
SOD1	yes	2463	chr21:31951461-31953923	+	ENSG00000142168	ENST00000270142	GTGGCCTGTTCAGGATCCTATA	CCGAGGACTGCAACGGAAA
SON	yes	2495	chr21:33834772-33837266	+	ENSG00000159140	ENST00000290239	AGCAGTGTAACCTAGGAGATTAAC	CGTCCGCTCCGTTCTCTC
STCH	yes	2478	chr21:14677339-14679816	-	ENSG00000155304	ENST00000285667	TACTTCCCAACCCCAGTGCC	GACTGTACCAGACGTGAGGC
SUMO3	yes	2388	chr21:45062530-45064917	-	ENSG00000184900	ENST00000332859	CGTAGCCCTTCCTGTACTGA	GCTCTGGGCGGTTGGAGA
SYNJ1	yes	2440	chr21:33022115-33024554	-	ENSG00000159082	ENST00000357345	CTGTGTGCGTGTTTAATTTTCCT	CCGGCTTGCTCACCTCTTC
TFF1	yes	2482	chr21:42659723-42662204	-	ENSG00000160182	ENST00000291527	TAAAATCCAGCTGGGCACAGT	CCCGAGCCCCGGATTTTAT
TFF2	yes	2470	chr21:42644180-42646649	-	ENSG00000160181	ENST00000291526	GAAGACGGTGGGTGAAGTGA	GTCCACTGCCGGGTATGTTTT
TFF3	yes	2541	chr21:42608540-42611080	-	ENSG00000160180	ENST00000291525	CACCACTTCATGCTGTCCTGTG	GCTCTGGCAGCCATGACCACC
TIAM1	yes	2485	chr21:31853117-31855601	-	ENSG00000156299	ENST00000286827	CCTGGCCAGCACATTGTTAATA	CGGTCGTCCGGGATCTCC
TMPRSS3	yes	2461	chr21:42689307-42691767	-	ENSG00000160183	ENST00000291532	CATCTGACACACGTGTTAACATG	CCAACGGCTTGCATCAAAATAA
TRPM2	yes	2468	chr21:44595091-44597558	+	ENSG00000142185	ENST00000300482	GAGCCGGGTGCGGTCTGT	TTCCAGACACCAGTGACTACTTT
TTC3	yes	2424	chr21:37364992-37367415	+	ENSG00000182670	ENST00000354749	GTGCAATCTCAGCTCACGG	TGAAGCGAGGTGCGTCAC
UBASH3A	yes	2541	chr21:42694574-42697114	+	ENSG00000160185	ENST00000319294	GTAGCTGGAACTAATGCCACCATGTCTTTT	CCCTTAGGCCAAGAAGCTCGC
USP25	yes	2484	chr21:16021771-16024254	+	ENSG00000155313	ENST00000285681	TGAAGGCCTTCCTCCTGTATC	GCGTTCGCACGTTCCTTTG
WRB	yes	2444	chr21:39671650-39674093	+	ENSG00000182093	ENST00000333781	TGGGGTGTGGTAGTGGACTTC	CAAGCGCCCAGAACAGG
ADARB1	–	2493	chr21:45316499-45318991	+	ENSG00000197381	ENST00000354687	CAGGGAGAGCGTCACAATAG	GCCAACCTGGCCGCTTGG
ADARB1	–	2539	chr21:45316398-45318936	+	ENSG00000197381	ENST00000354687	AACAGCTGGGAACGCCATGTC	CGCCTCAGCGCCTCAGCCC
AGPAT3	–	2281	chr21:44107169-44109449	+	ENSG00000160216	ENST00000327505	TCTGAGTAACTAGGATTACAAGAGA	GCGGCGATCCCACAGTTC
AGPAT3	–	2507	chr21:44106956-44109462	+	ENSG00000160216	ENST00000327505	TTTCTTTACCCTTGCCCATTCACTATC	CCACTTCCTGTGCGCGGCGA
ANKRD21	–	2487	chr21:13901921-13904407	+	ENSG00000166351	ENST00000360495	AGGGATAAGAGCAGAAATAAATGAA	CCAGGGGAGGCCGGTAGT
ANKRD21	–	2530	chr21:13894847-13897376	+	ENSG00000166351	ENST00000360495	GCAAATCTGAGGACAGGAAGAAATTAGGGA	GACGGAAATTTGAACCAAGAGAAATGGG
APP	–	2463	chr21:26464841-26467303	-	ENSG00000142192	ENST00000348990	GCCTCAGCCTCCCCAGTA	CTGTGCGAGTGGGATCCG
B3GALT5	–	2491	chr21:39951101-39953591	+	ENSG00000183778	ENST00000343118	TGGGCAGGCACCATCCATT	GCTTGGTTTGTTTACCAAAAATGG
BRWD1	–	2461	chr21:39607465-39609925	-	ENSG00000185658	ENST00000333229	TTAACTTTCTCAAATCCATCCTAAAATA	CGCGGGAGACGAAAAGTAG
C21orf104	–	2489	chr21:39317109-39319597	-	ENSG00000183412	ENST00000329367	GCCTTGCTCTGTAATAACATGAA	ACCCCCGCCCACAGTGTC
C21orf104	–	2507	chr21:39322896-39325402	-	ENSG00000183412	ENST00000329367	GGGAAGTGGCCAGGCATGAGAG	GAGCAATCTGATGACTCCGTTTCTGTTTC
C21orf105	–	2473	chr21:43218747-43221219	-	ENSG00000196847	ENST00000356939	TGAGGCTCCCCTGAGACAAA	AGCAACACACTGTTTTAATAAGCA
C21orf105	–	2528	chr21:43218757-43221284	-	ENSG00000196847	ENST00000356939	TCCCTCACAATTTGCTCACAGGAAAATTCC	TGTTTTAATAAGCACCTGGATGCAGACAGA
C21orf109	–	2496	chr21:29485849-29488344	+	ENSG00000198966	ENST00000361202	TTCAAATCTGGTTTGACTTCAACA	CAAAGAATCAGTCAATCAAAGATGG

Tabel continued

OVERVIEW Chromosome 21 Promoter Cloning, 0.5 kb Fragments

UCSC Assembly hg18

Mean size: 508
Min size: 217
Max size: 863

HGNC_Symbol	Cloned	Amplicon	Amplicon_Location	Strand	Ensembl_GeneID	Ensembl_TranscriptID	Primer_Sequence_fwd	Primer_Sequence_rev
ABCG1	yes	533	chr21:42508800-42509332	+	ENSG00000160179	ENST00000347800	AATTTCATGAGGGACACGAGGAACTGGCA	CGGAAGGTGAAAGGCATGTC
ADAMTS1	yes	571	chr21:27139564-27140134	-	ENSG00000154734	ENST00000284984	GCTGGGAAGGTGGAGAAGTGGGGT	ATAGCCCCTGGCTCGGAG
ADAMTS5	yes	528	chr21:27260677-27261204	-	ENSG00000154736	ENST00000284987	AAGGGACACAGACACGCCGCTTC	CACACTTGCTTGCAGGATTGA
ATP5J	yes	510	chr21:26029811-26030320	-	ENSG00000154723	ENST00000284971	GACAACCCATTGTTATCTCCTTGGTT	CTCGTTCCGGGGCCTTTTC
BACE2	yes	217	chr21:41461424-41461640	+	ENSG00000182240	ENST00000328735	GCTGACCTCCGAGTCACCCCC	CGGGAAACTCGAGTCCCG
C21orf119	yes	465	chr21:32686881-32687345	+	ENSG00000186837	ENST00000331283	TGCCCGGCACACGGAGGC	CCCAAGACTCTGGATCCCTATC
C21orf123	yes	507	chr21:45669419-45669925	-	ENSG00000183535	ENST00000330082	CTCCACACCTCACAGGGCAGC	GAGGACGGGGTGGGTCTGA
C21orf125	yes	640	chr21:43693708-43694347	+	ENSG00000188660	ENST00000342757	TGCAGCATACAGATCTGGGGGAGA	GGGGCAGAGGGCAATGGAC
C21orf129	yes	491	chr21:42008964-42009454	-	ENSG00000185481	ENST00000331508	GCCATTGACTGGAGGTCAGCAGGT	TCCTCACCTCTGCGTTCCC
C21orf13	yes	523	chr21:39739488-39740010	-	ENSG00000157578	ENST00000358268	TGCATTGTTTTGTCATCAGATTTTG	TGAGGGGGACGACGGCAG
C21orf30	yes	483	chr21:44703742-44704224	+	ENSG00000183637	ENST00000332734	GCAGGTGAACTGTTGCTGGCCT	GGACAGGGCCTCCGGTGT
C21orf42	yes	863	chr21:25725502-25726364	-	ENSG00000185433	ENST00000332587	TCCCAATATATTTATAAGAGGTCTGGGT	TAAAGAAAATGTAGATCAATTTTCAGATC
C21orf45	yes	446	chr21:32573285-32573730	-	ENSG00000159055	ENST00000290130	AGGGAAGATCGCTTAAGGCTAGGAG	AGAAGGGGCAAAAGCCTGAA
C21orf55	yes	704	chr21:33785722-33786425	-	ENSG00000177692	ENST00000314399	GCCTAAGGTGCAAGCAGAAGA	TAGACCATTTGCCACTTTACACT
C21orf62	yes	507	chr21:33107818-33108324	-	ENSG00000159100	ENST00000290186	AGAGAATAATGTCTATCTTTGGAGTTCA	TCTGCCTGGCCACGGATC
C21orf66	yes	494	chr21:33066005-33066498	-	ENSG00000159086	ENST00000331923	AGTCACACCTCAGGTCAAAGTT	CCCACGCGCTCTCAATACT
C21orf77	yes	573	chr21:32870050-32870622	-	ENSG00000186842	ENST00000334165	TCTTGCTGGAAATCGAGGATGGCT	CAAATGGAGAAAGGGGAACAAAG
C21orf88	yes	468	chr21:39906652-39907119	-	ENSG00000184809	ENST00000359876	GGGCGTGGCGTGAGTAGCGATAAG	GCCTCCTCACCTATCAGATTTC
C21orf91	yes	509	chr21:18113528-18114036	-	ENSG00000154642	ENST00000284881	TGACACCTGAATATGAACTTTCCTTCG	GCTCGGGTTCCTCCACTATT
CBR1	yes	563	chr21:36363647-36364209	+	ENSG00000159228	ENST00000290349	TCCCACCATTATTGGAAAATCTGGCT	CAGAGACTGCTCGAGTCTG
CBR3	yes	578	chr21:36428744-36429321	+	ENSG00000159231	ENST00000290354	CTCCCACTTGGGGGCACATTC	ACCGGGCTTCGGACCACC
CCT8	yes	466	chr21:29367838-29368303	-	ENSG00000156261	ENST00000286788	CATTCCCAGCGCTTCCTATGGT	AGGAGCACGCACAGCCTT
CHAF1B	yes	593	chr21:36679007-36679599	+	ENSG00000159259	ENST00000314103	TCTGCCTACCGACTGATTTACAGCC	AGAGGCACCGTCACCTCC
CHODL	yes	464	chr21:18538587-18539050	+	ENSG00000154645	ENST00000299295	TCACTGAGCAGGTTTTGGCTCTCA	TGCGCCCTGGTCCTGGAT
CLDN17	yes	557	chr21:30460811-30461367	-	ENSG00000156282	ENST00000286808	GATTACCATCTTAAGGAAAAGTGGCT	TTTACTTTGAAGACTGGTTCAATTC
CLDN8	yes	547	chr21:30510182-30510728	-	ENSG00000156284	ENST00000286809	TCACATGTGTCCCAAAGAATGTGG	TCTGTGCCACAGAAGAGAACA
CLIC6	yes	404	chr21:34963148-34963551	+	ENSG00000159212	ENST00000360731	CAAGGACTTCTGGTCAACCCCAGG	AGATCCCTGTCCTTGATCGG
COL6A2	yes	713	chr21:46341805-46342517	+	ENSG00000142173	ENST00000300527	CAAGCCCACCTGTGGTTTCCAGC	ACCTGGTCCCGAGGAGGC
CXADR	yes	546	chr21:17806626-17807171	+	ENSG00000154639	ENST00000284878	TCGGTAGCTGCAGAGGGGCTCTATC	AGGTGAACTCGCGTCACTTC
CYYR1	yes	280	chr21:26867490-26867769	-	ENSG00000166265	ENST00000299340	TGCCATACACTTTCTGTACTTCCCAATC	GGGTGGCCCCGACTCCTC
DSCR2	yes	408	chr21:39477262-39477669	-	ENSG00000183527	ENST00000331573	GCCCCTCTAGCCCGCAGTCCT	GCGCGGGCAATAAGTCCC
DSCR9	yes	614	chr21:37502109-37502722	+	ENSG00000182342	ENST00000333494	CTTTGTGTCAAACAGACCTGTCATTG	GGCTCACCCCAGAGATAAAG
FAM3B	yes	432	chr21:41610114-41610545	+	ENSG00000183844	ENST00000357985	GAGGTGGGGCGCCTGCTAG	GGCCAGTCGGGCAAATGG
GART	yes	380	chr21:33837070-33837449	-	ENSG00000159131	ENST00000290230	GGCCTGAGCCGGGGAACGAC	CCCCAGCCGCTCAGAAAC
HEMK2	yes	460	chr21:29179593-29180052	-	ENSG00000156239	ENST00000303775	AACCGCATCATCTTGCTTAAGGTTT	TGCCCTCACCCTCATGCG
HLCS	yes	567	chr21:37284326-37284892	-	ENSG00000159267	ENST00000336648	GGCAATCCCTCAGTTTCTAGTGGGC	ACTTCTGAGGCCGAAACGC
JAM2	yes	503	chr21:25933001-25933503	+	ENSG00000154721	ENST00000312957	TTTAAAGCACGCAACAGAGCATTCA	GGGAAACGCCTCCTGTTTTG
KCNJ6	yes	239	chr21:38210610-38210848	-	ENSG00000157542	ENST00000288309	CATGGAGCACCCACGTTTCCAG	ACGCGGCTCCGAGATAAAA
LIPI	yes	555	chr21:14501128-14501682	-	ENSG00000188992	ENST00000344577	AACATTTGCATAAATGAGTTGCTGA	ATCAGAAGTTCACTAGCTCTTTG
LRRC3	yes	510	chr21:44699263-44699772	+	ENSG00000160233	ENST00000291592	TGGGCGGAGATCTAGCCCTACTCA	GGCGAAAGGCGGAGACAG
MRPL39	yes	478	chr21:25901697-25902174	-	ENSG00000154719	ENST00000352957	CCCACTCCAGCCGAACTAAGAACC	CCCTCCTCCCCCGGAAAC
MX2	yes	499	chr21:41655309-41655807	+	ENSG00000183486	ENST00000330714	GATCATGTACCTGAGAGCATGTCC	CTTCACTTTTATGTGGCATATGAAC
OLIG1	yes	492	chr21:33363805-33364296	+	ENSG00000184221	ENST00000333063	ATGGGTGGAGCTTCAGGGTTGC	ACCCCTTTAAACCCGGCTTG
OLIG2	yes	452	chr21:33319650-33320101	+	ENSG00000185569	ENST00000333337	GCTGCGTGGGGCATCAGGTTAGT	GGCTCGGCCGGTTTTTATAG
PCP4	yes	490	chr21:40160763-40161252	+	ENSG00000183036	ENST00000328619	AAGGGAAACACAGCTCCTGGGAA	ACACTGCTCAGGGGTCGC
PRSS7	yes	564	chr21:18697846-18698409	-	ENSG00000154646	ENST00000284885	GCAACACTTTTGATTGTTTCCATAAG	AAGAGAATATAAATAATTCTACCAACTG
PTTG1IP	yes	439	chr21:45118173-45118611	-	ENSG00000183255	ENST00000330938	GGACATCGGCGGGAGCCCAG	GCCTGCGCGATAGGAGGAA
PWP2	yes	446	chr21:44351232-44351677	-	ENSG00000160220	ENST00000291576	GTCATCCACCCGCCTCAGTC	AGACACTTCCCGGGGGTCC
RBM11	yes	509	chr21:14509828-14510336	+	ENSG00000185272	ENST00000331453	CAAAAACTCATCTAGGGAGCACTTTCT	GCCCCTTCTCCGGCTTTAC
RCAN1	yes	486	chr21:34909272-34909757	-	ENSG00000159200	ENST00000313806	CAAAAGTGGGCCCCTGTGCTTCAA	GCTCCGCCGTTAACCCCC
RWDD2B	yes	487	chr21:29313588-29314074	-	ENSG00000156253	ENST00000286777	CTGGGAAATGACAACGGGGTTGGCA	GACCGGAAGCTGGGGCTC
SETD4	yes	582	chr21:36354560-36355141	-	ENSG00000185917	ENST00000344232	AGCCCATAATCAAAGTGGGGGTGG	GCCCAAGGCACGTGGAAATA
SH3BGR	yes	433	chr21:39745193-39745625	+	ENSG00000185437	ENST00000333634	GCCCATGGCTGGTCAGTTTCTGTT	CACTGGACCCTTTCCCTATCAC
SIM2	yes	536	chr21:36992807-36993342	+	ENSG00000159263	ENST00000290399	TTTCCCCTTTTCAGCTCTTGGACCTGCA	TGTGTGGCTAGGCGCAGCAC
SLC19A1	yes	621	chr21:45786763-45787383	-	ENSG00000173638	ENST00000311124	GTCAGGCAGGAGGGCGGG	GCGGACTCCGGGACTACA
SOD1	yes	604	chr21:31953320-31953923	+	ENSG00000142168	ENST00000270142	CCGAATTCTGCCAACCAAATAAGAAACT	CCGAGGACTGCAACGGAAA
SON	yes	496	chr21:33836771-33837266	+	ENSG00000159140	ENST00000290239	GGGGGCGGGGGCTGAAA	CGTCCGCTCCGTTCTCTC
STCH	yes	565	chr21:14677339-14677903	-	ENSG00000155304	ENST00000285667	GACCCTTCTACGGAGGATGGCTGG	GACTGATCCAGACGTGAGGC
SYNJ1	yes	465	chr21:33022115-33022579	-	ENSG00000159082	ENST00000357345	CTCCTCTCCCGTCAGGCCCGTG	CCGGCTTGCTCACCTCTTC
TTC3	yes	429	chr21:37366987-37367415	+	ENSG00000182670	ENST00000354749	GGGCAGGGAACAAGGCTCGC	TGAAGCGAGGTGCGTCAC
USP25	yes	574	chr21:16023681-16024254	+	ENSG00000155313	ENST00000285681	TTTCCCTTCAGGCAAGTTCAAGGC	GCGTTCGCACGTTCCTTTG
WRB	yes	420	chr21:39673674-39674093	+	ENSG00000182093	ENST00000333781	CTGGAACGAAAACGTAAGGC	CAAGCGCCCCAGAACAGG
ABCC13		448	chr21:14567579-14568026	+	ENSG00000155288	ENST00000318742	CCCCTTTGTACATCATGTGTCACCTTT	GGCCTGGAGGAGGCAAGG
ADARB1		529	chr21:45318463-45318991	+	ENSG00000197381	ENST00000354687	AACCTGGCCGCTTGGCCG	GCCAACCTGGCCGCTTGG
AGPAT3		331	chr21:44109119-44109449	+	ENSG00000160216	ENST00000327505	GGGACCCCAGCCTGGTCCTTC	GCGGCGATCCCACAGTTC
AIRE		446	chr21:44529726-44530171	+	ENSG00000160224	ENST00000291582	CCTCGCNTGGCCTCGGGT	CGCGCCGCTGTTATACCT
APP		697	chr21:26464841-26465537	-	ENSG00000142192	ENST00000348990	TCCAGGTTGCTCGTGCCTGCTTTT	CTGTGCGAGTGGGATCCG
ATP5O		445	chr21:34210007-34210451	-	ENSG00000159186	ENST00000290299	TAACAATGGGTAGTCATGGCCCAG	TGGCGTAACCGCTAGGTTC
BTG3		495	chr21:17907049-17907543	-	ENSG00000154640	ENST00000339775	CCCCGGGTCTAGAGAGCTGGGTCTA	CGCGTTGAGAGGACTGGC
C21orf105		545	chr21:43218747-43219291	-	ENSG00000196847	ENST00000356939	CCTGCAACCATGCCCGGCT	AGCAACACACTGTTTTAATAAGCA
C21orf122		646	chr21:45317370-45318015	-	ENSG00000183821	ENST00000333073	CTGTTCCACAGTTATACCAAGAAATGTT	CCTGACCTGTCCGGGACTG
C21orf2		464	chr21:44583683-44584146	-	ENSG00000160226	ENST00000339818	TCCTGAGCCNATTGGCGGCT	CCGCTGACACGTTGGCTC
C21orf25		478	chr21:42246843-42247320	-	ENSG00000157617	ENST00000329623	ACTGCGCCAGGTACAGGCCTACC	GCCTCGCCCCAACTTCCC
C21orf33		516	chr21:44377444-44377959	+	ENSG00000160221	ENST00000291577	GTGAGGACAGCNGGGTCGGAAG	GCACCCACCCCTCTACAAAC
C21orf56		591	chr21:46428691-46429281	-	ENSG00000160284	ENST00000330205	GNGTCTTGTTTTATCAGCAGCCGGG	AAATCACCTCAGTGTTGTTTCTTG
C21orf57		534	chr21:46530235-46530768	+	ENSG00000182362	ENST00000339195	GCTTGGGAGACTANAAACCCGCCC	CCCGGACGTACCGGTGTC
C21orf67		470	chr21:45184211-45184680	-	ENSG00000183250	ENST00000332379	TGTTTCTNGTGCTCCGCGCCA	AGGTGCGTCGGTGAAAGG
C21orf70		438	chr21:45183922-45184359	+	ENSG00000160256	ENST00000291634	GCCTCACGGCAGCCTGGTG	AAGCTGGTCCTGCGTTCTC
C21orf84		574	chr21:43722444-43723017	-	ENSG00000185186	ENST00000332440	GGCAGGTGGGGGATAAGTAGCCAA	CACCTTTGCCCTGGCTATGTA
C21orf90		416	chr21:44761101-44761516	+	ENSG00000182912	ENST00000330490	GAGACNTGTGGGAGCCCTGG	AACTGGTATGTGTCCTAGAGAGA
COL6A1		521	chr21:46225605-46226125	+	ENSG00000142156	ENST00000323891	ACCACACAGCGCGGCCAG	CGAGGCTGCCTTCTGCTC
CRYAA		544	chr21:43461708-43462251	+	ENSG00000160202	ENST00000291554	CAGCTTTGGTACCACGGGGACCTG	GAGGCTGGCAGGAGTCAG
CSTB		475	chr21:44020644-44021118	-	ENSG00000160213	ENST00000291568	GATGTNCTGGGTCTCGGCGGTG	AAGTAGGCGCTGGGGTCAC
DNMT3L		476	chr21:44506479-44506954	-	ENSG00000142182	ENST00000270172	GGTTAAGAGCTCAGTGGGGCTGGG	TTTGGGGACCTCAGTGGGGA
FTCD		486	chr21:46399933-46400418	-	ENSG00000160282	ENST00000355384	GAAAAGTTGGGGACGCATTCCACC	GCCGCCAGGGCCTTTATAC
H2BFS		547	chr21:43808975-43809521	+	ENSG00000197597	ENST00000361105	AGAGCCCCATCCCTGCAGTA	CGAGCAGCAGATCGTGAAA
HMGN1		450	chr21:39642871-39643320	-	ENSG00000205581	ENST00000288344	AAATGTTCGAACCCAATTCACCCC	TGGGAGAACCGGATGGAACC
HSF2BP		385	chr21:43903829-43904213	-	ENSG00000160207	ENST00000291560	ACAGCCTGAATTTTGGCAACCGAA	TTTGTGACGCCATCAGCCC
HUNK		282	chr21:32167207-32167488	+	ENSG00000142149	ENST00000270112	AGGGATTTGAAAGTGCACGGGCTG	GGTCCGGGTGCCTGGGTC
ICOSLG		590	chr21:44485234-44485823	-	ENSG00000160223	ENST00000291579	CTNTAACTGCGCTCCCGCCGC	CGGGAGATCGGCTCTAACT
IFNAR1		483	chr21:33618646-33619128	+	ENSG00000142166	ENST00000270139	CTGGGGACTCCCAACGCCACT	GCTGCCCCTCTTAGCTCTC
ITGB2		467	chr21:45173136-45173602	-	ENSG00000160255	ENST00000355153	GCATGCCACTCTTCTTTCTG	GGATGTGTCTCCAAGAAAAGTGT
KCNJ15		221	chr21:38550309-38550529	+	ENSG00000157551	ENST00000328656	TTCCNCCAGGAGGCAAGA	CAAAAGAAAACTGATGCCTTTTTAAA
LSS		555	chr21:46473123-46473677	-	ENSG00000160285	ENST00000356396	CNGCAGTGCTCTACGCCG	TCAGAGAGGCTCACCTCAAAG
MCM3AP		428	chr21:46529668-46530095	-	ENSG00000160294	ENST00000291688	TGACCAAATCGAAATGGCGNCTTA	GTACAAAATTAATTGGCTTCCTGAAA
MCM3APAS		528	chr21:46473095-46473622	+	ENSG00000187155	ENST00000333280	GCAGTGCTCTACGCCGCC	TGCCAGATGTCCGCACCC
MORC3		533	chr21:36613889-36614421	+	ENSG00000159256	ENST00000290384	CGCCCGGCCCNAGAACTCTTTTT	TACAAAGCCAGGAACGGCC
NDUFV3		453	chr21:43185984-43186436	+	ENSG00000160194	ENST00000340344	TCCTTGCCGCAGCAAACACG	CCGACTGCGCCAATTACAG
PDE9A		700	chr21:42946276-42946975	+	ENSG00000160191	ENST00000335512	CCGCAGGGTCTCCTCCGC	ACTCGGACTTTTTACTGTACTTTC
PDXK		556	chr21:43962899-43963454	+	ENSG00000160209	ENST00000291565	CGCAGACGCCTCCTTCTGACCT	AGACGCCTCCTTCTGACCT
PFKL		528	chr21:44543871-44544398	+	ENSG00000141959	ENST00000312352	GCAGCTTCTCCAGGTCCACCGC	GGCCCAGCTCGCACTCCC
POFUT2		486	chr21:45532267-45532752	-	ENSG00000186866	ENST00000349485	AGAAGCCGGAGGCCAGGACACT	CCCCGCCCCCGTGAAATG
PPIAL3		544	chr21:19151393-19151936	+	OTTHUMG00000007455	ENST00000358455	TGCAGAATAGATCACTGAATATGTCCAA	GCGTCTGCAAAGCCAAGTAA
PRDM15		435	chr21:42172612-42173046	-	ENSG00000141956	ENST00000269844	CCTTTGGAATGTGCAGAGCGGG	GGTCCTGCGGGTGGGACT
RIPK4		461	chr21:42060321-42060781	-	ENSG00000183421	ENST00000332512	GGCCAGGTGGGGGTGGNTAC	GCTCCGCCCTGTTACGTC
S100B		543	chr21:46849396-46849938	-	ENSG00000160307	ENST00000291700	TCTTCCTTGTCTCACCCCTTC	GTCGTCTCTGCTGCGGGT
SFRS15		453	chr21:32026249-32026701	-	ENSG00000156304	ENST00000286835	TTAAATCCTGCGCCCTTTGTTCGC	TCTCTCCAGCGGGATGGC
SUMO3		354	chr21:45062530-45062883	-	ENSG00000184900	ENST00000332859	GGAGCNGGCGAGTCACGCTCT	GCTCTGGGCGGTTGGAGA
TFF1		430	chr21:42659723-42660152	-	ENSG00000160182	ENST00000291527	ACCAGGGCGCAGATCACCTTGTT	CCCGAGCCCCGGATTTTAT
TFF2		438	chr21:42644180-42644617	-	ENSG00000160181	ENST00000291526	CAGGGCACANAGCCCCAGGAC	GTCCACTGCCGGGTATGTTTT
TFF3		333	chr21:42608739-42609071	-	ENSG00000160180	ENST00000291525	CTGGAGGACANCAAGGCCAGGA	AAAGGGGAGAGGTCAGTTCAG
TMEM1		354	chr21:44256266-44256619	+	ENSG00000160218	ENST00000291574	AGCCCCCAGCGGCGATAAGGAG	CCGCCGCATGACCATTTC
TMEM50B		547	chr21:33774064-33774610	-	ENSG00000142188	ENST00000270178	AGCACCNTCCCTACTCTGGGCA	AGACTGCTGCGCCACAAC
TMPRSS2		532	chr21:41801928-41802459	-	ENSG00000184012	ENST00000332149	CCTAGGCGGACTGGGGAGGGTC	CCGCCTCCTGCTTAGCTC
TPTE		448	chr21:10012778-10013225	-	ENSG00000166157	ENST00000361285	GCACAGCTGCCTTCTAACCCGC	CCGCCCCCTGCCTTGAAC
U2AF1		456	chr21:43400740-43401195	-	ENSG00000160201	ENST00000291552	CCGAAGATGGAGGCCAGATACTCC	GGGAGACGTCACCCGGAC
UBASH3A		412	chr21:42696638-42697049	+	ENSG00000160185	ENST00000319294	AGCTTGTTGGAGACCTTGGCGTAG	GGCTGAGGCTGTGAAGAAAAA
USP16		519	chr21:29318256-29318774	+	ENSG00000156256	ENST00000334352	GCGNATCACAAGGTCAGG	ATCGGCTCTGACGCTAGTAA
ZNF295		474	chr21:42303521-42303994	-	ENSG00000173276	ENST00000310826	CCTTTCCGCTCACACTCGGCTC	CATCTTGAATTTGACGTCATCCC

HGNC Symbol	Activity	RNA-seq expression HEK	+TSA activity	-FCS activity	Activated by TSA	Activated by -FCS	RNA-seq expression HEK	Normalized expression	RNA-seq reads HEK	UniGene tissue hits	Pol2 bind site	CpG island	TATA box	INR element	DPE element
SUM	56	87	81	96	28	40	87+16+79		87+16	156	47	83	14	12	70
ADAMTS1	1	1	1	1	0	0	YES	0,120904	535	34	1	1	1	0	1
ATP5J	1	1	1	1	0	0	YES	0,414027	366	39	0	1	0	0	0
C21orf119	1	1	1	1	0	0	YES	0,056962	45	15	1	1	0	0	0
C21orf13	1	0	1	1	0	0	uncertain	0,000416	1	17	1	1	0	1	1
C21orf45	0	1	1	1	1	1	YES	0,148014	205	32	1	1	0	0	0
C21orf55	1	1	1	1	0	0	YES	0,006235	13	9	1	1	0	0	0
C21orf91	1	1	1	1	0	0	YES	0,013748	72	26	1	1	0	0	1
CBR1	1	1	1	1	0	0	YES	0,20372	230	37	1	1	0	0	0
CCT8	1	1	1	1	0	0	YES	0,876969	891	43	1	1	0	0	0
CXADR	1	1	1	1	0	0	YES	0,038003	51	33	0	1	0	1	1
GART	1	1	1	1	0	0	YES	0,378117	1289	40	1	1	0	0	1
LRRC3	1	1	1	1	0	0	YES	0,013555	31	3	0	0	0	0	0
OLIG2	1	1	1	1	0	0	YES	0,008155	27	5	0	1	1	0	0
RCAN1	1	1	1	1	0	0	YES	0,047538	139	41	0	1	0	0	1
SETD4	1	1	1	1	0	0	YES	0,053744	150	28	1	1	0	0	0
SOD1	1	1	1	1	0	0	YES	1,001773	565	44	1	1	0	0	1
SON	1	1	1	1	0	0	YES	0,402818	3231	43	1	1	0	0	1
STCH	1	1	1	1	0	0	YES	0,15698	605	34	1	1	0	0	1
SYNJ1	1	1	1	1	0	0	YES	0,024749	155	28	1	1	0	1	0
TTC3	1	1	1	1	0	0	YES	0,281191	1199	44	1	1	0	0	1
PWP2	1	1	1	1	0	0	YES	0,203471	551	36	1	1	0	0	0
CHAF1B	0	1	1	1	1	1	YES	0,227021	410	27	0	1	0	0	1
COL6A2	1	1	1	1	0	0	YES	0,065606	235	35	0	1	1	0	0
HLCS	0	1	1	0	1	2	YES	0,03703	209	34	1	1	0	0	0
C21orf66	1	1	0	1	0	0	YES	0,102202	376	32	1	1	0	0	1
RBM11	0	1	0	1	2	1	YES	0,01118	18	8	0	1	0	0	0
ATP5O	1	1	1	1	0	0	YES	0,643159	456	42	1	0	0	0	1
BACH1	1	1	1	1	0	0	YES	0,050523	280	28	1	1	0	0	0
BTG3	1	1	1	1	0	0	YES	0,127176	168	32	1	1	0	0	0
C21orf115	1	0	1	1	0	0	no	0	0	0	0	0	0	0	1
C21orf33	1	1	1	1	0	0	YES	0,349788	248	35	1	1	0	0	1
C21orf56	1	1	1	1	0	0	YES	0,135906	162	24	0	0	0	0	0
C21orf57	1	1	1	1	0	0	YES	0,072905	67	27	1	1	0	0	1
C21orf58	1	1	1	1	0	0	YES	0,094927	247	28	1	1	0	0	0
C21orf67	1	1	1	1	0	0	YES	0,020788	19	7	1	1	0	0	1
C21orf70	1	1	1	1	0	0	YES	0,137821	43	21	1	1	0	0	0
C21orf86	1	1	1	1	0	0	YES	0,050847	24	14	0	0	0	0	1
CBS	1	1	1	1	0	0	YES	1,30965	2470	32	0	1	0	0	0
DIP2A	1	1	1	1	0	0	YES	0,089725	572	31	1	1	0	0	1
DONSON	1	1	1	1	0	0	YES	0,121648	245	26	1	1	0	0	0
DSCR4	1	0	1	1	0	0	no	0	0	3	0	0	0	0	1
DSCR8	1	0	1	1	0	0	no	0	0	6	0	0	0	0	1
HMGN1	1	1	1	1	0	0	YES	0,046753	18	45	1	1	0	0	1
IFNAR2	1	1	1	1	0	0	YES	0,017937	68	27	0	1	0	0	1
MCM3AP	1	1	1	1	0	0	YES	0,194194	970	34	1	1	0	0	1
NDUFV3	1	1	1	1	0	0	YES	0,173273	306	33	1	0	0	0	1
PCNT	1	1	1	1	0	0	YES	0,134176	1221	33	1	1	0	0	0
PIGP	1	1	1	1	0	0	YES	0,068627	63	28	1	1	0	0	0
POFUT2	1	1	1	1	0	0	YES	0,034321	154	32	1	1	0	0	1
PRDM15	1	1	1	1	0	0	YES	0,016506	65	24	0	1	0	0	0
RIPK4	1	1	1	1	0	0	YES	0,028813	109	26	0	1	0	1	1
RRP1	1	1	1	1	0	0	YES	0,243955	1140	36	1	1	0	0	1
RSPH1	1	0	1	1	0	0	no	0	0	11	0	1	0	0	0
TIAM1	1	1	1	1	0	0	YES	0,04618	217	30	1	1	0	0	0
KRTAP19-	1	1	0	1	0	0	YES	0,271186	48	1	0	0	0	0	0
KRTAP21-	1	0	0	1	0	0	no	0	0	0	0	0	0	0	0
PTTG1IP	1	1	1	1	0	0	YES	0,342248	883	45	0	1	0	0	0
KCNJ6	0	0	1	1	1	1	no	0	0	5	0	1	0	0	1
MRPL39	0	1	1	1	1	1	YES	0,097561	84	34	1	1	0	0	1
CHODL	0	0	0	0	2	2	uncertain	0,001678	4	10	0	1	0	1	1
SH3BGR	0	1	0	0	2	2	YES	0,015422	19	18	0	0	0	1	0
OLIG1	1	1	1	1	0	0	YES	0,003579	8	4	0	1	0	0	1
SIM2	1	1	1	1	0	0	YES	0,039689	143	11	0	1	0	0	0
SLC19A1	0	1	1	1	1	1	YES	0,028754	171	31	0	1	0	0	1
USP25	0	1	1	1	1	1	YES	0,058357	265	33	1	1	0	0	0
C21orf88	0	0	1	1	1	1	no	0	0	4	0	1	0	0	0
DSCR2	1	1	1	1	0	0	YES	0,188242	349	30	1	1	0	0	0
MX2	0	0	1	1	1	1	no	0	0	30	0	0	0	0	1
RWDD2B	0	1	1	1	1	1	YES	0,124841	196	36	1	1	0	0	1
CBR3	0	1	1	0	1	2	YES	0,009606	10	19	0	1	0	0	1
BACE2	0	1	1	0	1	2	YES	0,074951	114	38	0	1	0	0	0
PCP4	0	0	1	0	1	2	no	0	0	12	0	0	0	0	0
ADAMTS5	0	0	0	1	2	1	no	0	0	23	0	1	0	0	0
C21orf30	0	1	0	1	2	1	YES	0,008039	18	8	0	0	0	0	1
C21orf123	0	0	0	1	2	1	uncertain	0,000443	1	1	0	0	0	0	1
C21orf125	0	0	0	1	2	1	no	0	0	9	0	0	0	0	0
C21orf129	0	0	0	1	2	1	no	0	0	3	0	0	0	0	0
C21orf62	0	0	0	0	2	2	no	0	0	9	0	0	0	0	1
WRB	0	1	0	0	2	2	YES	0,114112	169	35	1	1	1	0	0
ABCG1	0	1	0	0	2	2	YES	0,003436	12	26	0	0	0	0	1
C21orf42	0	0	0	0	2	2	no	0	0	3	0	0	0	1	0
C21orf77	0	0	0	0	2	2	no	0	0	3	0	0	0	1	0
CLDN17	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
CLDN8	0	0	0	0	2	2	no	0	0	9	0	0	0	0	0
CLIC6	0	0	0	0	2	2	no	0	0	14	0	1	0	0	1
CYYR1	0	0	0	0	2	2	no	0	0	26	0	1	0	0	0
DSCR9	0	0	0	0	2	2	no	0	0	6	0	0	0	0	0
FAM3B	0	0	0	0	2	2	no	0	0	15	0	1	0	0	0
HEMK2	0	1	0	0	2	2	YES	0,039112	37	18	1	1	1	0	0

HGNC Symbol	Activity	RNA-seq expression HEK	+TSA activity	-FCS activity	Activated by TSA	Activated by -FCS	RNA-seq expression HEK	Normalized expression	RNA-seq reads HEK	UniGene tissue hits	Pol2 bind site	CpG island	TATA box	INR element	DPE element
JAM2	0	1	0	0	2	2	YES	0,005865	8	19	0	1	0	0	0
LIPI	0	0	0	0	2	2	no	0	0	4	0	0	0	0	0
PRSS7	0	0	0	0	2	2	no	0	0	5	0	0	0	0	0
C21orf87	0	0	1	1	1	1	no	0	0	5	0	1	0	0	1
ETS2	0	1	1	1	1	1	YES	0,040913	138	42	0	1	0	0	0
IFNGR2	0	1	1	1	1	1	YES	0,105623	216	36	0	1	0	0	1
KRTAP11-	0	0	1	1	1	1	no	0	0	0	0	0	1	0	0
AIRE	0	0	1	1	1	1	no	0	0	2	0	1	0	1	0
C21orf25	0	0	1	1	1	1	uncertain	0,000173	1	36	0	1	0	0	0
GRIK1	0	0	1	1	1	1	no	0	0	10	0	1	0	0	1
KRTAP13-.	0	0	1	1	1	1	no	0	0	0	0	0	0	0	0
KRTAP19-:	0	0	1	1	1	1	uncertain	0,00625	1	0	0	0	1	0	0
RUNX1	0	1	1	1	1	1	YES	0,003755	30	33	0	0	1	0	1
CLDN14	0	0	1	0	1	2	no	0	0	3	0	0	0	0	1
C21orf2	0	1	1	0	1	2	YES	0,020504	114	27	1	1	0	0	1
C21orf29	0	0	1	0	1	2	uncertain	0,000273	1	2	0	0	0	0	0
KRTAP13-;	0	0	1	0	1	2	no	0	0	1	0	0	0	0	0
RRP1B	0	1	1	0	1	2	YES	0,163025	388	30	1	1	0	0	0
B3GALT5	0	0	0	1	2	1	no	0	0	7	0	0	0	1	0
GABPA	0	1	0	1	2	1	YES	0,048804	100	27	0	1	0	0	1
H2BFS	0	0	0	1	2	1	no	0	0	2	0	0	0	0	0
TFF1	0	0	0	1	2	1	no	0	0	10	0	0	0	1	1
UBASH3A	0	0	0	1	2	1	no	0	0	11	0	0	0	0	0
C21orf93	0	0	0	1	2	1	no	0	0	4	0	0	0	0	1
FTCD	0	0	0	1	2	1	uncertain	0,001888	3	9	0	0	1	0	0
HSF2BP	0	1	0	1	2	1	YES	0,005409	9	10	1	1	0	0	1
KCNE1	0	0	0	1	2	1	no	0	0	12	0	0	0	0	0
MRAP	0	0	0	1	2	1	no	0	0	0	0	0	0	0	0
MX1	0	0	0	1	2	1	no	0	0	39	0	0	0	0	0
PRED58	0	0	0	1	2	1	uncertain	0,001669	1	2	0	0	0	0	1
PRMT2	0	1	0	1	2	1	YES	0,167018	475	42	0	0	0	0	1
SuMO3	0	1	0	1	2	1	YES	0,376982	642	41	0	1	0	1	1
C21orf19	0	1	0	0	2	2	YES	0,025253	10	0	0	0	0	0	0
DOPEY2	0	1	0	0	2	2	YES	0,020143	130	25	0	0	0	0	0
TFF3	0	0	0	0	2	2	no	0	0	21	0	0	0	0	0
TMPRSS3	0	0	0	0	2	2	no	0	0	15	0	0	0	0	0
ABCC13	0	0	0	0	2	2	no	0	0	2	0	0	0	0	1
C21orf100	0	0	0	0	2	2	no	0	0	1	0	0	0	0	0
C21orf109	0	0	0	0	2	2	no	0	0	5	0	0	0	0	1
C21orf110	0	0	0	0	2	2	no	0	0	4	0	0	0	0	0
C21orf114	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
C21orf116	0	0	0	0	2	2	uncertain	0,000465	1	0	0	0	0	0	0
C21orf120	0	0	0	1	2	1	no	0	0	0	0	0	0	0	0
C21orf121	0	0	0	0	2	2	uncertain	0,00114	1	3	0	1	0	0	0
C21orf128	0	0	0	0	2	2	no	0	0	1	0	0	0	0	0
C21orf15	0	0	0	0	2	2	no	0	0	10	0	0	0	0	0
C21orf32	0	0	0	0	2	2	uncertain	0,00678	2	4	0	0	0	0	1
C21orf34	0	0	0	0	2	2	uncertain	0,00333	3	21	0	0	0	0	1
C21orf41	0	0	0	0	2	2	no	0	0	7	0	0	0	0	0
C21orf74	0	0	0	0	2	2	no	0	0	3	0	0	0	0	0
C21orf82	0	0	0	0	2	2	no	0	0	0	0	0	0	0	1
C21orf84	0	0	0	0	2	2	no	0	0	6	0	0	0	0	1
C21orf9	0	0	0	0	2	2	no	0	0	3	0	0	0	0	0
C21orf90	0	1	0	0	2	2	YES	0,020529	52	4	0	0	0	0	0
C21orf94	0	0	0	0	2	2	no	0	0	2	0	0	0	0	0
C21orf99	0	1	0	0	2	2	YES	0,238024	159	2	0	0	0	0	0
CRYAA	0	0	0	0	2	2	no	0	0	7	0	0	0	0	0
DNMT3L	0	0	0	0	2	2	uncertain	0,000718	1	2	0	0	0	0	0
DSCAM	0	0	0	0	2	2	no	0	0	5	0	0	0	0	0
DSCR10	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
DSCR6	0	0	0	0	2	2	no	0	0	1	0	1	0	0	0
ERG	0	0	0	0	2	2	uncertain	0,00034	1	26	0	1	0	0	1
FGF7P2	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
IGSF5	0	0	0	0	2	2	no	0	0	7	0	0	0	0	0
IL10RB	0	1	0	0	2	2	YES	0,052903	82	37	1	1	0	0	1
ITGB2	0	0	0	0	2	2	no	0	0	35	0	0	0	0	0
KCNE2	0	0	0	0	2	2	uncertain	0,002674	2	4	0	0	0	0	1
KRTAP13-	0	0	0	0	2	2	no	0	0	0	0	0	1	0	0
KRTAP13-;	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
KRTAP15-	0	0	0	0	2	2	no	0	0	0	0	0	0	0	1
KRTAP19-;	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
KRTAP19-;	0	0	0	0	2	2	no	0	0	1	0	0	0	0	0
KRTAP19-(	0	0	0	0	2	2	no	0	0	0	0	0	0	0	1
KRTAP19-	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
KRTAP20-	0	0	0	0	2	2	no	0	0	0	0	0	1	0	0
KRTAP20-;	0	0	0	0	2	2	no	0	0	0	0	0	1	0	0
KRTAP22-	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
KRTAP23-	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
KRTAP6-1	0	0	0	0	2	2	no	0	0	0	0	0	1	0	0
KRTAP6-2	0	0	0	0	2	2	no	0	0	0	0	0	1	0	0
KRTAP8-1	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
MRPS6	0	1	0	0	2	2	YES	0,023414	358	42	0	0	0	0	0
NCAM2	0	1	0	0	2	2	YES	0,001206	5	13	0	0	0	0	0
NRIP1	0	1	0	0	2	2	YES	0,035642	123	34	0	0	0	0	1
PCBP3	0	1	0	0	2	2	YES	0,010997	16	17	0	0	0	0	0
PDXK	0	1	0	0	2	2	YES	0,029774	266	36	1	1	0	0	0
PFKL	0	1	0	0	2	2	YES	0,308538	842	41	0	1	0	0	1
PPIAL3	0	1	0	0	2	2	YES	2,872093	247	0	0	0	0	0	1
PRED15	0	0	0	0	2	2	no	0	0	0	0	0	0	0	0
S100B	0	0	0	0	2	2	no	0	0	18	0	0	0	0	1
SAMSN1	0	0	0	0	2	2	no	0	0	25	0	0	0	0	0
TFF2	0	0	0	0	2	2	uncertain	0,002	1	6	0	0	0	0	0
TRPM2	0	0	0	0	2	2	no	0	0	22	0	0	0	0	1

6.3. Contributions

Hans-Jörg Warnatz: optimized the cloning of reporter constructs, performed amplification and cloning of reporter constructs, primer design, contributed to conceptualization, to data analysis and was pivotal in writing the manuscript

Anna Guerasimova: was involved in experimental work with promoter construct preparation, sequencing and cell arrays and contributed to image analysis

Xi Cheng: performed part of the cell array experiments and was involved in plasmid preparation

Dominique Vanhecke: was pivotal in initial optimization of the cell array experiments and reporter constructs design

Andrew Hufton: implemented the tool for TFBSs enrichment analysis

Stefan Haas: performed a major part of the primer design for 2.5kb fragments

Wilfried Nietfeld: was involved in the generation of promoter-reporter constructs

Martin Vingron: was involved in conceptualization of this study

Michal Janitz: unknown

Marie-Laure Yaspo: was involved in conceptualization, supervision of the study and writing of the manuscript

Hans Lehrach: conceptualized and supervised this study

7. Manuscript III

7.1. Discovery of human-specific functional transcription factor binding sites by ChIP-seq and comparative genomics

Discovery of human-specific functional transcription factor binding sites by ChIP-seq and comparative genomics

Robert Querfurth[1]*, Hans-Jörg Warnatz[1], Robert Querfurth[1]*, Ralf Sudbrak[1], Hans Lehrach[1] and Marie-Laure Yaspo[1]

[1]Department for Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

*To whom correspondence should be addressed. Tel: +49 30 8413 1225; Email: querfurt@molgen.mpg.de

Background

Phenotypic differences of closely related species such as human and chimpanzee are most likely caused by differences in gene regulation. This has been postulated first over 30 years ago, however still only a handful of verified examples exist. To find *cis*-regulatory adaptations on the lineage leading to human, we performed ChIP-seq of the transcription factor GABPa in HEK293 cells of human origin. We explored the enriched regions for GABPa binding sites (BSs), and based on multiple species alignments, we searched for BSs that were fixed during hominid evolution on the lineage leading to human. To clarify the transcriptional impact of such lineage-specific sites, we performed promoter-reporter gene assays of wild type and mutated promoters in HEK293 and COS-1 cells. Human mutated promoter-reporter constructs were modified by one or two single nucleotide mutations to mimic the ancestral state devoid of the GABPa BS. On the other hand, chimpanzee and rhesus constructs were modified to mimic the human-specific GABPa binding site.

Results

We identified 11,619 GABPa BSs within 5,797 of the 6,208 regions bound by GABPa as determined by ChIP-seq. 224 GABPa BSs are specific to human, while another 53 have been fixed before the split of human and chimpanzee. We selected and cloned four gene promoters with sites specific to human and one promoter with BSs specific to both human and chimpanzee. Reversion of human BSs to the ancestral states resulted in significantly lower reporter-gene activities compared to the wild type in three of the five cases, while mimicking the human BS in chimpanzee and rhesus led to significantly increased reporter gene activities in all cases.

Conclusion

Our analysis shows that ChIP-seq data can be used to identify linage-specific transcription factor binding sites (TFBSs) of functional relevance. Functional promoter analysis shows that the promoters of ZNF398, ZNF425, ZNF197 and ANTXR1 are differently regulated in human and chimpanzee, while the TMBIM6 promoter gained a functional GABPa BS in hominids. The rapidly increasing amount of transcription factors (TFs) being analyzed by ChIP as well as genomes being sequenced will allow for new insights into whole regulatory pathway adaptations, and understanding will further advance by incorporating gene ontology (GO)

annotation and expression profiles. We demonstrate here that TF-ChIP-seq combined with comparative genomics can be a powerful tool to trace evolutionary adaptations at single base pair resolution.

Abbreviations:
BS: binding site, TF: transcription factor, TFBS: transcription factor binding site, ChIP: chromatin immunoprecipitation, ZF: zinc finger, TSS: transcriptional start site, SNM: single nucleotide mutation

Introduction

Regulation of gene expression is considered as one of the major mechanisms shaping the phenotypic appearance of organisms [1, reviewed in 2, 3]. In particular, transcriptional initiation and elongation are of central importance to overall gene expression levels [4-8]. Both processes are thought to be regulated by transcription factors (TFs) that bind to specific DNA motifs of 5-15 base pairs termed transcription factor binding sites (TFBS). Even though there is some sequence variation in sites recognized by a certain TF, residing nucleotide substitutions can have great impact on TF affinities and transcription levels [9]. Different studies aimed at identifying such *cis*-regulatory changes in human and primates [10, 11], however experimentally supported examples are sparse. This is partly due to laborious experimental approaches and, of course, the fact that many *cis*-regulatory changes will only be relevant during development to regulate precise spacio-temporal gene expression patterns.

Previous studies, if not driven by the interest in a particular gene [12-14], were either entirely bioinformatic, as for example the search for certain substitution patterns in multiple species alignments [11, 15-17], or were based on differences in gene expression patterns of related species [18-21]. In both cases, further pinpointing of functional substitutions is difficult, as *de novo* TFBS prediction is not trivial, producing many false positives. The main problem here is evoked by inaccurate binding models [22-24] and the strong context-dependency of many TFs [25]. Now, the recently introduced method of chromatin immunoprecipitation followed by hybridization on microarrays (ChIP-chip) or massively parallel sequencing (ChIP-seq) permits genome-wide identification of *ex vivo* and *in vivo* TFBSs at high resolution, including the possibility to derive high-quality models of TF binding preferences. This data describing active TFBSs can subsequently be used in conjunction with multiple species alignments to search for sites that are specific to the species under investigation.

Here, we investigated the TFBSs of the GA binding protein transcription factor alpha subunit (GABPa) that possesses a binding motif which has been confirmed by several studies [9, 26, 27]. GABPa belongs to the ets family of DNA-binding factors and regulates a broad range of genes involved in cell cycle control, apoptosis, differentiation, hormonal regulation and other critical cellular functions [28]. Therefore, the likelihood to find human-specific BSs is higher than for TFs regulating only a small number of genes. Also, the DNA-binding domain of GABPa is entirely conserved in primates, mouse, dog and cow, rendering BS adaptations due to changes in protein structure unlikely. GABPa is known as a potent transcriptional activator and also regulates more than half of all bi-directional promoters [29]. In addition, repetitions of the GABPa BS influence transcription levels in a synergistic manner [30]. However, most

important to our approach is the finding that GABPa preferentially binds in close proximity to the transcriptional start site (TSS) [27], allowing to evaluate potential BS alterations straightforwardly by promoter-reporter gene assays.

In this study, we set out to demonstrate the practicability of combining data on experimentally supported transcription factor binding sites with comparative genomics to find functionally relevant substitutions. As shown in Figure 5, we used human HEK293 cells to perform ChIP-seq of the transcription factor GABPa, and searched the obtained set of functional binding sites for those that have evolved recently in the lineage leading to human. In this, we used the ChIP-seq peak regions to search for residing GABPa BSs and reconstructed the corresponding ancestral DNA sequences along the UCSC 44-vertebrate alignments. Subsequently, we identified human- and hominid-specific BSs by evaluation of the phylogenetic depth to which the human BS can be traced back. We tested the functionality of human-specific BSs by comparing the strength of wild type and mutated promoters from human, chimpanzee and rhesus. Four wild type gene promoters were selected and cloned, and in parallel, the newly evolved BSs were reversed to their ancestral states by site-directed mutagenesis. We also cloned the orthologous promoters for chimpanzee and rhesus and introduced the human specific GABPa BS. All wild type and mutated constructs were subjected to promoter reporter gene assays in HEK293 cells and in african green monkey-derived COS-1 cells to test the impact of the identified human- and hominid-specific substitutions on gene transcription.

Results

Identification of GABPa binding sites in 5,797 genomic regions

The main application of ChIP-seq is the identification of genomic regions that are enriched in specifically precipitated DNA (Figure 1A). To find regions of high sequencing read density (or peaks) within 6.96 million reads from GABPa ChIP-seq, we used the peak calling software QuEST (Figure B) [27]. We found 6,208 genomic peaks of GABPa reads, of which 80% can be mapped to transcripts within 600bps equally surrounding the transcriptional start sites (TSSs). Extension to 10kb centered on the TSSs results in 85% of peaks mappable to 18,832 UCSC transcripts, corresponding to 5,310 Entrez genes. As shown in Figure A, the majority of peaks was found to be located close to the nearest transcript start site. To identify the fraction of genes that is regulated by GABPa in HEK293 cells, we used previously published transcriptome sequencing (RNA-seq) data for the same cell line [31]. We found 49,245 UCSC transcripts with RNA-seq reads in two or more of the exons (in cases of transcripts consisting of one to three exons, only one exon needed to be matched by ChIP-seq reads). This number of transcripts corresponds to 15,101 Entrez genes, indicating that ~35% of the expressed gene-promoters are bound by GABPa.

In order to derive a GABPa consensus binding site from the ChIP-seq peak regions, we used DNA sequences of 200bps equally surrounding the 6,208 peak centers as input for the *de novo* motif discovery algorithm MEME [32]. A consensus binding sequence and a position specific weight matrix (PWM) were built based on 6,031 peaks (97% of peaks containing GABPa BSs) (Figure 1C). The PWM-contributing sites are preferentially located close to the peak centers (Figure 2B), indicating proper peak-calling from ChIP-seq reads. The identified PWM is very similar to the GABPa PWMs found in the TFBS databases JASPAR and TRANSFAC and it is almost identical to that found by Valoujev et al., who previously performed a similar experiment in Jurkat cells (Figure 3) [27]. Under default parameters, MEME assumes that each peak contains zero or one sequence motif. This assumption is advantageous to find non-repetitive motif elements. However, as more than one motif is likely present in each peak region, it is necessary to search for additional BSs, which can be done with the motif alignment and scan tool MAST (Figure 1D). The MAST analysis revealed 11,619 PWM hits in 5,797 peak regions of 200bps, with the majority of peaks containing two BSs, closely followed by peaks with single sites (Figure 1C).

Extraction of 224 human-specific GABPa binding sites

Based on the predicted 11,619 GABPa BSs (11 bps in length) within the ChIP-seq peaks, we extracted 11,008 multiple species alignments from UCSC MultiZ vertebrate alignments of 44 species (Figure 1E). For the remaining 611 BSs regions, there was either no alignment available, or the aligned regions were not contiguous. We were interested in BSs that emerged during human and hominid speciation. Using the eight available non-human primate genomes (Chimpanzee, Gorilla, Orangutan, Macaque, Marmoset, Tarsier and two prosimian species), we aimed at finding sites that are specific to four lineages, namely to human on the one hand, but also to the Hominini (Human and Chimpanzee), Homininae (Homini and gorilla) and Hominidae (Homininae and Orang-utan) lineages. For this, we reconstructed the ancestral sequences along the phylogeny of 34 mammalian species of the UCSC 44-vertebrate alignments using ANCESTORS (Figure 1F) [33]. The approach implemented in ACESTORS is suitable for reconstructing ancestral sequences including the most likely scenario of insertions and deletions observed in alignments, while retaining an extremely high degree of accuracy [33]. For the hominid lineage, no ancestral sequence was reconstructed for 65 BSs due to missing aligned sequences of more distantly related species, while all other alignments were obtained as expected. To search the reconstructed ancestral sequences for the presence of GABPa consensus sequences, we applied MAST using the human-derived GABPa PWM (Figure 1G).

We found 224 human specific BSs corresponding to 219 ChIP-seq peaks and 227 genes. For Hominini, we found 57 BSs, for Homininae 244 BSs and for Hominids 310 BSs. 41 peaks with human specific BSs were not mapped to known genes. Manual inspection of those peaks revealed that 23 are located in close proximity to ESTs that are not yet annotated by UCSC and therefore likely harbor true BSs. BS appearances for all ancestral branches leading to human are shown in Supplementary Figure S1.

Enriched gene categories associated with human-specific GABPa binding sites

We used the Database for Annotation, Visualization and Integrated Discovery (DAVID) to assess the 227 gene promoters that gained GABPa BS on the human lineage for enrichment of gene ontology associations, tissues with gene expression and protein domains [34, 35]. We found enrichment in genes involved in RNA processing (GO:0006396; $p=8.18^{-03}$), genes expressed in mammary gland ($p=2.53^{-03}$) and pineal gland ($p=7.08^{-03}$), and enrichment in genes containing a KRAB zinc finger (ZF) protein domain ($p=3.33^{-02}$). The full list of enriched categories can be found in Supplementary Table 1. KRAB zinc finger proteins are a class of genes specific to tetrapodes [36] and appear to have expanded on the primate lineage [37].

Interestingly, 35% of all HEK293-expressed genes are bound by GABPa, but 65% of the 277 expressed KRAB zinc finger genes are bound by GABPa.

Based on the analyses of ChIP-seq peak intensities, BS locations and gene expression strengths, we selected five promoters for further experimental studies. Also, we manually inspected transcriptional start sites and gene expression by exploring RNA-seq and RNA polymerase II ChIP-seq data from a previous study involving the same cell line [31]. The selected candidate promoters should comprise cases of repeated BSs, that have been reported to synergistically increase transcription levels [30], and also a bi-directional promoter, since GABPa is known to direct bidirectional transcription [29].

Functional analysis of newly evolved GABPa binding sites using reporter gene assays

Among the promoters with recently evolved GABPa BSs, we chose the promoters of ZNF197, ANTXR1 and TMBIM6 and the bi-directional promoter of ZNF398/ZNF425 for further analyses. ZNF197 was chosen as representative of the KRAB-ZF family and for the presence of two BSs, of which one is conserved among mammals, while the other is specific to humans. The anthrax toxin receptor-1 gene (ANTXR1) harbors three GABPa BS in the human promoter, but only two in chimpanzee and rhesus. In addition, this gene is highly expressed in HEK293 cells, and RNAi experiments showed strong down-regulation upon GABPa knockdown (data not shown). Even though we were particularly interested in human-specific BS gain, the TMBIM6 promoter harboring a hominid-specific BS was included due to a strong ChIP-seq peak and strong expression of the corresponding gene. Interestingly, MAST analysis predicted another GABPa BS next to the hominid-specific BS, which is deeply conserved but does not match the GABPa core consensus sequence “GGAA” (see Figure 4), a variation that was found in only 0.94% of all 11,008 BSs. Lastly, the bi-directional promoter of the KRAB ZF genes ZNF398/ZNF425, with TSSs located ~130bps apart, was chosen for being the only case in our analysis with two overlapping GABPa BSs, caused by two single nucleotide mutations specific to humans. This promoter was cloned in both directions to account for bi-directional transcription.

For each promoter, two fragments were cloned, one representing the wild type, the other a mutated form. Orthologous promoters were cloned from Human, Chimpanzee and Macaque genomic DNA. For human mutated forms, the BSs were modified by one or two single nucleotide mutations (SNMs) to mimic the ancestral state incompatible with GABPa binding. Inversely, for chimpanzee and macaque, the original sequences were modified to generate the human-specific GABPa BSs. All wild type (wt) and

mutated promoters were cloned into a modified *firefly* luciferase reporter gene vector pGL3 (see methods for details) and verified by whole-insert sequencing. Reporter gene expression was measured in human HEK293 cells and COS-1 cells derived from african green monkey and normalized to a co-transformed plasmid stably expressing *Renilla* luciferase. Figure 5 shows average *firefly* to *Renilla* ratios for all cloned fragments. Results are further summarized in Figure 6, including differences in activities of mutated and wt promoters and sequences of wt and mutated BSs. A genomic view of ChIP-seq peaks, cloned fragments, sequence differences to the human reference sequence and BS predictions can be found in Supplementary Figure S2.

The human ZNF197 promoter, harboring one conserved and one specific GABPa BSs, did not change activity upon BS reversion to the ancestral state (Figure 5A). Yet, human wt activity was significantly higher than chimpanzee and rhesus activities. Here, introduction of a SNM, creating the human specific BS, resulted in significant increase in activity in both cell lines, lifting reporter activities almost to the level of the human wt sequence. The human ANTXR1 promoter harbors two conserved BSs, plus one that is human-specific. Wt expression of chimpanzee and rhesus, carrying only two BSs, was significantly lower in at least one of the two cell lines (Figure 5B). SNM of the human specific BS, creating the ancestral state, caused significantly decreased reporter activity in COS-1 cells, while the observed decrease was not significant in HEK293 cells. Introduction of the human BS into chimpanzee and macaque promoters raised activity levels significantly in three of the four cases, namely for chimpanzee in both cell lines, while only in HEK293 cells for the macaque promoter.

The human promoter of TMBIM6 harbors a GABPa BS that is specific to hominids and another BS in close proximity that does not contain the GGAA core motif, even though it is highly conserved among mammals (see Figure 4). Human wt promoter activity was found to be significantly higher than chimpanzee and macaque activities in both cell lines, while rhesus activities were lower than chimpanzee (Figure 5C). Disruption of the hominid-specific BS in human lowered activity slightly below chimpanzee wt activity, while disruption of the chimpanzee BS lowered activity below macaque wt activity. On the other hand, introduction of the hominid-specific site into the macaque promoter resulted in very significant activity increase, lifting intensities above chimpanzee wt activity.

The bi-directional promoter of ZNF398/ZNF425 contains two overlapping human-specific GABPa BSs. For wt and in direction of ZNF398, promoters of human, chimpanzee and rhesus showed similar activities in HEK293 cells, while in COS-1 cells, activities were significantly

different (Figure 5D). The reversion of the human BS locus to the chimpanzee sequence by introduction of two SNMs resulted in more than two-fold reduction in activity in both cell lines. Vice versa, introduction of the human sequence into chimpanzee and rhesus promoters resulted in a very significant increase in activity in both cell lines. We observed similar effects of this fragment in ZNF425 direction, but to a lower, yet still very significant degree (Figure 5E).

In summary and regarding both cell lines, introduction of human GABPa BSs into chimpanzee or rhesus promoters resulted in significant increase in reporter gene expression in 17 of 18 cases. On the other hand, disruption of GABPa BSs in human and chimpanzee promoters led to significant decrease of reporter gene activity in 9 out of 12 cases. In no case, we observed opposite effects, since BS introduction never led to significant activity decrease, and BS disruption did not result in any significant activity increases.

Discussion

To identify functional TFBSs that were gained during hominid and human evolution, we have performed ChIP-seq of the transcription factor GABPa from human HEK293 cells. The search for over-represented sequence motifs within the TF-bound regions resulted in a GABPa consensus binding motif almost identical to that identified by a similar approach [27]. Despite the differences in experimental protocols, ChIP antibodies and cell lines, the near-perfect agreement of the derived binding preferences shows the high accuracy of the ChIP-seq approach.

GABPa regulates a significant fraction of human genes

More than one third of the promoters of genes expressed in HEK293 cells are bound by GABPa. The finding that more than 90% of these promoters also harbour one or more GABPa binding sites underlines the importance of this sequence motif in proximal promoter regions and the impact of GABPa on gene regulation. Considering that a TATA box is present in less than 22% of all human promoters [38], our results indicate that functional GABPa BSs reside in a comparable if not greater fraction of all human gene promoters.

Screening of the central 200bps of each ChIP-seq peak region revealed the presence of two and more BSs in almost 60% of the peaks. It is likely that the majority of the predicted sites contribute to transcriptional regulation, as GABPa is known to form heterotetramers composed of two GABPa and two GABPb subunits, to bind tandem repeats of the GGAA consensus motif [26]. In addition, it has been speculated that accumulations of BSs, including highly degenerate inexact versions, provide a favorable landscape attracting transcription factors to high-affinity sites [9, 39]. Therefore, newly emerged sites can also be functional despite the presence of deeply conserved BSs within promoters.

Human-specific GABPa binding sites are enriched for genes potentially important for human evolution

To find BSs that are specific to human or hominids, we reconstructed the ancestral sequences for the 11,008 human GABPa binding sites in HEK293 cells based on the UCSC 44-vertebrate whole genome alignments. This approach relies on the accuracy of the UCSC alignments. UCSC multiZ alignments of human-chimpanzee and human-macaque have been estimated to be problematic (while not necessarily wrong) for 0.004% and 0.02% of the aligned nucleotides, respectively [40]. Theoretically, this would imply, for 11,008 human-macaque alignments corresponding to 11bp of each GABPa binding site, that a fraction of 24 nucleotides was problematically aligned. However, this fraction is likely even smaller, as most problematic alignments have been found in intronic and intergenic regions [40], while the majority of the GABPa BSs reside in proximal promoter regions, where mammalian genomic sequences are particularly conserved [41], allowing for very accurate overall alignments.

Among the genes with BSs specific to the human lineage, we found enrichment in genes involved in RNA processing, genes expressed in mammary and pineal gland, and enrichment in genes containing a KRAB zinc finger protein domain. Even though the enrichments were not significant after correction for multiple testing, corresponding genes have likely been subjected to selective pressure during hominid evolution. For example, evolutionary changes in milk composition can be caused by regulatory mutations accounting for different needs of newborns for nutritional and immunological components [42]. Similarly, genes expressed in the pineal gland involved in circadian rhythm, growth, puberty and aging [43] have likely undergone adaptive evolution. Also, KRAB zinc fingers, a relatively young class of transcription factors proliferating through gene duplications and segmental duplications [44], are prone to acquire new sets of regulatory sequences.

Reporter gene assays with wild type and mutated promoters confirm the functionality of newly evolved GABPa binding sites

To test whether the identified sites play a role in transcription regulation, we carried out dual luciferase reporter assays of human, chimpanzee and macaque promoters in human HEK293 and african green monkey-derived COS-1 cells. The relative reporter activities observed in the monkey cell line were almost identical to those observed in the human cell line. This finding is supported by a recent study using an aneuploid mouse strain carrying an extra copy of human chromosome 21, which revealed that virtually all human transcription factor-binding locations

found in human hepatocytes were recapitulated across the entire human chromosome 21 within the aneuploid mouse hepatocytes [45]. Therefore, the results derived here from transfections of COS-1 cells can be regarded as controls for the assays in HEK293 cells, and vice versa.

In general, promoter-reporter gene assays are of great value to the functional characterization of regulatory elements. Within a cellular environment, these assays can be more or less representative for the regulation of endogenous expression, depending on the type of gene-promoter under investigation. For tightly regulated genes important during organismal development, cell lines can be of limited use to study promoter responses, as developmental signals may not be present. On the other hand, gene promoters involved in mechanisms of general importance to cellular function and survival can be studied for species-specific endogenous expression using cell lines, since intracellular signals ensuring cellular homeostasis govern transcriptional output of these genes to a greater extent than for developmental or environment-responsive genes. Hence, we do not emphasize to draw conclusions on inter-species differences in wt promoter strengths for the transcription factors ZNF197, ZNF398 and ZNF425, as these genes likely represent developmentally regulated genes with complex activation patterns. The same is true for ANTXR1, which represents a transmembrane adhesion molecule linking the actin cytoskeleton to collagen I fibers [46]. ANTXR1 is widely expressed, above all in endothelial cells, and is involved in angiogenesis [47]. Importantly, ANTXR1 has been shown to be a docking protein for *Bacillus anthracis* toxin, the causative agent of the anthrax disease.

The case of TMBIM6 might be different, as this is an anti-apoptotic protein protecting the cell against apoptosis induced by endoplasmatic reticulum stress (ER-stress) through reduction of the accumulation of reactive oxygen species (ROS) at the ER membrane [48]. Moreover, according to UniGene EST profiles, TMBIM6 is strongly expressed in all tissues [49]. TMBIM6 is more likely regulated by intracellular signals involved in homeostasis and hence, differences in reporter activities of orthologous wt promoters are presumable informative.

Regarding the mutation analyses for ZNF197, we found that introduction of a human-specific GABPa BS into chimpanzee and macaque promoters resulted in a significant and consistent increase in reporter activity, while we did not observe an activity decrease when disrupting the newly evolved BS in the human promoter. These findings could indicate the presence of additional mutations, allowing the binding of one or more factors that compensate the activating property of the new GABPa BSs. If a compensating factor depends on GABPa to fulfill its function, deletion of the new GABPa BS would have no effect.

Even if the transcriptional output is maintained, regulation of human ZNF197 expression might have changed. In theory, cases like this one could reflect a scenario where an increase in gene expression was beneficial at some time during evolution, while at a later period the evolutionary pressure was released again. Since that time, additional *cis*-regulatory mutations may have been fixed that compensate for the effect of the formerly beneficial mutation.

Similar to ZNF197, the disruption of the human-specific GABPa BS within the ANTXR1 promoter had no effect, while its introduction into the chimpanzee promoter showed significant activity increase. Again, this finding is indicative for a functional human-specific BS whose impact on transcription is compensated by further *cis*-regulatory mutations in human. Indeed, both promoters (ZNF197 and ANTXR1) harbor additional human-specific mutations in less then 100bp distance to the newly evolved GABPa BSs. In general, compensation does not necessarily render human-specific BSs irrelevant, as under different conditions, these BSs might still have a functional impact. In cell lines, it has been shown that susceptibility to anthrax toxin is influenced by the level of ANTXR1 expression [50]. In addition, subcutaneous injection of *B. anthracis* spores in mice significantly reduced ANTXR1 mRNA expression in lung, heart, stomach, skin, brain and muscle [51]. Hence, alterations in ANTXR1 regulation might play an important role in dealing with *B. anthracis* infection.

The TMBIM6 promoter might be of particular interest in respect to hominid and human evolution. Significant differences were found in wt promoter strengths of human, chimpanzee and macaque, which can be partly explained by a hominid-specific GABPa BS, as indicated by the mutational analyses. The human wt promoter drives higher reporter activity compared to the chimpanzee wt promoter, even though both species share a GABPa consensus BS. However, the human promoter (including exon 1) harbors two additional SNMs in very close proximity that might account for the observed difference (see Figure 4). The second GABPa BS predicted within the promoter does not mach the core GGAA motif, but this site is likely functional according to deep conservation and the fact that the core consensus is present in four species. TMBIM6 is an interesting candidate due to its function as reducer of ER-stress-induced accumulations of reactive oxygen species [48]. ER stress has been implicated in the development of diabetes, atherosclerosis and in many of the aging-related neurodegenerative diseases, such as Alzheimer's, amyotrophic lateral sclerosis and Parkinson's [52]. Therefore, changes in the regulation of TMBIM6 expression might play a role in extending the life spans of hominids.

The human genes ZNF398 and ZNF425, located head to head only ~130bps apart, were found regulated by GABPa, as disruption of two overlapping GABPa BSs residing in this bi-

directional promoter resulted in >2-fold activity reduction in the direction of ZNF398 and >1.2-fold reduction in the direction of ZNF425. This finding reflects another property of GABPa, namely regulation of bi-directional transcription [29]. The orthologous promoters of chimpanzee and rhesus, which are devoid of GABPa BS, show even stronger activities than the human wt promoter upon BS introduction, lifting activity levels well above human wt levels. To adjust the sequence of the macaque to the human BS sequence, six mutations were necessary, accompanied by a three bp deletion. Interestingly, this strong intervention had only a moderate effect in direction of ZNF425, while in the direction of ZNF398, we observed an almost 5-fold increase in activity. Taken together, this bi-directional promoter gained regulation through GABPa in human with stronger impact in the direction of ZNF398, while the orthologous chimpanzee and rhesus promoters were found to be regulated differently.

In summary, our experiments demonstrate that newly evolved functional TFBSs can be identified using ChIP-seq data together with comparative genomic analysis, which is reflected by the expected results of elevated reporter-gene expression in case of BS introduction and decreased expression in case of BS deterioration. Notably, a study during which GABPa BSs were introduced into six promoters previously unregulated by GABPa found only one of the six promoters activated [29], indicating that the introduction of GABPa BSs *per se* is mostly insufficient to affect gene expression. However, we find that the introduction of human-specific GABPa BSs into chimpanzee and rhesus promoters consistently elevated reporter gene activity, indicating that BSs need to be placed in the right context to exert an influence on gene expression. To our knowledge, this is the first approach that is capable to reliably identify newly evolved and functional TFBSs at high accuracy on a genomic scale.

Methods

Chromatin immunoprecipitation-sequencing

We performed ChIP-seq according to a published protocol [53]. Briefly, 5×10^8 HEK293 cells were cross linked for 10 min at room temperature with 1% formaldehyde, nuclei were prepared following the published protocol and chromatin was sheared to 100-500 bp size by 45 cycles of 30 sec on/off at the highest amplitude using a Bioruptor water bath sonicator (Diagenode). Nuclear extracts were immunoprecipitated with 10 µg rabbit anti-GABP-α (H-180X, Santa Cruz Biotechnology sc-22810) and 70 µl Protein G-Dynabeads (Invitrogen). After washing of beads, protein-DNA complexes were eluted, crosslinks were reversed overnight, and DNA was purified as published. For sequencing library preparation, 2 ng ChIP DNA and 10 ng Input DNA were subjected to end-repair, addition of Adenin bases and ligation of sequencing adapters, followed by DNA amplification through PCR and subsequent gel purification for sequencing on an Illumina Genome Analyzer GA2 according to the manufacturer's protocol for 36 bp reads. Reads were aligned to the human genomic sequence (hg18) using Eland, resulting in 6,955,499 GABPa ChIP reads with unique match to the genome (allowing up to two mismatches) and 2,948,346 corresponding reads from the input DNA.

Peak calling, gene mapping, MEME and MAST analysis

ChIP-seq reads was analyzed in three steps as published previously [27]. Briefly, we used the peak-calling algorithm QuEST to find enriched regions within the mapped ChIP-seq reads (see Supplementary Methods). Peaks were mapped to all UCSC known transcripts that start in a distance of 5kb to each side of the peak. For mapping, we used 65,297 UCSC known transcripts [54] mapping to 20,101 Entrez genes. UCSC transcript IDs were converted to Entrez gene IDs using UCSCs knownToLocusLink table. 260 peaks were mapped to UCSC transcripts that do not correspond to Entrez genes. For 934 peaks that could not be mapped to UCSC genes, we searched the human EST database and found 545 peaks that map within 5kb upstream to an EST starts. After extraction of peak-associated sequences comprising 200bp surrounding each peak center via the UCSC table browser [55], we applied MEME [32] to identify over-represented motifs. Using default parameters, MEME assumes that each sequence contains zero or one motif. The derived position weight matrix (PWM) was then used to run the MEME tool

MAST that reports the occurrences of PWM hits for each sequence in the input set at a particular stringency (set to p=0.001).

Multiple species alignment extraction and conversion

UCSC provides the 44way-vertebrate alignments in a multiple alignment format (MAF) that consists of short blocks (1-200bp) of multiple alignments, which can be concatenated. We extracted the alignments corresponding to GABPa BSs within the ChIP peak regions via UCSCs table browser [55] and converted the MAF-formatted alignments into the commonly used FASTA format, while excluding non-syntenic blocks and species with missing sequence data (e.g. insertions not included in the MAF alignments).

Ancestral sequence reconstruction

Ancestral sequences were calculated using ANCESTORS [56] obtained from http://ancestor.bioinfo.uqam.ca/programs/anc.tar. The program requires a multiple species alignment and a phylogenetic tree including branch lengths. To calculate branch lengths, all alignments were concatenated and run through BASEML, a maximum likelihood-based program of the PAML package [57]. The nucleotide substitution model HKY was used in both programs. Phylogeny was taken from UCSC (phyloP44wayPlacMammal) available at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/vertebrate.mod.

Cloning and plasmid preparation

PCR primers were designed using the Primer3 online service and extended by 29bp Gateway attB tails (Invitrogen) at the 5' end of each primer. Touch-down PCR was performed as described previously [58], except for the supplementation of each reaction with 0.001U *Pfu* polymerase. Mutations were introduced by primer-mediated mutagenesis. To facilitate cloning, the Gateway cloning cassette (Invitrogen) was amplified with forward primer attP1 and reverse attP2 and cloned into the pGL3 reporter vector (Promega). PCR products were purified and cloned upstream of the luciferase gene in the modified pGL3 vector using BP Clonase II Enzyme Mix (Invitrogen) following the manufacturers instructions. Plasmids were transformed into the *E. coli* strain GM2929. Inserts of positive clones were sequenced by the Services in Molecular Biology Company (Berlin, Germany). DNA concentration was measured on a Nanodrop UV spectrophotometer (NanoDrop Technologies) and standardized to 50 ng/µL for transfections.

Cell culture, transient transfection, and reporter gene activity assays

HEK293 and COS-1 cells were cultivated in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 100 U/ml penicillin/G-streptomycin (Biochrom) and 10% heat-inactivated fetal bovine serum (Biochrom) at 37°C and 5% CO_2. We seeded ~15,000 (HEK293) and ~5,000 (COS-1) cells per well in clear-bottom 96-well plates (Costar). Twenty-four hours after seeding, we co-transfected 150ng of experimental firefly luciferase plasmid together with 10ng of *Renilla* luciferase control plasmid (pRL-TK, Promega) in five replicates using Lipofectamine 2000 following the manufacturer's recommendations. Cells were lysed 24 hours post-transfection. We measured firefly luciferase and *Renilla* luciferase activities using the Centro LB960 luminometer (Berthold) and the Dual Luciferase Kit (Promega). We followed the protocol suggested by the manufacturer with the exception of injecting 25µl each of the firefly luciferase and *Renilla* luciferase substrate reagents. All measurements were performed at least in three technical and two biological replicates, including new dilution and concentration adjustments of reporter plasmids.

Figures

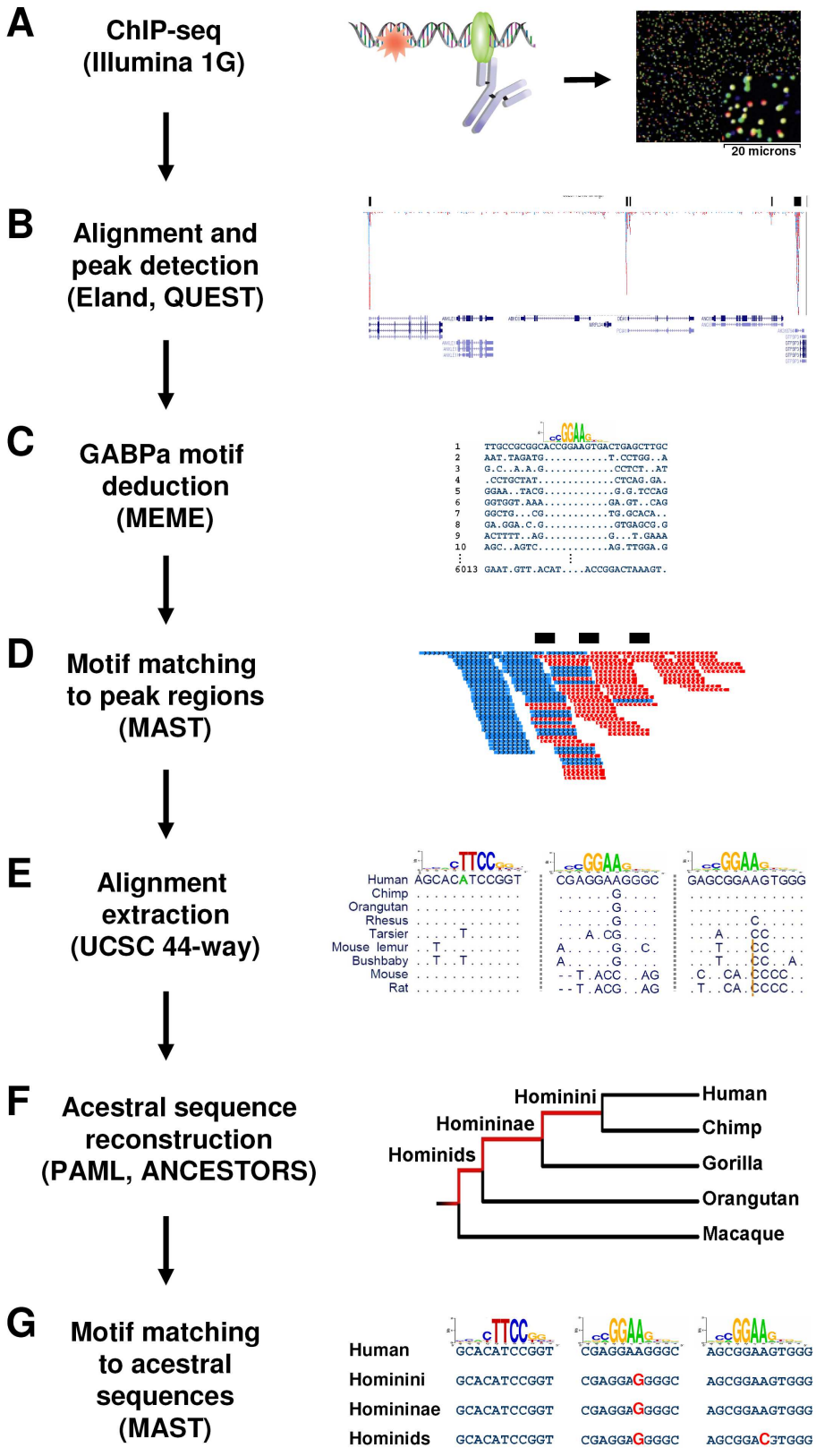


Figure 1. Overview on the procedure

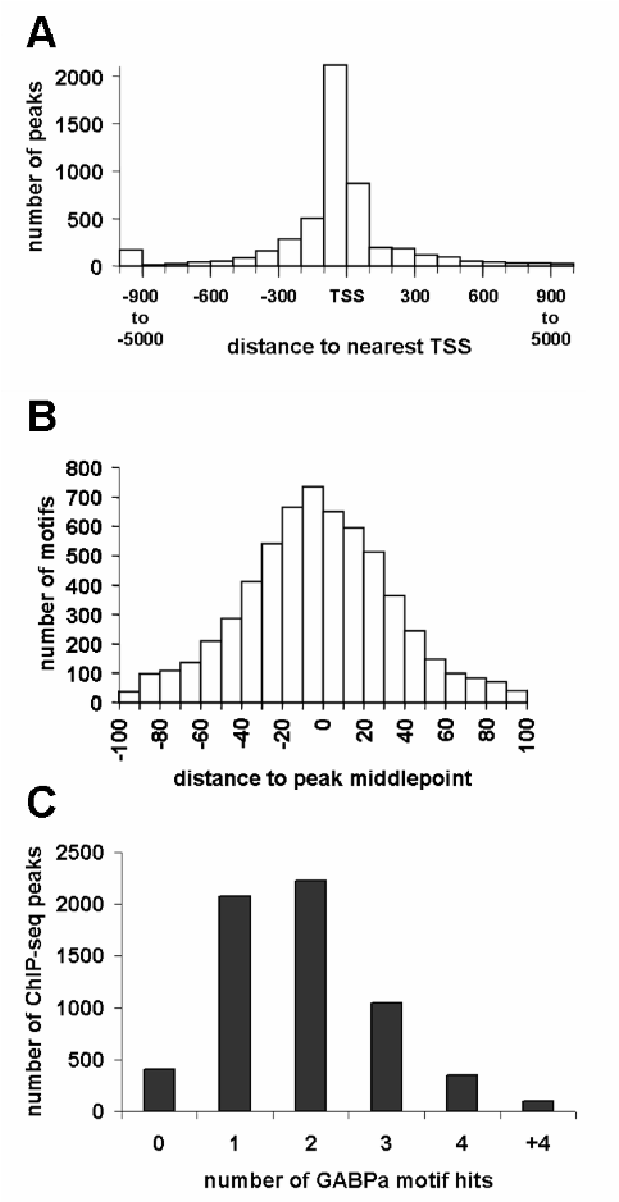


Figure 2. GABPa peaks map close to gene starts and harbor GABPa BS residing closely to the peak centers. (A) The histogram shows the distance of peak calls to the nearest transcriptional start sites (TSSs) of UCSC genes within 10kb centered on the TSS. The horizontal axis shows the base pairs surrounding the TSS. Negative values represent upstream, positive downstream regions. (B) The histogram shows the distance of the sites contributing to the MEME motif (6,031 of 6,208 in total) to the ChIP peak centers. (C) The histogram shows the distribution of motif occurrences within 200 bp surrounding the ChIP peak centers.

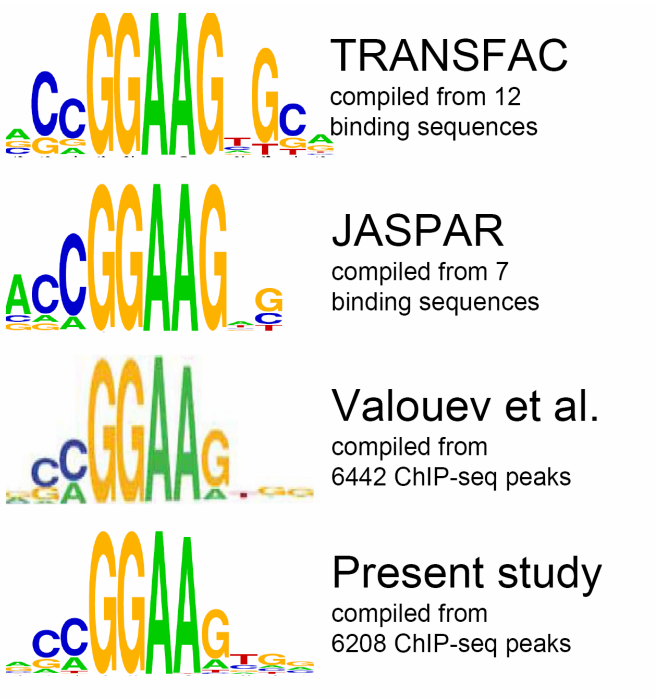


Figure 3. Comparison of GABPa motifs from different studies and databases. Sequence logos represent the different position weight matrices.

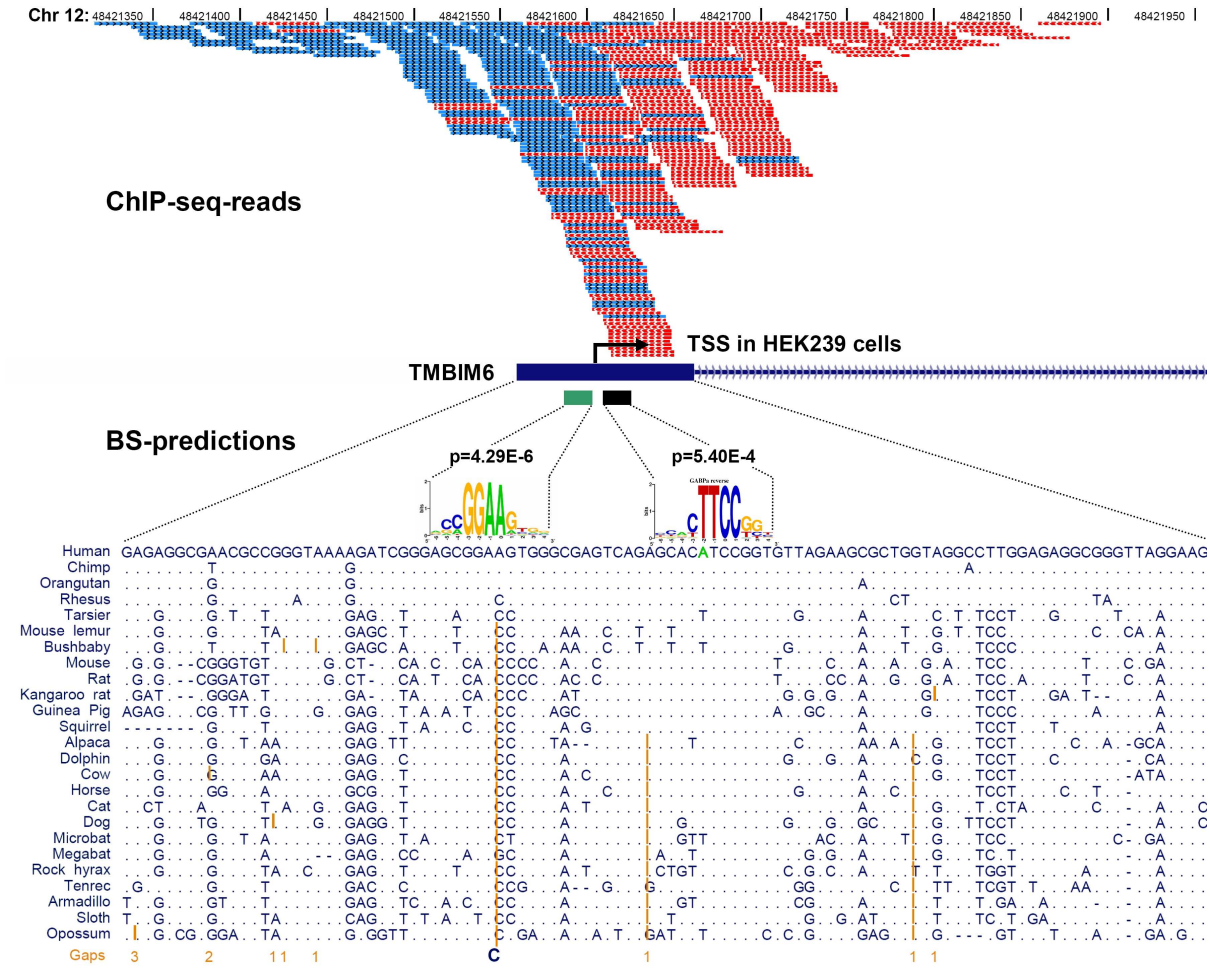


Figure 4. Genomic view of GABPa ChIP-seq reads spanning the TMBIM6 promoter including GABPa binding site predictions and multiple species alignment of the first exon. ChIP-seq reads are colored in blue (forward reads) and red (reverse reads). The first exon (5'UTR) is shown as blue bar with a black arrow indicating the transcriptional start site (TSS) in HEK293 cells as determined by RNA-seq. GABPa binding site predictions are shown as green and black boxes. Within the blowup in the lower part, including the UCSC multiple species alignment of exon 1, BSs are shown as sequence logos of the GABPa PWM aligned to their matching positions. Within the alignment, dots indicate identity to the human reference sequence, while orange vertical bars indicate bases that are not depicted. Orange numbers below represent the sum of bases not depicted. The blue (C) illustrates the presence of a single cytosine in all non-haplorhini at the indicted site.

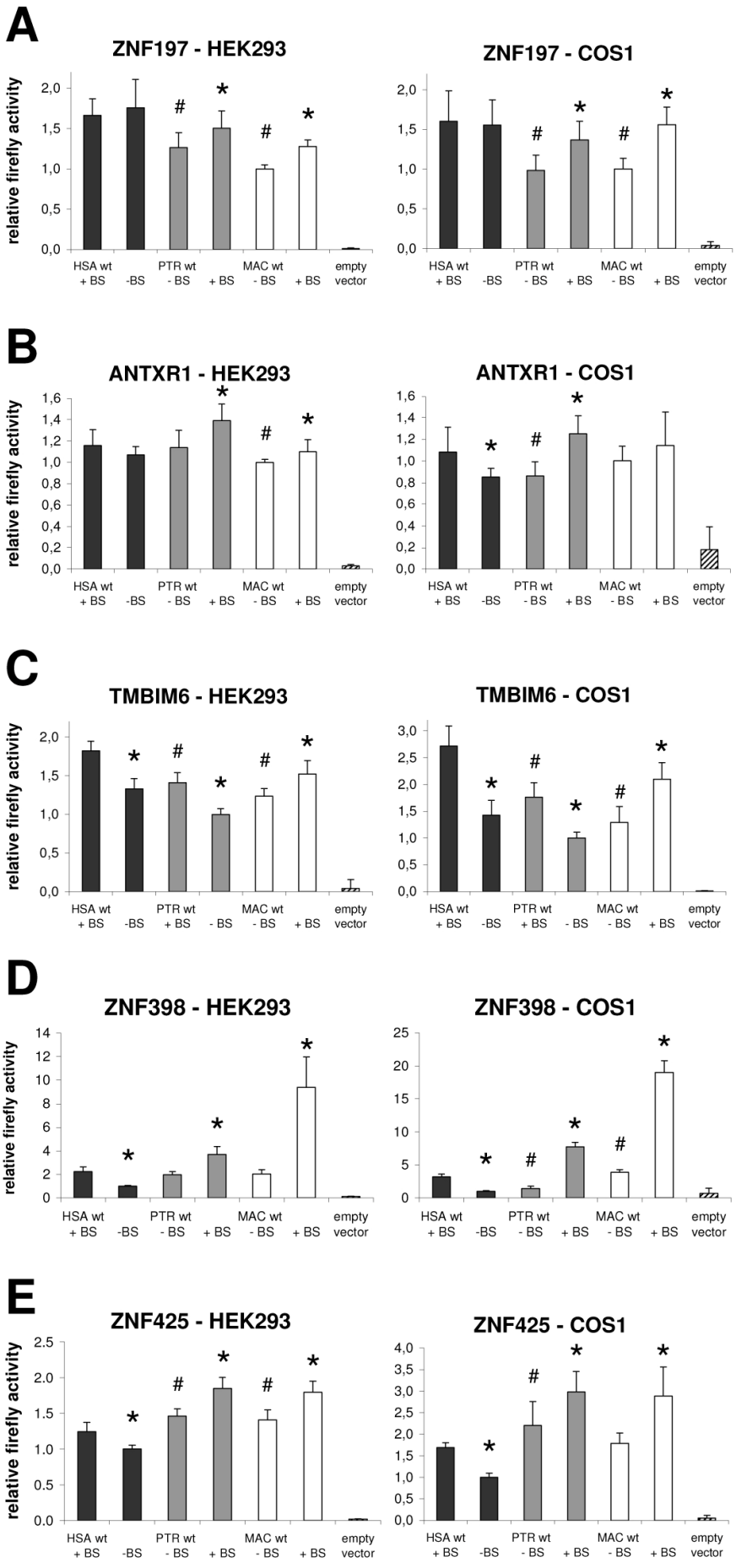

A
ZNF197 - HEK293
ZNF197 - COS1
B
ANTXR1 - HEK293
ANTXR1 - COS1
C
TMBIM6 - HEK293
TMBIM6 - COS1
D
ZNF398 - HEK293
ZNF398 - COS1
E
ZNF425 - HEK293
ZNF425 - COS1
relative firefly activity
HSA wt + BS
-BS
PTR wt - BS
+ BS
MAC wt - BS
+ BS
empty vector

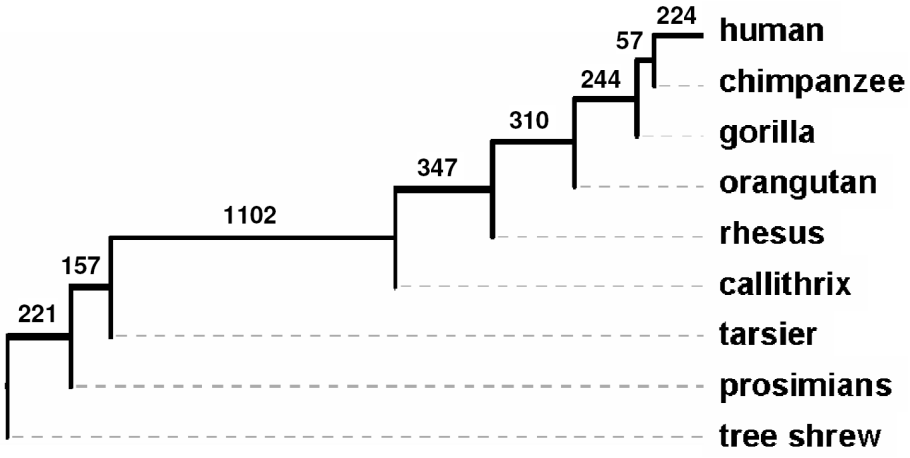
Figure 5. Normalized *firefly* luciferase activity of human, chimpanzee and rhesus wild type (wt) and mutated (mut) promoters. Bars represent average *firefly* to *Renilla* ratio in black for human (HSA), in grey for chimpanzee (PTR) and in white for macaque (MAC) promoters. For each species, the left column refers to the wild type and the right bar to the mutated promoter. (+BS) or (–BS) indicate presence or absence of GABPa binding sites in wt promoters and indicate introduction or disruption of sites in mutated promoters. For each promoter, measured activities were normalized to the construct with the lowest promoter activity level in HEK293 cells (set to one). Standard errors were calculated from at least six replicates. (*) indicates significant differences between wt and mutated promoter activities according to a one-tailed Welch's test, while (#) indicates significant difference of wt chimpanzee or macaque promoters compared to human wt activity, according to a two-tailed Welch's test. The raw data for all constructs are available in Supplemental Table 1-5.

Promoter (length)	# of BS per peak	Species	Wild type (wt)	Mutated (mut)	Log2 ratios (mut/wt) HEK293	COS-1
ZNF398	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -1.16 ***	↓ -1.69 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.92 ***	↑ 2.45 ***
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 2.22 ***	↑ 2.28 ***
ZNF425	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -0.32 ***	↓ -0.76 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.34 ***	↑ 0.44 **
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.35 ***	↑ 0.69 ***
ZNF197	2	HSA	TGCCGGAAGGGC	TGCCGCAAGGGC	→ 0.08 ns	→ -0.04 ns
(430bp)	1	PTR	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.25 *	↑ 0.47 ***
	1	MAC	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.35 ***	↑ 0.64 ***
ANTXR1	3	HSA	GCGAGGAAGGGC	GCGAGGAGGGGC	↓ -0.11 ns	↓ -0.34 *
(633bp)	2	PTR	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.29 **	↑ 0.54 ***
	2	MAC	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.14 *	↑ 0.19 ns
TMBIM6	1 (2)	HSA	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.45 ***	↓ -0.93 ***
(576bp)	1 (2)	PTR	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.49 ***	↓ -0.81 ***
	0 (1)	MAC	GAGCGGAcGTGG	GAGCGGAAGTGG	↑ 0.30 ***	↑ 0.70 ***

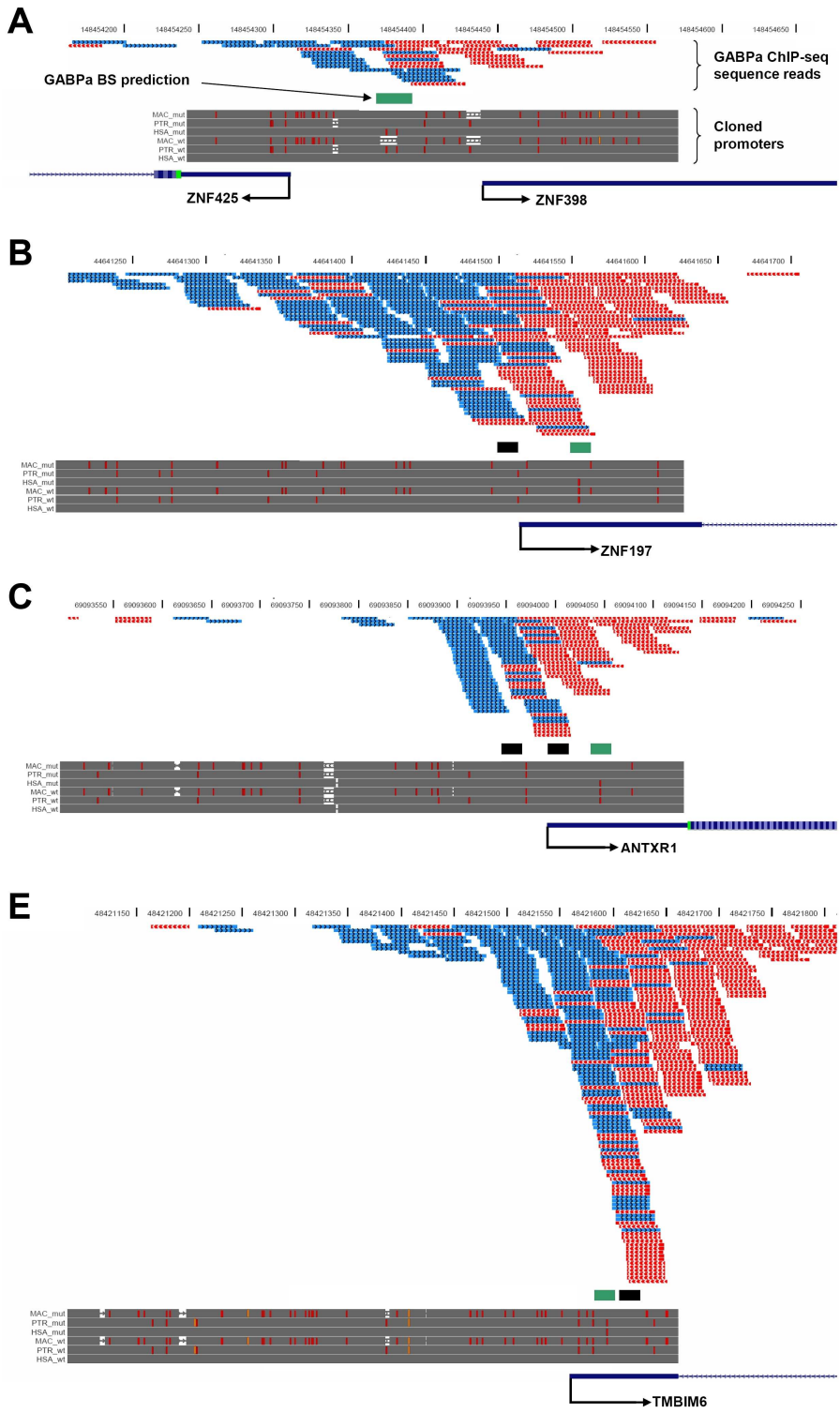
Figure 6. The introduction and disruption of GABPa consensus binding sites significantly influence reporter gene activities. For each gene, the number of predicted binding sites within 200bp surrounding the peak centers is indicated. Species are denoted by HSA – Homo sapiens, PTR – Pan troglodytes (chimp) and MAC – Macaca mulatta (macaque). Sequences are shown for wild type and mutated sites. Underlined bases indicate differences from the human wt sequence. Mutated bases are coloured in green or red indicating generation or disruption of a GABPa BS, respectively. Green arrows depict higher activity of mutated over wt promoter, red arrows indicate lower activity, and yellow arrows represent no change. Differences in mutated and wt promoter activities are given as log2 ratios of average luciferase to *Renilla* ratios. Significance levels, as determined by Welch's t-test for unequal variances, are indicated as (*) $P < 0.05$, (**) $P < 0.01$, (***) $P < 0.001$ and (ns) not significant.

Category	Source	Term	Count	PValue
Biological Process	GENE ONTOLOGY	RNA processing (GO:0006396)	22	8,18E-03
Biological Process	GENE ONTOLOGY	RNA metabolic process (GO:0016070)	59	2,52E-02
Biological Process	GENE ONTOLOGY	Regulation of Wnt receptor signaling pathway (GO:0030111)	3	3,31E-02
Cellular Component	GENE ONTOLOGY	Nuclear envelope (GO:0005635)	8	3,41E-02
Tissue Expression	CGAP EST	Mammary gland (16621)	9	2,53E-03
Tissue Expression	CGAP EST	Pineal gland (898)	14	7,08E-03
Tissue Expression	CGAP SAGE	Eye (1363)	20	4,39E-02
Tissue Expression	GNF U133A	PB-CD19+Bcells	123	2,35E-02
Tissue Expression	GNF U133A	Thymus	33	3,97E-02
Protein Domain	SMART	KRAB (SM00349)	14	2,45E-02
Protein Domain	INTERPRO	KRAB box (IPR001909)	14	3,03E-02
Protein Domain	PFAM	KRAB (PF01352)	14	3,33E-02

Supplementary Table S1. Enrichments in biological processes, cellular components, tissue expression and protein domains are shown for 229 human genes harbouring specific GABPa binding sites. The enrichment analysis was performed using DAVID functional analysis with 5,310 GABPa-regulated genes as background set.



Supplementary Figure S1. The phylogenetic tree shows GABPa BSs gained on the ancestral lineages leading to human.



Supplementary Figure S2. Genomic view of ChIP-seq reads for gene promoters analyzed in promoter reporter assays, including BS predictions, cloned promoters and gene starts. ChIP-seq reads are represented as blue and red dashes, representing forward and reverse reads, respectively. GABPa BS predictions within 200bp surrounding the peak centers are indicated as small boxes in black or in green where the BS is specific to human or hominids. Cloned promoters are represented as grey horizontal bars. Red dashes indicate mismatches to the human wild type sequence, while orange dashes indicate insertions.

References

1. Britten, R.J. and E.H. Davidson, *Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty.* Q Rev Biol, 1971. **46**(2): p. 111-38.
2. Wray, G.A., *The evolutionary significance of cis-regulatory mutations.* Nat Rev Genet, 2007. **8**(3): p. 206-16.
3. King, M.C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-16.
4. Lemon, B. and R. Tjian, *Orchestrated response: a symphony of transcription factors for gene control.* Genes Dev, 2000. **14**(20): p. 2551-69.
5. Roeder, R.G., *The role of general initiation factors in transcription by RNA polymerase II.* Trends Biochem Sci, 1996. **21**(9): p. 327-35.
6. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes.* Mol Biol Evol, 2003. **20**(9): p. 1377-419.
7. Muse, G.W., et al., *RNA polymerase is poised for activation across the genome.* Nat Genet, 2007. **39**(12): p. 1507-11.
8. Zeitlinger, J., et al., *RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo.* Nat Genet, 2007. **39**(12): p. 1512-6.
9. Lin, J.M., et al., *Transcription factor binding and modified histones in human bidirectional promoters.* Genome Res, 2007. **17**(6): p. 818-27.
10. Chabot, A., et al., *Using reporter gene assays to identify cis regulatory differences between humans and chimpanzees.* Genetics, 2007. **176**(4): p. 2069-76.
11. Haygood, R., et al., *Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution.* Nat Genet, 2007. **39**(9): p. 1140-4.
12. Rockman, M.V., et al., *Ancient and recent positive selection transformed opioid cis-regulation in humans.* PLoS Biol, 2005. **3**(12): p. e387.
13. Huby, T., et al., *Functional analysis of the chimpanzee and human apo(a) promoter sequences: identification of sequence variations responsible for elevated transcriptional activity in chimpanzee.* J Biol Chem, 2001. **276**(25): p. 22209-14.
14. Romanelli, M.G., et al., *Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5' flanking region.* Genomics, 2009. **93**(3): p. 227-34.
15. Bird, C.P., et al., *Fast-evolving noncoding sequences in the human genome.* Genome Biol, 2007. **8**(6): p. R118.

16. Pollard, K.S., et al., *Forces shaping the fastest evolving regions in the human genome.* PLoS Genet, 2006. **2**(10): p. e168.
17. Taylor, M.S., et al., *Rapidly evolving human promoter regions.* Nat Genet, 2008. **40**(11): p. 1262-3; author reply 1263-4.
18. Enard, W., et al., *Intra- and interspecific variation in primate gene expression patterns.* Science, 2002. **296**(5566): p. 340-3.
19. Gilad, Y., et al., *Expression profiling in primates reveals a rapid evolution of human transcription factors.* Nature, 2006. **440**(7081): p. 242-5.
20. Blekhman, R., et al., *Gene regulation in primates evolves under tissue-specific selection pressures.* PLoS Genet, 2008. **4**(11): p. e1000271.
21. Lin, L., et al., *Using high-density exon arrays to profile gene expression in closely related species.* Nucleic Acids Res, 2009. **37**(12): p. e90.
22. Barbulescu, K., et al., *New androgen response elements in the murine pem promoter mediate selective transactivation.* Mol Endocrinol, 2001. **15**(10): p. 1803-16.
23. Horie-Inoue, K., et al., *Identification of novel steroid target genes through the combination of bioinformatics and functional analysis of hormone response elements.* Biochem Biophys Res Commun, 2006. **339**(1): p. 99-106.
24. Verrijdt, G., A. Haelens, and F. Claessens, *Selective DNA recognition by the androgen receptor as a mechanism for hormone-specific regulation of gene expression.* Mol Genet Metab, 2003. **78**(3): p. 175-85.
25. Davidson, E.H., *The regulatory genome : gene regulatory networks in development and evolution*. 2006, Burlington, MA ; San Diego: Academic. xi, 289 p.
26. Batchelor, A.H., et al., *The structure of GABPalpha/beta: an ETS domain- ankyrin repeat heterodimer bound to DNA.* Science, 1998. **279**(5353): p. 1037-41.
27. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nat Methods, 2008. **5**(9): p. 829-34.
28. Rosmarin, A.G., et al., *GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions.* Blood Cells Mol Dis, 2004. **32**(1): p. 143-54.
29. Collins, P.J., et al., *The ets-related transcription factor GABP directs bidirectional transcription.* PLoS Genet, 2007. **3**(11): p. e208.
30. Yu, M., et al., *GA-binding protein-dependent transcription initiator elements. Effect of helical spacing between polyomavirus enhancer a factor 3(PEA3)/Ets-binding sites on initiator activity.* J Biol Chem, 1997. **272**(46): p. 29060-7.
31. Sultan, M., et al., *A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.* Science, 2008. **321**(5891): p. 956-60.

32. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009.
33. Diallo, A.B., V. Makarenkov, and M. Blanchette, *Exact and heuristic algorithms for the Indel Maximum Likelihood Problem.* J Comput Biol, 2007. **14**(4): p. 446-61.
34. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
35. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
36. Birtle, Z. and C.P. Ponting, *Meisetz and the birth of the KRAB motif.* Bioinformatics, 2006. **22**(23): p. 2841-5.
37. Huntley, S., et al., *A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors.* Genome Res, 2006. **16**(5): p. 669-77.
38. Gershenzon, N.I. and I.P. Ioshikhes, *Synergy of human Pol II core promoter elements revealed by statistical sequence analysis.* Bioinformatics, 2005. **21**(8): p. 1295-300.
39. Zhang, C., et al., *A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome.* Nucleic Acids Res, 2006. **34**(8): p. 2238-46.
40. Prakash, A. and M. Tompa, *Measuring the accuracy of genome-size multiple alignments.* Genome Biol, 2007. **8**(6): p. R124.
41. Mahony, S., et al., *Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome.* Genome Biol, 2007. **8**(5): p. R84.
42. Lemay, D.G., et al., *The bovine lactation genome: insights into the evolution of mammalian milk.* Genome Biol, 2009. **10**(4): p. R43.
43. Pierpaoli, W., *Neuroimmunomodulation of aging. A program in the pineal gland.* Ann N Y Acad Sci, 1998. **840**: p. 491-7.
44. Hamilton, A.T., et al., *Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes.* Genome Res, 2006. **16**(5): p. 584-94.
45. Wilson, M.D., et al., *Species-specific transcription in mice carrying human chromosome 21.* Science, 2008. **322**(5900): p. 434-8.
46. Werner, E., A.P. Kowalczyk, and V. Faundez, *Anthrax toxin receptor 1/tumor endothelium marker 8 mediates cell spreading by coupling extracellular ligands to the actin cytoskeleton.* J Biol Chem, 2006. **281**(32): p. 23227-36.
47. Hotchkiss, K.A., et al., *TEM8 expression stimulates endothelial cell adhesion and migration by regulating cell-matrix interactions on collagen.* Exp Cell Res, 2005. **305**(1): p. 133-44.

48. Kim, H.R., et al., *Bax inhibitor 1 regulates ER-stress-induced ROS accumulation through the regulation of cytochrome P450 2E1*. J Cell Sci, 2009. **122**(Pt 8): p. 1126-33.
49. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2009. **37**(Database issue): p. D5-15.
50. Taft, S.C. and A.A. Weiss, *Toxicity of anthrax toxin is influenced by receptor expression.* Clin Vaccine Immunol, 2008. **15**(9): p. 1330-6.
51. Xu, Q., E.D. Hesek, and M. Zeng, *Transcriptional stimulation of anthrax toxin receptors by anthrax edema toxin and Bacillus anthracis Sterne spore.* Microb Pathog, 2007. **43**(1): p. 37-45.
52. Naidoo, N., *ER and aging-Protein folding and the ER stress response.* Ageing Res Rev, 2009. **8**(3): p. 150-9.
53. Schmidt, D., et al., *ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions.* Methods, 2009. **48**(3): p. 240-8.
54. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-46.
55. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
56. Blanchette, M., et al., *Reconstructing large regions of an ancestral mammalian genome in silico.* Genome Res, 2004. **14**(12): p. 2412-23.
57. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
58. Ralser, M., et al., *An efficient and economic enhancer mix for PCR.* Biochem Biophys Res Commun, 2006. **347**(3): p. 747-51.
59. MacArthur, S. and J.F. Brookfield, *Expected rates and modes of evolution of enhancer sequences.* Mol Biol Evol, 2004. **21**(6): p. 1064-73.
60. Smith, N.G., M. Brandstrom, and H. Ellegren, *Evidence for turnover of functional noncoding DNA in mammalian genome evolution.* Genomics, 2004. **84**(5): p. 806-13.
61. Stone, J.R. and G.A. Wray, *Rapid evolution of cis-regulatory sequences via local point mutations.* Mol Biol Evol, 2001. **18**(9): p. 1764-70.
62. Dermitzakis, E.T. and A.G. Clark, *Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.* Mol Biol Evol, 2002. **19**(7): p. 1114-21.
63. Balmer, J.E. and R. Blomhoff, *Evolution of transcription factor binding sites in mammalian gene regulatory regions: handling counterintuitive results.* J Mol Evol, 2009. **68**(6): p. 654-64.
64. Jin, W., et al., *The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster.* Nat Genet, 2001. **29**(4): p. 389-95.

7.2. Supplemental material

HEK293					firefly/renilla ratios																
	average assay 1	average assay 2	average assay 3	average assay 4	Assays																
TMBIM6																					
normalization divisor	**5,93**	**1,83**	**12,88**		**1**				**2**								**3**				
Human	11,19	3,27	23,34		11,88	10,84	11,13	10,91	3,48	3,39	3,68	2,97	2,95	3,25	3,03	3,40	3,27	20,94	22,96	25,34	25,06	22,40
Human mutated	8,36	2,34	17,08		9,60	8,06	8,11	7,69	2,38	2,16	2,40	2,25	2,53	1,99	2,13	2,91	2,34	18,25	17,10	17,92	14,62	17,49
Chimpanzee	8,13	2,52	19,30		8,26	8,53	8,10	7,65	2,84	2,57	2,92	2,52	2,10	2,38	2,66	2,14	2,52	19,31	18,55	20,82	20,38	17,44
Chimpanzee mutated	5,93	1,83	12,88		6,02	5,76	6,71	5,22	1,92	1,67	1,93	1,70	1,66	1,73	2,02	1,97	1,83	13,30	13,93	11,71	11,95	13,51
Macaque	7,12	2,29	15,86		7,65	6,60	7,32	6,91	2,72	2,30	2,30	2,14	1,99	2,13	2,11	2,62	2,29	15,55	16,27	16,28	16,01	15,17
Macaque mutated	9,52	2,89	17,06		8,42	9,34	11,37	8,97	3,09	3,02	3,26	2,53	2,66	2,91	2,85	2,81	2,89	15,95	17,25	17,47	16,03	18,58
Empty vector	0,07	0,13	0,08		0,07	0,07	0,07	0,08	0,02	0,02	0,02	0,02	0,02	0,02	0,03	0,95	0,13	0,06	0,08	0,08	0,08	0,09
ZNF398																					
normalization divisor	**0,40**	**0,83**	**0,06**		**1**					**2**			**3**								
Human	0,73	2,27	0,14		0,71	0,69	0,70	0,70	0,87	2,13	2,45	2,23	0,15	0,14	0,13						
Human mutated	0,40	0,83	0,06		0,39	0,35	0,39	0,40	0,45	0,81	0,88	0,79	0,06	0,06	0,06						
Chimpanzee	0,84	1,35	0,12		0,79	0,79	0,78	0,90	0,95	1,54	1,41	1,10	0,12	0,12	0,12						
Chimpanzee mutated	1,23	3,62	0,24		1,40	1,26	1,08	1,18	1,22	3,94	3,58	3,34	0,25	0,24	0,24						
Macaque	0,71	2,10	0,11		0,73	0,67	0,61	0,79	0,74	1,81	2,06	2,44	0,11	0,11	0,11						
Macaque mutated	2,76	8,42	0,76		2,74	2,51	2,80	2,72	3,01	7,83	8,41	9,01	0,73	0,78	0,76						
Empty vector	0,05	0,07	0,01		0,04	0,05	0,06	0,06	0,05	0,07	0,07	0,07	0,00	0,01	0,01						
ZNF425																					
normalization divisor	**2,17**	**5,63**	**0,48**	**0,49**	**1**					**2**			**3**			**4**					
Human	2,51	7,67	0,62	0,60	2,46	2,23	2,67	2,52	2,69	6,62	7,40	8,98	0,62	0,65	0,61	0,63	0,57	0,61			
Human mutated	2,17	5,63	0,48	0,49	2,23	2,07	2,16	2,24	2,14	5,84	4,90	6,16	0,48	0,50	0,47	0,52	0,48	0,47			
Chimpanzee	3,09	8,42	0,71	0,72	3,21	2,88	2,74	3,21	3,41	7,84	9,56	7,87	0,73	0,70	0,71	0,70	0,74	0,72			
Chimpanzee mutated	4,15	10,33	0,86	0,90	4,66	3,95	3,72	4,08	4,33	9,03	11,99	9,98	0,88	0,83	0,86	0,87	0,90	0,92			
Macaque	2,99	7,12	0,70	0,78	3,07	2,82	2,91	3,04	3,12	6,04	7,93	7,38	0,70	0,67	0,72	0,80	0,77	0,76			
Macaque mutated	4,02	8,98	0,88	0,93	3,80	3,59	4,11	4,43	4,16	8,99	8,16	9,78	0,85	0,88	0,90	0,92	0,94	0,93			
Empty vector	0,05	0,07	0,01	0,01	0,04	0,05	0,06	0,06	0,05	0,07	0,07	0,07	0,01	0,01	0,01	0,00	0,01	0,01			
ZNF197																					
normalization divisor	**8,89**	**0,55**	**0,61**		**1**			**2**			**3**										
Human	14,75	1,05	0,88		14,24	14,29	15,71	1,02	1,07	1,05	0,82	0,91	0,90								
Human mutated	11,95	1,16	1,12		11,01	12,94	11,91	1,22	1,16	1,11	1,01	1,13	1,21								
Chimpanzee	9,88	0,75	0,81		9,70	7,57	12,37	0,73	0,82	0,71	0,80	0,81	0,80								
Chimpanzee mutated	10,96	0,92	0,98		10,78	10,67	11,42	0,86	0,97	0,94	0,97	1,01	0,96								
Macaque	8,89	0,55	0,61		8,73	9,39	8,56	0,54	0,53	0,58	0,57	0,66	0,59								
Macaque mutated	11,52	0,67	0,81		10,90	11,10	12,54	0,63	0,70	0,68	0,80	0,82	0,81								
Empty vector	0,07	0,01	0,01		0,07	0,07	0,07	0,01	0,01	0,01	0,00	0,01	0,01								
ANTXR1																					
normalization divisor	**3,17**	**0,24**	**0,25**		**1**			**2**			**3**										
Human	4,25	0,30	0,25		3,91	4,12	4,71	0,29	0,30	0,30	0,24	0,26	0,26								
Human mutated	3,56	0,27	0,27		3,33	3,69	3,66	0,28	0,27	0,27	0,25	0,29	0,25								
Chimpanzee	3,17	0,30	0,31		3,30	3,04	3,15	0,31	0,31	0,29	0,30	0,34	0,31								
Chimpanzee mutated	5,16	0,34	0,32		4,51	5,33	5,63	0,33	0,35	0,34	0,32	0,31	0,33								
Macaque	3,40	0,24	0,26		3,23	3,39	3,58	0,24	0,24	0,24	0,26	0,26	0,26								
Macaque mutated	3,43	0,29	0,28		2,93	3,66	3,68	0,28	0,30	0,28	0,29	0,27	0,28								
Empty vector	0,07	0,01	0,01		0,07	0,07	0,07	0,006	0,011	0,01	0,00	0,01	0,01								

Supplementary table 1. Firefly to Renilla ratios observed in HEK293 cells.

COS1				firefly/renilla ratios												
	average assay 1	average assay 2	average assay 3	Assays												
TMBIM6																
normalization divisor	**15,85**	**2,00**		**1**				**2**								
Human	44,87	2,00		37,07	42,30	43,65	56,47	15,85	13,15	15,61	16,58	17,57	15,51	13,41	19,83	15,94
Human mutated	23,66	15,94		21,91	24,84	28,95	18,92	8,31	9,20	7,79	10,98	10,19	8,30	4,48	7,47	8,34
Chimpanzee	31,50	8,34		33,52	29,18	25,22	38,07	9,20	10,35	7,72	9,76	10,00	9,70	10,65	11,07	9,81
Chimpanzee mutated	15,85	9,81		15,11	18,90	13,91	15,49	4,95	5,82	6,54	6,43	6,08	5,00	6,45	6,80	6,01
Macaque	24,43	6,01		23,51	32,58	25,47	16,15	5,81	6,88	7,80	6,33	8,23	6,78	7,22	6,14	6,90
Macaque mutated	29,78	6,90		36,60	26,51	22,73	33,30	13,82	12,76	12,67	11,45	12,85	12,76	13,59	16,56	13,31
Empty vector	0,24	13,31		0,23	0,23	0,23	0,26	0,05	0,06	0,05	0,04	0,05	0,04	0,04	0,08	0,05
ZNF398																
normalization divisor	**0,06**	**0,15**		**1**			**2**									
Human	0,18	0,49		0,15	0,21	0,19	0,47	0,52	0,49							
Human mutated	0,06	0,15		0,06	0,06	0,06	0,13	0,17	0,13							
Chimpanzee	0,06	0,25		0,07	0,06	0,07	0,25	0,24	0,27							
Chimpanzee mutated	0,47	1,10		0,45	0,46	0,49	1,07	1,25	0,98							
Macaque	0,21	0,62		0,21	0,21	0,22	0,60	0,63	0,62							
Macaque mutated	1,17	2,65		1,11	1,15	1,26	2,70	2,90	2,34							
Empty vector	0,07	0,03		0,09	0,11	0,02	0,04	0,03	0,03							
ZNF425																
normalization divisor	**0,60**	**1,68**	**2,04**	**1**			**2**			**3**						
Human	1,06	2,80	3,37	1,03	1,16	1,00	2,99	2,86	2,56	3,47	3,43	3,20				
Human mutated	0,60	1,68	2,04	0,64	0,62	0,55	1,75	1,41	1,89	1,91	2,22	2,00				
Chimpanzee	1,63	3,60	3,66	1,92	1,79	1,18	3,93	2,81	4,06	3,77	3,82	3,39				
Chimpanzee mutated	2,15	4,73	5,29	2,20	2,09	2,16	5,14	4,29	4,77	4,81	5,48	5,58				
Macaque	1,22	2,70	3,55	1,35	1,08	1,22	2,93	2,45	2,73	3,93	3,20	3,53				
Macaque mutated	2,26	4,19	4,97	2,14	2,47	2,18	4,46	4,15	3,97	4,88	5,45	4,58				
Empty vector	0,07	0,02	0,03	0,09	0,11	0,02	0,02	0,02	0,02	0,04	0,03	0,03				
ZNF197																
normalization divisor	**0,80**	**1,72**	**2,00**	**1**			**2**			**3**						
Human	1,61	3,07	2,82	1,75	1,54	1,54	2,95	3,03	3,23	1,96	3,42	3,07				
Human mutated	1,52	2,74	3,09	1,43	1,64	1,50	2,79	2,17	3,27	2,78	3,06	3,43				
Chimpanzee	0,96	1,72	2,00	0,81	1,11	0,95	1,71	1,77	1,70	1,78	2,07	2,16				
Chimpanzee mutated	0,99	3,00	3,04	0,84	1,07	1,07	2,99	2,48	3,52	3,38	3,41	2,32				
Macaque	0,80	2,00	2,25	0,66	0,81	0,93	1,80	1,84	2,36	2,61	1,99	2,13				
Macaque mutated	1,19	3,15	3,60	1,12	1,40	1,06	2,81	3,31	3,33	3,08	4,47	3,25				
Empty vector	0,07	0,02	0,03	0,09	0,11	0,02	0,02	0,02	0,02	0,04	0,03	0,03				
ANTXR1																
normalization divisor	**0,14**	**0,30**	**0,33**	**1**			**2**			**3**						
Human	0,15	0,42	0,43	0,13	0,15	0,17	0,39	0,45	0,42	0,39	0,51	0,40				
Human mutated	0,15	0,30	0,33	0,14	0,15	0,14	0,29	0,29	0,32	0,30	0,31	0,37				
Chimpanzee	0,14	0,31	0,33	0,14	0,13	0,15	0,37	0,30	0,27	0,34	0,30	0,36				
Chimpanzee mutated	0,21	0,45	0,47	0,19	0,20	0,25	0,42	0,43	0,50	0,48	0,48	0,44				
Macaque	0,18	0,32	0,39	0,16	0,17	0,21	0,28	0,30	0,38	0,41	0,32	0,45				
Macaque mutated	0,15	0,48	0,43	0,15	0,18	0,13	0,52	0,52	0,41	0,42	0,44	0,42				
Empty vector	0,07	0,02	0,03	0,09	0,11	0,02	0,017	0,025	0,024	0,04	0,03	0,03				

Supplementary table 2. Firefly to Renilla ratios observed in COS-1 cells.

HEK293	Normalized intensities																			
TMBIM6	**1**					**2**									**3**					
Human	2,00	1,83	1,88	1,84	1,89	1,90	1,86	2,01	1,63	1,62	1,78	1,66	1,86	1,79	1,63	1,78	1,97	1,95	1,74	
Human mutated	1,62	1,36	1,37	1,30	1,41	1,30	1,19	1,31	1,23	1,38	1,09	1,17	1,60	1,28	1,42	1,33	1,39	1,14	1,36	
Chimpanzee	1,39	1,44	1,37	1,29	1,37	1,55	1,41	1,60	1,38	1,15	1,30	1,46	1,17	1,38	1,50	1,44	1,62	1,58	1,35	
Chimpanzee mutated	1,02	0,97	1,13	0,88	1,00	1,05	0,92	1,06	0,93	0,91	0,95	1,10	1,08	1,00	1,03	1,08	0,91	0,93	1,05	
Macaque	1,29	1,11	1,23	1,17	1,20	1,49	1,26	1,26	1,17	1,09	1,17	1,15	1,44	1,25	1,21	1,26	1,26	1,24	1,18	
Macaque mutated	1,42	1,58	1,92	1,51	1,61	1,69	1,65	1,79	1,39	1,46	1,59	1,56	1,54	1,59	1,24	1,34	1,36	1,24	1,44	
Empty vector	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,52	0,07	0,00	0,01	0,01	0,01	0,01	
ZNF398	**1**					**2**			**3**											
Human	1,80	1,74	1,77	1,78	2,20	2,58	2,96	2,69	2,47	2,34	2,18									
Human mutated	0,99	0,89	0,99	1,00	1,13	0,98	1,06	0,95	0,96	1,05	0,99									
Chimpanzee	1,99	2,01	1,96	2,28	2,41	1,87	1,71	1,33	2,02	2,03	2,03									
Chimpanzee mutated	3,55	3,19	2,73	2,99	3,08	4,77	4,34	4,05	4,21	4,06	3,98									
Macaque	1,84	1,68	1,54	2,01	1,88	2,19	2,49	2,96	1,91	1,90	1,86									
Macaque mutated	6,94	6,36	7,09	6,88	7,62	9,49	10,18	10,91	12,24	12,99	12,76									
Empty vector	0,10	0,14	0,16	0,14	0,12	0,08	0,09	0,09	0,08	0,19	0,17									
ZNF425	**1**					**2**			**3**			**4**								
Human	1,13	1,03	1,23	1,16	1,24	1,18	1,31	1,59	1,28	1,34	1,25	1,27	1,16	1,25						
Human mutated	1,03	0,96	1,00	1,03	0,99	1,04	0,87	1,09	1,00	1,03	0,97	1,07	0,99	0,95						
Chimpanzee	1,48	1,33	1,26	1,48	1,57	1,39	1,70	1,40	1,51	1,45	1,46	1,43	1,50	1,47						
Chimpanzee mutated	2,15	1,82	1,71	1,88	2,00	1,60	2,13	1,77	1,83	1,72	1,78	1,77	1,84	1,88						
Macaque	1,42	1,30	1,34	1,40	1,44	1,07	1,41	1,31	1,45	1,39	1,48	1,63	1,56	1,56						
Macaque mutated	1,75	1,66	1,90	2,04	1,92	1,60	1,45	1,74	1,77	1,82	1,87	1,87	1,91	1,90						
Empty vector	0,02	0,02	0,03	0,03	0,02	0,01	0,01	0,01	0,01	0,02	0,02	0,01	0,02	0,02						
ZNF197	**1**			**2**			**3**													
Human	1,60	1,61	1,77	1,84	1,93	1,90	1,35	1,49	1,47											
Human mutated	1,24	1,46	1,34	2,19	2,09	2,00	1,66	1,87	1,99											
Chimpanzee	1,09	0,85	1,39	1,32	1,48	1,27	1,32	1,34	1,32											
Chimpanzee mutated	1,21	1,20	1,28	1,55	1,76	1,69	1,60	1,66	1,58											
Macaque	0,98	1,06	0,96	0,98	0,96	1,05	0,94	1,08	0,97											
Macaque mutated	1,23	1,25	1,41	1,14	1,27	1,22	1,32	1,34	1,33											
Empty vector	0,01	0,01	0,01	0,01	0,02	0,02	0,01	0,02	0,02											
ANTXR1	**1**			**2**			**3**													
Human	1,15	1,21	1,39	1,21	1,26	1,27	0,94	0,99	1,01											
Human mutated	0,98	1,09	1,07	1,16	1,12	1,15	0,97	1,13	0,98											
Chimpanzee	0,97	0,89	0,93	1,30	1,30	1,21	1,14	1,29	1,20											
Chimpanzee mutated	1,33	1,57	1,66	1,38	1,48	1,43	1,22	1,21	1,25											
Macaque	0,95	1,00	1,05	0,99	1,02	0,99	0,99	1,00	1,01											
Macaque mutated	0,86	1,08	1,08	1,16	1,28	1,19	1,12	1,04	1,09											
Empty vector	0,02	0,02	0,02	0,03	0,05	0,04	0,02	0,04	0,04											

Supplementary table 3. Normalized firefly to renilla ratios for HEK293 cell. For each promoter, measured activities were normalized to the construct with the lowest promoter activity level in HEK293 cells (set to one).

COS	Normalized intensities													
TMBIM6	**1**					**2**								
Human	2,34	2,67	2,75	3,56	2,83	2,64	2,19	2,60	2,76	2,92	2,58	2,23	3,30	2,65
Human mutated	1,38	1,57	1,83	1,19	1,49	1,38	1,53	1,30	1,83	1,70	1,38	0,75	1,24	1,39
Chimpanzee	2,11	1,84	1,59	2,40	1,99	1,53	1,72	1,28	1,63	1,66	1,61	1,77	1,84	1,63
Chimpanzee mutated	0,95	1,19	0,88	0,98	1,00	0,82	0,97	1,09	1,07	1,01	0,83	1,07	1,13	1,00
Macaque	1,48	2,05	1,61	1,02	1,54	0,97	1,15	1,30	1,05	1,37	1,13	1,20	1,02	1,15
Macaque mutated	2,31	1,67	1,43	2,10	1,88	2,30	2,12	2,11	1,91	2,14	2,12	2,26	2,76	2,22
Empty vector	0,01	0,01	0,01	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
ZNF398	**1**			**2**										
Human	2,48	3,49	3,28	3,24	3,55	3,33								
Human mutated	1,02	1,04	0,94	0,90	1,18	0,92								
Chimpanzee	1,16	0,98	1,14	1,68	1,67	1,85								
Chimpanzee mutated	7,55	7,86	8,38	7,33	8,55	6,72								
Macaque	3,51	3,50	3,72	4,13	4,33	4,24								
Macaque mutated	18,71	19,44	21,32	18,51	19,86	16,01								
Empty vector	1,44	1,90	0,32	0,24	0,18	0,22								
ZNF425	**1**			**2**			**3**							
Human	1,70	1,91	1,66	1,78	1,70	1,52	1,70	1,68	1,57					
Human mutated	1,07	1,02	0,91	1,04	0,84	1,12	0,93	1,09	0,98					
Chimpanzee	3,17	2,96	1,95	2,33	1,67	2,41	1,84	1,87	1,66					
Chimpanzee mutated	3,63	3,46	3,57	3,05	2,55	2,83	2,35	2,68	2,73					
Macaque	2,23	1,78	2,02	1,74	1,45	1,62	1,92	1,56	1,73					
Macaque mutated	3,54	4,08	3,60	2,65	2,46	2,36	2,39	2,67	2,24					
Empty vector	0,14	0,19	0,03	0,01	0,01	0,01	0,02	0,01	0,02					
ZNF197	**1**			**2**			**3**							
Human	2,19	1,93	1,93	1,47	1,52	1,61	0,87	1,52	1,37					
Human mutated	1,80	2,06	1,89	1,39	1,08	1,64	1,24	1,36	1,53					
Chimpanzee	1,01	1,39	1,20	0,85	0,88	0,85	0,79	0,92	0,96					
Chimpanzee mutated	1,05	1,35	1,35	1,49	1,24	1,76	1,50	1,52	1,03					
Macaque	0,82	1,01	1,16	0,90	0,92	1,18	1,16	0,89	0,95					
Macaque mutated	1,40	1,76	1,32	1,41	1,65	1,66	1,37	1,99	1,45					
Empty vector	0,11	0,14	0,02	0,01	0,01	0,01	0,02	0,01	0,01					
ANTXR1	**1**			**2**			**3**							
Human	0,74	0,82	0,92	1,23	1,41	1,31	0,99	1,28	1,01					
Human mutated	0,79	0,82	0,80	0,90	0,90	0,99	0,77	0,78	0,94					
Chimpanzee	0,76	0,70	0,84	1,15	0,94	0,84	0,86	0,76	0,91					
Chimpanzee mutated	1,05	1,08	1,38	1,30	1,33	1,57	1,23	1,21	1,10					
Macaque	0,89	0,96	1,16	0,88	0,94	1,18	1,05	0,81	1,13					
Macaque mutated	0,83	0,98	0,73	1,61	1,62	1,28	1,07	1,11	1,07					
Empty vector	0,47	0,62	0,10	0,05	0,08	0,07	0,09	0,07	0,08					

Supplementary table 4. Normalized firefly to renilla ratios for COS-1 cell. For each promoter, measured activities were normalized to the construct with the lowest promoter activity level in COS-1 cells (set to one).

	HEK293				COS-1			
	Average	SD	log2 mut/wt	Welch's test 1-tail, 2-tailed	Average	SD	log2 mut/wt	Welch's test 1-tail, 2-tailed
TMBIM6								
Human	1,82	0,12			2,72	0,37		
Human mutated	1,33	0,14	**-0,45**	**5,32E-14**	1,43	0,28	**-0,93**	**1,19E-10**
Chimpanzee	1,41	0,13		**5,56E-12**	1,76	0,27		**5,99E-08**
Chimpanzee mutated	1,00	0,08	**-0,49**	**4,73E-13**	1,00	0,11	**-0,81**	**1,34E-08**
Macaque	1,23	0,10		**1,62E-17**	1,29	0,30		**3,70E-11**
Macaque mutated	1,52	0,18	**0,30**	**4,53E-07**	2,09	0,31	**0,70**	**1,10E-07**
Empty vector	0,04	0,12			0,01	0,00		
ZNF398								
Human	2,23	0,42			3,23	0,39		
Human mutated	1,00	0,06	**-1,16**	**8,88E-07**	1,00	0,11	**-1,69**	**6,60E-06**
Chimpanzee	1,97	0,28		**1,03E-01**	1,41	0,36		**7,82E-06**
Chimpanzee mutated	3,72	0,65	**0,92**	**6,41E-07**	7,73	0,68	**2,45**	**3,53E-08**
Macaque	2,02	0,40		**2,56E-01**	3,91	0,37		**1,15E-02**
Macaque mutated	9,41	2,56	**2,22**	**9,25E-07**	18,98	1,77	**2,28**	**1,17E-06**
Empty vector	0,12	0,04			0,72	0,75		
ZNF425								
Human	1,25	0,13			1,69	0,11		
Human mutated	1,00	0,06	**-0,32**	**2,07E-06**	1,00	0,09	**-0,76**	**1,72E-10**
Chimpanzee	1,46	0,10		**6,83E-05**	2,21	0,55		**2,37E-02**
Chimpanzee mutated	1,85	0,15	**0,34**	**3,29E-08**	2,98	0,47	**0,44**	**2,79E-03**
Macaque	1,41	0,14		**3,76E-05**	1,78	0,24		**3,07E-01**
Macaque mutated	1,80	0,15	**0,35**	**9,03E-08**	2,89	0,67	**0,69**	**4,56E-04**
Empty vector	0,02	0,01			0,05	0,07		
ZNF197								
Human	1,66	0,20			1,60	0,38		
Human mutated	1,76	0,35	**0,08**	**2,43E-01**	1,55	0,32	**-0,04**	**3,87E-01**
Chimpanzee	1,26	0,19		**5,41E-04**	0,98	0,19		**1,04E-03**
Chimpanzee mutated	1,50	0,21	**0,25**	**1,11E-02**	1,37	0,24	**0,47**	**8,89E-04**
Macaque	1,00	0,05		**6,03E-06**	1,00	0,14		**1,27E-03**
Macaque mutated	1,28	0,08	**0,35**	**2,89E-07**	1,56	0,22	**0,64**	**1,05E-05**
Empty vector	0,01	0,01			0,04	0,05		
ANTXR1								
Human	1,16	0,15			1,08	0,24		
Human mutated	1,07	0,08	**-0,11**	**7,24E-02**	0,85	0,08	**-0,34**	**1,14E-02**
Chimpanzee	1,14	0,16		**2,56E-01**	0,86	0,13		**3,32E-02**
Chimpanzee mutated	1,39	0,16	**0,29**	**2,07E-03**	1,25	0,17	**0,54**	**2,88E-05**
Macaque	1,00	0,03		**7,43E-04**	1,00	0,14		**4,02E-01**
Macaque mutated	1,10	0,11	**0,14**	**1,55E-02**	1,14	0,31	**0,19**	**1,14E-01**
Empty vector	0,03	0,01			0,18	0,21		

Supplementary table 5. Average intensities, standard deviations (SD) and log2 ratios of mutated (mut) to wild type (wt) activities and significance levels for HEK293 and COS1 cells. Significant differences between wt and mutated promoter activities were calculated according to a one-tailed Welch's test, while significance of difference of wt chimpanzee or macaque promoters compared to human wt activity, was calculated according to a two-tailed Welch's test.

7.3. Contributions

Hans-Jörg Warnatz: performed the ChIP-seq and peak finding, and contributed to writing the manuscript

Ralf Sudbrak: was involved in discussion and writing of the manuscript

Hans Lehrach: was involved in discussion and writing of the manuscript

Marie-Laure Yaspo: was involved in discussion and writing of the manuscript

8. Discussion

The overall context of the presented manuscripts comprises promoter analysis by experimental and bioinformatics means. Among the first experimental obstacles encountered was the difficulty in amplification of GC-rich promoters, which evoked the optimization of PCR conditions including to puzzle out a PCR enhancer mix.

The derived protocol was pivotal to the amplification of human chromosome 21 promoters. The amplified promoters were cloned for studying promoter activities under different conditions, to elucidate the impact of different promoter elements, and to examine possibilities and limitations of the cell-array technology. Several findings of this study were valuable to the design of the next study aiming at the identification of human and hominid specific TFBSs.

First, cell-array technology is not suitable for the quantification of promoter strength. However, an important finding was that promoter fragments of ~0.5 kb in length were always sufficient to drive reporter gene expression. The analysis and integration of 2nd-generation sequencing data from RNA-seq and RNAPII ChIP-seq, allows accurate mapping of TSSs and quantification of expression. Profiting from these findings it was possible for me to select and precisely clone promoters with human- and hominid-specific GABPa binding sites.

8.1. Promoter analysis

The only definitive means of promoter characterization involves cloning of putative promoter regions, followed by *in vivo* functional assays, typically by transient or stable transformation together with a reporter gene [8]. The first step requires the localization of promoters, which due to their predictable location immediately upstream of TSS can be achieved relatively straightforward.

However, in the human genome TSS annotation is far from complete [114]. The difficulties in reliable TSS annotation originate from the 3′ bias in isolation and synthesis of cDNAs [115] and the existence of alternative promoters regulating alternative mRNA isoforms [116]. Knowledge of 5'UTR length and alternative promoter usage are valuable pieces of information to verify annotated TSS coordinates prior to promoter studies, allowing for accurate cloning to enhance experimental readout. In this respect, genome-wide profiling of regions bound by components of the PIC and mapping of active genes (by ChIP-seq and RNA-seq) within the organism or cell line under investigation represent valuable resources to explore gene activity patterns prior to single-gene functional studies. On the long term, RNA-seq and ChIP-seq will widely replace real-time- and RACE-PCR as well as techniques developed to capture full-length mRNAs for mapping of 5'ends, such as 5' SAGE (5'-end serial analysis of gene expression) [117] and CAGE (cap analysis gene expression) [118].

Among the bottlenecks of large-scale promoter studies are amplification and cloning, especially when aiming at studying large promoter fragments including distal promoter regions. Amplification is frequently hampered by high GC content [119] especially found in CpG islands, which locate close to the TSS of the majority of the human genes. Since CpG islands are prone to form super-structures, display high melting temperatures, and re-hybridization of complementary strands occurs quickly, they can strongly inhibit PCR [120].

The first manuscript presents a PCR enhancer mix that, together with the corresponding PCR protocol and primer design, represents an efficient strategy for the amplification of such regions. The components of the enhancer mix contribute to lowering the melting temperatures and thereby inhibit secondary structure formation and re-hybridization. Aside from that, to keep temperatures during PCR cycles high, we designed primers with melting temperatures in the range of 68-72°C. Finally, by using a touchdown PCR program, implying a successively lower annealing temperature for each cycle, we ensure accurate initial annealing and thereby specific amplification of the desired target region. The importance of this effective protocol is underlined by the finding that a rate-limiting step in gene regulation is activation of stalled

RNAPII [27], which occurs primarily at CpG islands [33]. RNAPII stalling is widespread, occurring at thousands of genes that respond to stimuli and developmental signals [13, 27, 32]. According to this, CpG-rich promoters might even more shift into the focus of functional promoter studies.

Our analysis on activities of cloned 2.5 kb fragments of human chromosome 21 promoters in HEK293 cells, including 2nd-generation sequencing data, resulted in several relevant findings. As might be expected, we found transiently transfected promoters active if the corresponding gene was endogenously expressed and the cloned fragment covered the employed TSS in HEK293 cells.

We further tested if transfected promoter constructs respond to external stimuli by treatment of cells with Trichostatin A (TSA) or depletion of fetal calf serum (FCS). Depletion of serum represents a stress condition that induces cell type-specific responses affecting cell cycle regulation, apoptosis, cell growth, and cell differentiation [121-123]. Indeed, serum depletion activated 40 of the previously inactive promoters. Among these promoters, we searched for common transcription factors binding sites that might explain these responses. Interestingly, we found a significant enrichment of NF-kappaB binding sites, a factor that is known to be activated upon serum starvation [124].

On the other hand, we treated cells with Trichostatin A, a histone deacetylase inhibitor that activates transcription from repressed chromosomal regions and also has been shown to activate the transcription of genes silenced by DNA methylation through inhibition of DNA methyltransferase DNMT1 [125]. Among the promoters activated by TSA treatment, we find a highly significant fraction containing CpG islands, indicating that promoter-reporter constructs are sensible to endogenous DNA methylation-mediated silencing. Recently, it was reported that endogenous CpG methylation can occur in less than an hour [126], which could indicate that methylase-deficient *E.coli* strains are not necessary for promoter reporter gene assays, since methylation patterns will be adjusted endogenously after transfection.

Another observation was that truncation of promoters to the proximal ~0.5 kb hardly resulted in loss of activity. We observed the same activity patterns in 2/3 (41) of the tested long and corresponding short promoters. However, truncation frequently resulted in loss of the potential to respond to external stimuli, as the activity of 21 long promoters changed following one of the treatments, while only three corresponding truncated promoters responded. This finding hints towards the presence of *cis*-regulatory response elements residing in distal promoter regions. Hence, proximal promoter regions are sufficient to drive gene expression under standard cell culture conditions. However, long promoters are more likely to integrate endogenous signaling pathways into reporter gene expression than short ones.

A relevant observation in respect to reporter gene assays, discussed in the third manuscript, was the finding that activity patterns of the different promoter constructs in human HEK293 cells were almost entirely reproduced in african green monkey derived COS-1 cells. This observation is in line with the findings of Wilson et al. who used hepatocytes from an aneuploid mouse strain that carried the human chromosome 21 to test whether interspecies differences in transcriptional regulation are primarily caused by *cis*- or *trans*-acting mechanisms. They found that: "*Virtually all transcription factor-binding locations, landmarks of transcription initiation, and the resulting gene expression observed in human hepatocytes were recapitulated across the entire human chromosome 21 in the mouse hepatocyte nucleus*" [127]. Therefore, also in transient transfection assays, *cis*-acting elements seem to be largely responsible for directing transcriptional output, allowing to study promoter activities of related species, especially primates, within the same cell line.

Even though *ex vivo* promoter reporter gene assays cannot account for endogenous signaling during organism development, together our findings further underline the importance and suitability of transient transfection assays in studying *ex vivo* promoter activity and response. In particular, this might be valid for studying genes involved in cellular homeostasis, as they rely to a greater extent on intracellular signals and are less likely targets of signals passed through developing organisms.

8.2. Lineage-specific transcription factor binding sites

Numerous examples within different organisms underline the fact that mutations in *cis*-regulatory regions cause a variety of interesting and ecologically significant phenotypic differences in morphology, physiology and behavior [64]. Hence, the identification of TFBS alterations with functional consequences is fundamental to the understanding of species-specific traits and evolution. The major obstacles in discovering *cis*-regulatory adaptations are the pinpointing of potentially relevant substitutions and subsequently, their functional validation. Genome-wide bioinformatics approaches alone suffer mainly from false positive predictions caused by the shortness and high sequence degeneration of many TFBSs as well as their strong context-dependency. However, regarding particular genes or regions of interest, TFBS prediction can be successfully applied [128, 129].

Today, the evaluation of regulatory mutations within an organism, including the entire array of functional consequences, is not possible, as this would imply to monitor not only direct effects on the regulated genes, but also all downstream effects during development and life. However, for model organisms, such as *C. elegans* and *D. melanogaster*, reporter-gene assays are successfully in use to trace expression patterns during development and life [130, 131]. For many model organisms, efficient approaches exist for the delivery of reporter gene constructs, yet for higher organisms, especially mammals, such assays are work-intensive, time-consuming and not applicable to chimpanzees or humans. For humans, the only way to assay the impact of regulatory mutations affecting the binding of a specific TF is *ex vivo*, while placing a mutated promoter-reporter and a wild-type control construct into a human cell line, where it is exposed to the array of transcription factors that is also encountered by the endogenous promoter [8]. As mentioned above, such assays cannot account for all types of transcriptional regulation, especially not those occurring during development. However, our findings indicate that promoter-reporter constructs can potentially integrate various regulatory mechanisms, including CpG methylation, nucleosome derangement and RNAPII stalling, allowing for suitable mapping of endogenous transcriptional regulation.

For our approach in finding human- and hominid-specific TFBSs, we chose to study the endogenous binding of the transcription factor GABPa in human HEK293 cells. Besides the considerable pre-existing knowledge on this TF, it is ideal for the functional validation of candidate TFBS alterations for two reasons. First, GABPa is a strong transcriptional activator, and second, GABPa binds preferentially in close proximity to the TSS, which is important for

functional evaluation, since cloning of long inserts and cell transfection with large plasmids is more complicated [129].

GABPa is known to bind to proximal promoters of thousands of genes in different cell lines [105, 132]. In line with this, our analysis revealed that one third of the genes expressed in HEK293 cells show signals of GABPa bound to their proximal promoters. This region is pivotal for transcription initiation as underlined by several findings. The proximal promoter is considerably conserved among mammals and remarkably enriched for transcription factor binding sites, which becomes even more pronounced when considering only phylogenetically conserved TFBSs [17, 133, 134]. Furthermore, residing SNPs are more likely involved in transcriptional regulation than others residing further upstream [135]. In addition, SNPs add up to 72% of known functional *cis*-regulatory mutations in human [69]. Therefore, we aimed at identifying single nucleotide mutations that occurred during human evolution and have created functional GABPa binding sites.

The approach is based on the characterization of genomic regions that are bound *ex vivo* by GABPa. However, it appears unlikely that a significant fraction of the GABPa regulated gene-promoters are not regulated by GABPa *in vivo*, at least in one of the hundreds of cell types. Therefore, the derived binding preferences of GABPa very likely picture *in viv*o preferences, and in addition a significant fraction of the thousands of gene-promoters recognized by GABPa in HEK293 cells, will be similarly GABPa-bound *in vivo*.

The implemented bioinformatics approach to analyze ChIP-seq data and identify human specific TFBSs is straightforward. We used DNA regions that were bound by GABPa in HEK293 cells to calculate a GABPa consensus-binding matrix. Subsequently, the same regions were scanned to find all occurrences of this consensus sites at a particular threshold. Then, we obtained multiple species alignments from USCS whole genome alignments, corresponding to the predicted human binding sites within the GABPa bound regions. To address the question, which sites evolved on the lineage leading to humans, we reconstructed the ancestral sequences of human binding sites based on the multiple species along the entire mammalian phylogeny. For this purpose, we used ACESTORS, an algorithm that has been shown suitable for reconstructing ancestral sequences, including the most likely scenario of insertions and deletions observed in alignments, while retaining an extremely high degree of accuracy [136, 137]. Next, we searched the ancestral sequences for the presence of GABPa BSs to find those BSs that have emerged in the human, hominini, hominiae or hominid lineages. For these four lineages, we found 224, 57, 244 and 310 specific BSs, out of 11,008 sites in total.

However, the particular focus of this study lies in sites specific to the human species. The annotation of human-specific sites is most reliable, as the reconstruction of the hominini

sequence is accurate, depending solely on genome sequences and not on reconstructed sequences, as is the case for the deeper lineages of homininae and hominid.

To address the question whether the genes associated with 224 human specific BSs show any functional relations, we searched for enrichment in corresponding gene ontology terms, tissue expression patterns and protein domains. We found enrichment of genes involved in RNA processing, genes expressed in mammary and pineal gland, and of genes containing a KRAB protein domain. Adaptations in the regulation of these genes have likely occurred during human speciation, as for example different needs of newborns for nutritional and immunological components require changes in milk composition (Lemay, Lynn et al. 2009). Similarly, gene expression in the pineal gland, involved in circadian rhythm, growth, puberty and aging (Pierpaoli 1998), has likely changed during human speciation. The KRAB domain serves to recruit histone deacetylase complexes to regions surrounding the DNA-binding sites [138], leading to repression of transcription [138-140]. KRAB-associated zinc finger proteins thus function as potent transcriptional repressors [141]. This functional similarity is not very specific. However, KRAB zinc-fingers represent a group of genes specific to tetrapodes [142] and have expanded in primates, mainly driven through gene duplication [143].

New genes are believed to be free to evolve, including for new sets of regulatory elements [144], and since GABPa represents a strong transcriptional activator [145], here BS gain might indicate that evolution favored higher transcription rates of the KRAB genes. Together these findings hint towards further functional similarities of KRAB-ZFs beyond their general transcription repressor activity. Another interesting finding was that GABPa binds to the promoters of only one third of the genes expressed in HEK293 cells, but to the promoters of 65% of the expressed KRAB-ZFs. This very significant enrichment indicates a general role of GABPa in regulation of KRAB-ZF expression in HEK293 cells and deserves further investigation, in particular *in vivo* and in respect to development.

To test whether human specific GABPa BS, identified through our approach, influence gene expression, four promoters were functionally tested by dual luciferase assays, including a bi-directional promoter. These promoters correspond to five genes, ANTXR1 and TMBIM6, and three KRAB ZFs, namely ZNF197 and ZNF398/ZNF425 located head to head. Functional testing involved the comparison of wild type promoters of human, chimpanzee and macaque and testing for the influence of human-specific BSs by site directed mutagenesis. For this, human promoters were modified by single nucleotide mutations to mirror the chimpanzee sequence devoid of the GABPa BS, while vice versa, chimpanzee and macaque promoters were mutated to build the human specific BS.

The creation of human specific BSs within chimpanzee and macaque promoter backgrounds consistently resulted in significantly elevated reporter gene activities. On the other hand, disruption of the human specific sites in human promoters resulted in significant decrease in three cases (including both directions of the bidirectional promoter), while no significant activity decrease was observed in two cases. For these two gene promoters of ZNF197 and ANTXR1, it is possible that other human-specific mutations create or modify adjacent BSs for factors that are compensating the activating potential of the specific GABPa BSs. Indeed, within both of the cloned promoter fragments of ZNF197 and ANTXR1, three such mutations exist, which can be addressed in subsequent experiments. On the other hand, the human bi-directional promoter of ZNF398/ZNF425 and the promoter of TMBIM6 showed significant activity decrease when BSs were ancestralized. Hence, no compensating mutations reside within the cloned promoter fragments, rendering a functional importance of the human-specific TFBS *in vivo* more likely compared to ZNF197 and ANTXR1.

TMBIM6 is special within the analysis, as human and chimpanzee promoters share a GBAPa BS that is absent in non-hominid primates. Still, the human promoter drives higher reporter gene activity than the chimpanzee promoter, while the macaque promoter results in even lower expression. The cloned human promoter differs in seven substitutions from the chimpanzee promoters, some of which will be responsible for the differences in human/chimpanzee promoter strengths. Taken together, the cloned TMBIM6 promoter gained a functional GABPa BS in hominids, while the human promoter gained one or more additional BSs that further increase promoter strength. TMBIM6 is known as an anti-apoptotic protein protecting from apoptosis induced by endoplasmatic reticulum stress (ER-stress) by reducing accumulation of reactive oxygen species (ROS) [146]. ER-stress has been implicated in the development of diabetes, atherosclerosis and in many of the aging-related neurodegenerative diseases, such as Alzheimer's, amyotrophic lateral sclerosis and Parkinson's [147]. Hence, changes in the regulation of TMBIM6 expression might play a role in allowing long life spans of hominids and particularly for man. Together these findings render the regulation of TMBMI6 expression an interesting subject for further investigation.

In summary, this work presents an efficient approach to the identification of lineage-specific TFBSs, with evidence for functional impact of identified sites on transcription regulation. Limitations of this strategy rest in the capacities for functional testing and in ChIP experiments, as suitable antibodies are not yet available for the vast majority of TFs. However, new approaches are on the way, including expression of TFs fused to short epitope tags for efficient imunoprecipitations [148]. On the other hand, ChIP-seq studies uncovering thousands of *in vivo* BSs of single TFs will allow for refined bioinformatic models of TF binding preferences, lifting TFBSs predictions to the next level, away from *in vitro*-derived binding models. Finally,

bringing together lineage-specific TFBSs with the growing body of data on expression profiles, protein interactions, gene functions, regulatory pathways and disease associations, as exemplified in this work, will reveal many more mutations involved in disease and the evolution of species-specific traits.

9. Bibliography

1. Arbeitman, M.N., et al., *Gene expression during the life cycle of Drosophila melanogaster.* Science, 2002. **297**(5590): p. 2270-5.
2. White, K.P., et al., *Microarray analysis of Drosophila development during metamorphosis.* Science, 1999. **286**(5447): p. 2179-84.
3. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.* Nature, 2001. **409**(6819): p. 533-8.
4. Mody, M., et al., *Genome-wide gene expression profiles of the developing mouse hippocampus.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8862-7.
5. Kayo, T., et al., *Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5093-8.
6. Alberts, B., J.H. Wilson, and T. Hunt, *Molecular biology of the cell*. 5th ed., Reference ed. ed. 2008, New York, N.Y. ; Abingdon: Garland Science. xxxiii, 1601, [90] p.
7. Lewin, B., *Genes VIII*. Instructor's ed. ed. 2004, Upper Saddle River, N.J.: Pearson Prentice Hall. xxi, 1027 p.
8. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes.* Mol Biol Evol, 2003. **20**(9): p. 1377-419.
9. Carey, M. and S.T. Smale, *Transcriptional regulation in eukaryotes : concepts, strategies, and techniques*. 1999, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
10. Latchman, D.S., *Gene regulation : a eukaryotic perspective*. 4th ed. ed. 2002, Cheltenham: Nelson Thornes. xviii, 323 p., [8] p. of plates.
11. Lemon, B., et al., *Selectivity of chromatin-remodelling cofactors for ligand-activated transcription.* Nature, 2001. **414**(6866): p. 924-8.
12. White, R.J., *Gene transcription : mechanisms and control*. 2001, Oxford: Blackwell Science. xii, 273 p.
13. Muse, G.W., et al., *RNA polymerase is poised for activation across the genome.* Nat Genet, 2007. **39**(12): p. 1507-11.
14. Zeitlinger, J., et al., *RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo.* Nat Genet, 2007. **39**(12): p. 1512-6.
15. Sims, R.J., 3rd, R. Belotserkovskaya, and D. Reinberg, *Elongation by RNA polymerase II: the short and long of it.* Genes Dev, 2004. **18**(20): p. 2437-68.
16. Juven-Gershon, T., et al., *The RNA polymerase II core promoter - the gateway to transcription.* Curr Opin Cell Biol, 2008. **20**(3): p. 253-9.
17. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
18. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1412-7.
19. Bock, C., et al., *CpG island mapping by epigenome prediction.* PLoS Comput Biol, 2007. **3**(6): p. e110.
20. Wu, J.Q. and M. Snyder, *RNA polymerase II stalling: loading at the start prepares genes for a sprint.* Genome Biol, 2008. **9**(5): p. 220.

21. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
22. Ramirez-Carrozzi, V.R., et al., *Selective and antagonistic functions of SWI/SNF and Mi-2beta nucleosome remodeling complexes during an inflammatory response.* Genes Dev, 2006. **20**(3): p. 282-96.
23. Ramirez-Carrozzi, V.R., et al., *A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling.* Cell, 2009. **138**(1): p. 114-28.
24. Weinmann, R., H.J. Raskas, and R.G. Roeder, *Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection.* Proc Natl Acad Sci U S A, 1974. **71**(9): p. 3426-39.
25. Matsui, T., et al., *Multiple factors required for accurate initiation of transcription by purified RNA polymerase II.* J Biol Chem, 1980. **255**(24): p. 11992-6.
26. Margaritis, T. and F.C. Holstege, *Poised RNA polymerase II gives pause for thought.* Cell, 2008. **133**(4): p. 581-4.
27. Gilmour, D.S., *Promoter proximal pausing on genes in metazoans.* Chromosoma, 2009. **118**(1): p. 1-10.
28. Gilmour, D.S. and J.T. Lis, *RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells.* Mol Cell Biol, 1986. **6**(11): p. 3984-9.
29. Wu, C., et al., *Purification and properties of Drosophila heat shock activator protein.* Science, 1987. **238**(4831): p. 1247-53.
30. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells.* Cell, 2007. **130**(1): p. 77-88.
31. Hargreaves, D.C., T. Horng, and R. Medzhitov, *Control of inducible gene expression by signal-dependent transcriptional elongation.* Cell, 2009. **138**(1): p. 129-45.
32. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
33. Hendrix, D.A., et al., *Promoter elements associated with RNA Pol II stalling in the Drosophila embryo.* Proc Natl Acad Sci U S A, 2008. **105**(22): p. 7762-7.
34. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.
35. Thomas, M.C. and C.M. Chiang, *The general transcription machinery and general cofactors.* Crit Rev Biochem Mol Biol, 2006. **41**(3): p. 105-78.
36. Chi, T., *A BAF-centred view of the immune system.* Nat Rev Immunol, 2004. **4**(12): p. 965-77.
37. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
38. Rezai-Zadeh, N., et al., *Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1.* Genes Dev, 2003. **17**(8): p. 1019-29.
39. Roeder, R.G. and W.J. Rutter, *Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms.* Nature, 1969. **224**(5216): p. 234-7.
40. Roeder, R.G. and W.J. Rutter, *Specific nucleolar and nucleoplasmic RNA polymerases.* Proc Natl Acad Sci U S A, 1970. **65**(3): p. 675-82.

41. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* EMBO J, 2004. **23**(20): p. 4051-60.
42. Gershenzon, N.I. and I.P. Ioshikhes, *Synergy of human Pol II core promoter elements revealed by statistical sequence analysis.* Bioinformatics, 2005. **21**(8): p. 1295-300.
43. Smale, S.T. and J.T. Kadonaga, *The RNA polymerase II core promoter.* Annu Rev Biochem, 2003. **72**: p. 449-79.
44. Chalkley, G.E. and C.P. Verrijzer, *DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator.* EMBO J, 1999. **18**(17): p. 4835-45.
45. Shao, H., et al., *Core promoter binding by histone-like TAF complexes.* Mol Cell Biol, 2005. **25**(1): p. 206-19.
46. Muller, F., M.A. Demeny, and L. Tora, *New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors.* J Biol Chem, 2007. **282**(20): p. 14685-9.
47. Saunders, A., L.J. Core, and J.T. Lis, *Breaking barriers to transcription elongation.* Nat Rev Mol Cell Biol, 2006. **7**(8): p. 557-67.
48. Rasmussen, E.B. and J.T. Lis, *In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes.* Proc Natl Acad Sci U S A, 1993. **90**(17): p. 7923-7.
49. Peterlin, B.M. and D.H. Price, *Controlling the elongation phase of transcription with P-TEFb.* Mol Cell, 2006. **23**(3): p. 297-305.
50. Yang, Z., et al., *Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4.* Mol Cell, 2005. **19**(4): p. 535-45.
51. Triezenberg, S.J., *Structure and function of transcriptional activation domains.* Curr Opin Genet Dev, 1995. **5**(2): p. 190-6.
52. Torchia, J., C. Glass, and M.G. Rosenfeld, *Co-activators and co-repressors in the integration of transcriptional responses.* Curr Opin Cell Biol, 1998. **10**(3): p. 373-83.
53. Gstaiger, M., et al., *A B-cell coactivator of octamer-binding transcription factors.* Nature, 1995. **373**(6512): p. 360-2.
54. Shiama, N., *The p300/CBP family: integrating signals with transcription factors and chromatin.* Trends Cell Biol, 1997. **7**(6): p. 230-6.
55. Wolffe, A.P., *Transcriptional regulation in the context of chromatin structure.* Essays Biochem, 2001. **37**: p. 45-57.
56. Bjorklund, S. and C.M. Gustafsson, *The yeast Mediator complex and its regulation.* Trends Biochem Sci, 2005. **30**(5): p. 240-4.
57. Kornberg, R.D., *Mediator and the mechanism of transcriptional activation.* Trends Biochem Sci, 2005. **30**(5): p. 235-9.
58. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.
59. Locker, J., *Transcription factors*. Human molecular genetics series. 2001, Chichester: BIOS. xvi, 336 p.
60. Azumi, K., et al., *Gene expression profile during the life cycle of the urochordate Ciona intestinalis.* Dev Biol, 2007. **308**(2): p. 572-82.
61. Wilkins, A.S., *The evolution of developmental pathways*. 2002, Sunderland, Mass.: Sinauer Associates. xvii, 603 p.
62. Kammandel, B., et al., *Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity.* Dev Biol, 1999. **205**(1): p. 79-97.

63. Pirkkala, L., P. Nykanen, and L. Sistonen, *Roles of the heat shock transcription factors in regulation of the heat shock response and beyond.* FASEB J, 2001. **15**(7): p. 1118-31.
64. Wray, G.A., *The evolutionary significance of cis-regulatory mutations.* Nat Rev Genet, 2007. **8**(3): p. 206-16.
65. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.
66. Carroll, S.B., J.K. Grenier, and S.D. Weatherbee, *From DNA to diversity : molecular genetics and the evolution of animal design*. 2001, Malden, Mass. ; Oxford: Blackwell Science. xvi, 214 p.
67. Wray, G.A. and C.J. Lowe, *Developmental regulatory genes and echinoderm evolution.* Syst Biol, 2000. **49**(1): p. 28-51.
68. Knight, J.C., *Regulatory polymorphisms underlying complex disease traits.* J Mol Med, 2005. **83**(2): p. 97-109.
69. Rockman, M.V. and G.A. Wray, *Abundant raw material for cis-regulatory evolution in humans.* Mol Biol Evol, 2002. **19**(11): p. 1991-2004.
70. He, G., et al., *Interleukin-10 -1082 promoter polymorphism is associated with schizophrenia in a Han Chinese sib-pair study.* Neurosci Lett, 2006. **394**(1): p. 1-4.
71. Ye, S., et al., *Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression.* J Biol Chem, 1996. **271**(22): p. 13055-60.
72. Beyzade, S., et al., *Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction.* J Am Coll Cardiol, 2003. **41**(12): p. 2130-7.
73. Tournamille, C., et al., *Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals.* Nat Genet, 1995. **10**(2): p. 224-8.
74. Hamblin, M.T. and A. Di Rienzo, *Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus.* Am J Hum Genet, 2000. **66**(5): p. 1669-79.
75. Moraes, M.O., et al., *Interleukin-10 promoter single-nucleotide polymorphisms as markers for disease susceptibility and disease severity in leprosy.* Genes Immun, 2004. **5**(7): p. 592-5.
76. Enoch, M.A., et al., *5-HT2A promoter polymorphism -1438G/A, anorexia nervosa, and obsessive-compulsive disorder.* Lancet, 1998. **351**(9118): p. 1785-6.
77. Caspi, A., et al., *Role of genotype in the cycle of violence in maltreated children.* Science, 2002. **297**(5582): p. 851-4.
78. Enattah, N.S., et al., *Identification of a variant associated with adult-type hypolactasia.* Nat Genet, 2002. **30**(2): p. 233-7.
79. Huby, T., et al., *Functional analysis of the chimpanzee and human apo(a) promoter sequences: identification of sequence variations responsible for elevated transcriptional activity in chimpanzee.* J Biol Chem, 2001. **276**(25): p. 22209-14.
80. Rockman, M.V., et al., *Positive selection on a human-specific transcription factor binding site regulating IL4 expression.* Curr Biol, 2003. **13**(23): p. 2118-23.
81. Brinkman-Van der Linden, E.C., et al., *Human-specific expression of Siglec-6 in the placenta.* Glycobiology, 2007. **17**(9): p. 922-31.

82. Rockman, M.V., et al., *Ancient and recent positive selection transformed opioid cis-regulation in humans.* PLoS Biol, 2005. **3**(12): p. e387.
83. Prabhakar, S., et al., *Accelerated evolution of conserved noncoding sequences in humans.* Science, 2006. **314**(5800): p. 786.
84. Prabhakar, S., et al., *Human-specific gain of function in a developmental enhancer.* Science, 2008. **321**(5894): p. 1346-50.
85. Galas, D.J. and A. Schmitz, *DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.* Nucleic Acids Res, 1978. **5**(9): p. 3157-70.
86. Connaghan-Jones, K.D., A.D. Moody, and D.L. Bain, *Quantitative DNase footprint titration: a tool for analyzing the energetics of protein-DNA interactions.* Nat Protoc, 2008. **3**(5): p. 900-14.
87. Garner, M.M. and A. Revzin, *A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system.* Nucleic Acids Res, 1981. **9**(13): p. 3047-60.
88. Shimomura, O., F.H. Johnson, and Y. Saiga, *Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea.* J Cell Comp Physiol, 1962. **59**: p. 223-39.
89. Gould, S.J. and S. Subramani, *Firefly luciferase as a tool in molecular and cell biology.* Anal Biochem, 1988. **175**(1): p. 5-13.
90. Benita, Y., et al., *Regionalized GC content of template DNA as a predictor of PCR success.* Nucleic Acids Res, 2003. **31**(16): p. e99.
91. Lee, D.H., et al., *Functional characterization of core promoter elements: the downstream core element is recognized by TAF1.* Mol Cell Biol, 2005. **25**(21): p. 9674-86.
92. Thompson, W., et al., *Decoding human regulatory circuits.* Genome Res, 2004. **14**(10A): p. 1967-74.
93. Pavesi, G., et al., *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W199-203.
94. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009.
95. Loots, G.G., et al., *Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.* Science, 2000. **288**(5463): p. 136-40.
96. Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.
97. Yuh, C.H., et al., *Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin.* Dev Biol, 2002. **246**(1): p. 148-61.
98. Harr, R., M. Haggstrom, and P. Gustafsson, *Search algorithm for pattern match analysis of nucleic acid sequences.* Nucleic Acids Res, 1983. **11**(9): p. 2943-57.
99. Knuppel, R., et al., *TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins.* J Comput Biol, 1994. **1**(3): p. 191-8.
100. Wingender, E., *The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.* Brief Bioinform, 2008. **9**(4): p. 326-32.

101. Barbulescu, K., et al., *New androgen response elements in the murine pem promoter mediate selective transactivation.* Mol Endocrinol, 2001. **15**(10): p. 1803-16.
102. Verrijdt, G., A. Haelens, and F. Claessens, *Selective DNA recognition by the androgen receptor as a mechanism for hormone-specific regulation of gene expression.* Mol Genet Metab, 2003. **78**(3): p. 175-85.
103. Horie-Inoue, K., et al., *Identification of novel steroid target genes through the combination of bioinformatics and functional analysis of hormone response elements.* Biochem Biophys Res Commun, 2006. **339**(1): p. 99-106.
104. Massie, C.E. and I.G. Mills, *ChIPping away at gene regulation.* EMBO Rep, 2008. **9**(4): p. 337-43.
105. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nat Methods, 2008. **5**(9): p. 829-34.
106. Solomon, M.J., P.L. Larsen, and A. Varshavsky, *Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene.* Cell, 1988. **53**(6): p. 937-47.
107. Solomon, M.J. and A. Varshavsky, *Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures.* Proc Natl Acad Sci U S A, 1985. **82**(19): p. 6470-4.
108. Buck, M.J. and J.D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.* Genomics, 2004. **83**(3): p. 349-60.
109. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
110. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
111. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009.
112. King, M.C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-16.
113. Stedman, H.H., et al., *Myosin gene mutation correlates with anatomical changes in the human lineage.* Nature, 2004. **428**(6981): p. 415-8.
114. Cooper, S.J., et al., *Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.* Genome Res, 2006. **16**(1): p. 1-10.
115. Kimmel, A.R. and S.L. Berger, *Preparation of cDNA and the generation of cDNA libraries: overview*. Methods Enzymol, 1987. **152**: p. 307-16.
116. Landry, J.R., D.L. Mager, and B.T. Wilhelm, *Complex controls: the role of alternative promoters in mammalian genomes.* Trends Genet, 2003. **19**(11): p. 640-8.
117. Hashimoto, S., et al., *5'-end SAGE for the analysis of transcriptional start sites.* Nat Biotechnol, 2004. **22**(9): p. 1146-9.
118. Shiraki, T., et al., *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15776-81.
119. Hube, F., et al., *Improved PCR method for amplification of GC-rich DNA sequences.* Mol Biotechnol, 2005. **31**(1): p. 81-4.
120. Henke, W., et al., *Betaine improves the PCR amplification of GC-rich DNA sequences.* Nucleic Acids Res, 1997. **25**(19): p. 3957-8.

121. Leicht, M., et al., *Mechanism of cell death of rat cardiac fibroblasts induced by serum depletion.* Mol Cell Biochem, 2003. **251**(1-2): p. 119-26.
122. Li, G., et al., *A novel cellular survival factor--the B2 subunit of vacuolar H+-ATPase inhibits apoptosis.* Cell Death Differ, 2006. **13**(12): p. 2109-17.
123. Cooper, S.J., et al., *Serum response factor binding sites differ in three human cell types.* Genome Res, 2007. **17**(2): p. 136-44.
124. Grimm, S., et al., *Bcl-2 down-regulates the activity of transcription factor NF-kappaB induced upon apoptosis.* J Cell Biol, 1996. **134**(1): p. 13-23.
125. Januchowski, R., et al., *Trichostatin A down-regulate DNA methyltransferase 1 in Jurkat T cells.* Cancer Lett, 2007. **246**(1-2): p. 313-7.
126. Kangaspeska, S., et al., *Transient cyclical methylation of promoter DNA.* Nature, 2008. **452**(7183): p. 112-5.
127. Wilson, M.D., et al., *Species-specific transcription in mice carrying human chromosome 21.* Science, 2008. **322**(5900): p. 434-8.
128. Romanelli, M.G., et al., *Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5' flanking region.* Genomics, 2009. **93**(3): p. 227-34.
129. Rico, D., et al., *Identification of conserved domains in the promoter regions of nitric oxide synthase 2: implications for the species-specific transcription and evolutionary differences.* BMC Genomics, 2007. **8**: p. 271.
130. Gompel, N., et al., *Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila.* Nature, 2005. **433**(7025): p. 481-7.
131. Wagmaister, J.A., et al., *Identification of cis-regulatory elements from the C. elegans Hox gene lin-39 required for embryonic expression and for regulation by the transcription factors LIN-1, LIN-31 and LIN-39.* Dev Biol, 2006. **297**(2): p. 550-65.
132. Wallerman, O., et al., *Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing.* Nucleic Acids Res, 2009.
133. Yokoyama, K.D., U. Ohler, and G.A. Wray, *Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships.* Nucleic Acids Res, 2009. **37**(13): p. e92.
134. Mahony, S., et al., *Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome.* Genome Biol, 2007. **8**(5): p. R84.
135. Buckland, P.R., et al., *Strong bias in the location of functional promoter polymorphisms.* Hum Mutat, 2005. **26**(3): p. 214-23.
136. Diallo, A.B., V. Makarenkov, and M. Blanchette, *Exact and heuristic algorithms for the Indel Maximum Likelihood Problem.* J Comput Biol, 2007. **14**(4): p. 446-61.
137. Blanchette, M., et al., *Aligning multiple genomic sequences with the threaded blockset aligner.* Genome Res, 2004. **14**(4): p. 708-15.
138. Ayyanathan, K., et al., *Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation.* Genes Dev, 2003. **17**(15): p. 1855-69.
139. Abrink, M., et al., *Conserved interaction between distinct Kruppel-associated box domains and the transcriptional intermediary factor 1 beta.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1422-6.

140. Pengue, G. and L. Lania, *Kruppel-associated box-mediated repression of RNA polymerase II promoters is influenced by the arrangement of basal promoter elements.* Proc Natl Acad Sci U S A, 1996. **93**(3): p. 1015-20.
141. Huntley, S., et al., *A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors.* Genome Res, 2006. **16**(5): p. 669-77.
142. Birtle, Z. and C.P. Ponting, *Meisetz and the birth of the KRAB motif.* Bioinformatics, 2006. **22**(23): p. 2841-5.
143. Hamilton, A.T., et al., *Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes.* Genome Res, 2006. **16**(5): p. 584-94.
144. Conrad, B. and S.E. Antonarakis, *Gene duplication: a drive for phenotypic diversity and cause of human disease.* Annu Rev Genomics Hum Genet, 2007. **8**: p. 17-35.
145. Collins, P.J., et al., *The ets-related transcription factor GABP directs bidirectional transcription.* PLoS Genet, 2007. **3**(11): p. e208.
146. Kim, H.R., et al., *Bax inhibitor 1 regulates ER-stress-induced ROS accumulation through the regulation of cytochrome P450 2E1.* J Cell Sci, 2009. **122**(Pt 8): p. 1126-33.
147. Naidoo, N., *ER and aging-Protein folding and the ER stress response.* Ageing Res Rev, 2009. **8**(3): p. 150-9.
148. Nishiyama, A., et al., *Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors.* Cell Stem Cell, 2009. **5**(4): p. 420-33.

10. Appendix

10.1. Abbreviations

5'UTR:	5-prime untranslated region
BRE:	Basic recognition element
BSs:	Binding sites
CGIs:	CpG islands
ChIP:	Chromatin immunoprecipitation
ChIP-chip:	Chromatin immunoprecipitation followed by DNA microarray hybridization
ChIP-seq:	Chromatin immunoprecipitation followed by massively parallel sequencing
CTD:	C-terminal domain
DNA:	Deoxyribonucleic acid
DPE:	Downstream promoter element
DSIF:	DRB sensitivity inducing factor
EMSA:	Electrophoretic mobility shift assay
EST:	Expressed sequence tag
FCS:	Fetal calf serum
GABPa:	GA binding protein
GO:	Gene ontology
GTFs:	General transcription factors
GTM:	General transcription machinery
HEK293:	Human embryonic kidney cell line 293
HSA:	Homo sapiens
HSA21:	Human chromosome 21
INR:	Initiator element
MAC:	Macaca mulatta or Rhesus monkey
NELF:	Negative elongation factor
PCR:	Polymerase chain reaction
PIC:	Preinitiation complex
Pol IIA:	RNA Polymerase II polypeptide A (hypophosphorylated form)
P-TEFb:	Positive transcription elongation factor b
PTR:	Pan troglodytes or Chimpanzee
PWM:	Position weight matrix
RNAPII:	RNA Polymerase II
RNA:	Ribonucleic acid
RNA-seq:	Massively paralell sequencing of cDNA
ROS:	Reactive oxygen species
TBP:	TATA binding protein
TF:	Transcription factor
TFBSs:	Transcrition factor binding site
TSA:	Trichostatin A
TSSs:	Transcription start site
ZF:	Zinc-finger transcription factor

10.2. Curriculum Vitae

Not included in electronic version

Publikationen:

Ralser M.*, Querfurth R.*, Warnatz HJ., Lehrach H., Yaspo ML., Krobitsch S. 2006. An efficient and economic enhancer mix for PCR. Biochem Biophys Res Commun. 1;347(3):747-51. Reproduced by permission of ELSEVIER. *Equal contribution

Warnatz HJ.*, Querfurth R.*, Guerasimova A.*, Cheng X., Vanhecke D., Hufton A., Haas S., Nietfeld W., Vingron M., Janitz M., Lehrach H., Yaspo ML. 2009. Analysis of activities, response patterns and *cis*-regulatory elements of human chromosome 21 gene promoters. In preparation. *Equal contribution

Querfurth R., Warnatz HJ., Sudbrak R., Lehrach H., Yaspo ML. 2009. Discovery of human-specific functional transcription factor binding sites by ChIP-seq and comparative genomics. In preparation.

Polak P., Querfurth R., Arndt P. 2009. The Evolution of Transcription Associated Biases of Mutations across Vertebrates. BMC Genomics. Under revision.

Cheng X., Guerasimova A., Manke T., Rosenstiel P., Haas S., Warnatz HJ., Querfurth R., Nietfeld W., Vanhecke D., Lehrach H., Yaspo ML., Janitz M. Screening of human gene promoter activities using transfected-cell arrays. 2009. Gene. 450(1-2): p. 48-54.

10.3. Acknowledgements

At first, I would like to thank Marie-Laure Yaspo for giving me trust, freedom and support for doing my project. Likewise, I thank Hans Lehrach. He was the first I met when applying for a PhD position, and said “ich sehe da kein Problem” and sent me to meet with Marie-Laure.

Many thanks to Prof. Rupert Mutzel for altruistically agreeing to review this work.

During the last years, I’ve learned many things from Hans-Jörg Warnatz, while I could always count on his help, thank you!

Thanks to Paz Polak, who always had an ear and brain for my stuff. Paz, Alon Magen, Andrew Hufton and Helge Roider spent a lot of time to help me with computational problems, I am very grateful for that.

Besides, there are many people I’d like to thank for helping with materials, suggestions and discussions, or just for being nice creating such a good and friendly atmosphere for everyone to work in. Forgive me if I forgot somebody to mention, it’s already late.

There are: Cornelia Lange, Markus Ralser, Daniela Balzereit, Daniela Köster, Hannes Luz, Ute Nonnhoff, Linda Hallen, Holger Klein, Mario Drungowski, Tobias Nolden, Florian Mertes, Ruben Rosenkranz, Serguei Baranov, Marc Sultan, Illaria Piccini, David Rozado, Silke Wehrmeyer, Clara Schaefer, Christian Linke and Guifré Ruiz.

Also I’d like to thank Hans Zischler for encouraging me to go to the MPIMG, luckily it worked out.

My deepest gratitude to my family Mam, Burki, Rici, Gila and Miri may you live long, happy and healthy!

10.4. Selbständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbst verfasst habe sowie keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Robert Querfurth Berlin, Oktober 2009