4

DISCUSSION

4.1 On the annotation of genes and ORFs on Hsa21

4.1.1 Protein-coding versus non-protein-coding genes

A comprehensive catalog of annotated genes is a fundamental prerequisite for functional analysis of the transcripts and proteins encoded by the genome. Seven years after the initial publication of the DNA sequence and gene catalog of human chromosome 21 (Hattori *et al.* 2000), experimental and computational analysis of the human transcriptome is still an ongoing process. Although the annotation is considered to be almost complete for canonical protein coding genes, which are conserved over significant evolutionary distances, it has become clear that not only protein-coding genes are transcribed, but that a larger fraction of the human genome is transcriptionally active.

A genome-wide study on transcription using tiling arrays interrogating nearly all non-repetitive human genomic DNA identified 10,595 transcribed sequences not detected by any other method (Bertone *et al.* 2004). A recent detailed survey of 1% of the human genome reported by the ENCODE consortium found that “the human genome is pervasively transcribed, such that the majority of its bases are associated with at least one primary transcript, and many transcripts link distal regions to established protein-coding loci” (Birney *et al.* 2007). Regarding chromosome 21, in-depth analysis of its transcriptome by high-density oligonucleotide tiling arrays showed that 49% of the observed transcription was outside of any known annotation (Kampa *et al.* 2004).

Many functional sequences in the genome seem not to be conserved over significant evolutionary distances, and evidence from the ENCODE project suggests that many functional elements show no detectable level of sequence constraint (Pheasant and Mattick 2007). Functional non-protein-coding transcripts (ncRNAs) include structural RNAs (e.g. the long-known transfer RNAs, ribosomal RNAs and small nuclear RNAs) and more recently discovered regulatory RNAs (e.g. miRNAs). Conservative reports suggest that the mammalian genome encodes about 300 to 500 miRNA genes

(Landgraf *et al.* 2007; Saini *et al.* 2007), others estimate larger numbers. Approximately 50% of miRNAs are expressed from ncRNAs, whereas the rest are located mostly in the introns of coding genes (Saini *et al.* 2007). The miRNAs encoded on Hsa21 will also have to be analyzed for their potential involvement in the development of the Down syndrome phenotype. For example, it has been recently proposed that lower blood pressure in trisomy 21 is partially caused by the overexpression of miRNA-155, which is encoded on Hsa21, leading to allele-specific underexpression of one of its target genes, angiotensin II receptor type 1 (Sethupathy *et al.* 2007).

4.1.2 On the quality of the gene annotations used in this study

The work described here focused on the subset of protein-coding genes of Hsa21 and their functional annotation. The basis for gene annotation was the most recent gene catalog available at the beginning of the project (Watanabe *et al.* 2004). The quality of this gene catalog can be estimated by comparison with the most recent version of the human genome (NCBI build 36.2). This NCBI version annotates 352 genes on chromosome 21 in the 'Genes on sequence' data set. Among these genes, there are 55 pseudogenes and 57 uncharacterized loci. The remaining 240 genes, which are approved by the HGNC (HUGO Gene Nomenclature Committee), constitute the known genes also present in the EntrezGene database. Comparison with the Hsa21 gene catalog shows that 232 of all currently known 240 genes were part of the analysis, illustrating the high quality of the gene annotations used during the work described here.

4.1.3 On the quality of the ORF annotations used in this study

Analysis of the transcript sequences showed that a complete open reading frame (ORF) was annotated for 82% of the 284 Hsa21 genes. The remaining transcripts contained partial or no ORFs at all. For each gene, the longest ORF was chosen to include as many protein domains as possible for further analyses. 26 out of 27 highly paralogous keratin-associated genes were excluded because of their strong resemblance, making cloning almost impossible. Manual annotation of the remaining genes resulted in annotated ORFs for 206 different genes from this catalog.

The resulting annotation data for 206 complete ORFs ('Hsa21 full ORFs') was used for all downstream analyses. A comparison with the current data set from NCBI's consensus coding sequence (CCDS) project, which became available in late 2005, shows that 205 CCDS overlap with the 206 manually annotated Hsa21 ORFs.

4.2 ORF cloning as resource for functional genomics

4.2.1 Approaches in functional genomics and proteomics

Functional genomics approaches have proven very successful in the study of DNA and transcript sequences, which have been analyzed in great detail in many healthy and disease states of an ever-growing number of organisms and tissues. On the protein level, the accomplishments are still more modest, mostly because protein analysis is intrinsically more complex due to the higher complexity of protein sequences (and structures) in opposition to nucleic acids. Where DNA and RNA can be conveniently and specifically detected using hybridization (in PCR and on microarrays) or cloning followed by sequencing, specific protein detection is a more challenging task. Three main approaches for specific protein detection have been applied so far.

'Sequencing proteomics' relies on protein separation techniques coupled with mass spectrometry analysis methods, either qualitative (Mann and Wilm 1994; Patterson and Aebersold 1995) or quantitative (Han *et al.* 2001). These techniques promise future proteome-wide protein detection and quantification assays and are already also applied for protein interaction detection using protein complex pull-outs (Kuster *et al.* 2001).

'Affinity proteomics' is based on protein-specific affinity reagents, such as antibodies, antibody fragments or aptamers, allowing specific detection and purification of proteins. These reagents can be used for a wide range of biochemical and functional studies. Most antibodies have been reported from single gene studies, but also more general tools for genome-based proteomics have been proposed (Larsson *et al.* 2000). In a pilot project, specific antibodies were generated for 54 chromosome 21 proteins, out of 168 attempted (Agaton *et al.* 2003). These antibodies have been used for protein profiling of chromosome 21 gene products in human tissues (ProteomeBinders initiative, part of the antibody initiative of the Human Proteome Organization HUPO, see http://www.proteinatlas.org/).

In the work described here, a ‘clone-based proteomics’ approach was applied for functional genomics analysis of Hsa21 protein functions. After recombinatorial cloning of each ORF in a ‘gene-by-gene’ approach, protein fusions or marker sequences were artificially attached by recombinatorial subcloning of the corresponding ORFs in suitable expression vectors. The artificial fusions allowed systematic analysis of subcellular protein localizations (Hu *et al.* 2006) and protein network analysis using the yeast two-hybrid system (Fields and Song 1989; Auerbach *et al.* 2002).

4.2.2 Hsa21 ORF cloning for clone-based proteomics

The most failure-prone step in ORF cloning is the process of creating the master entry clones. Here, the Gateway recombinatorial cloning method was chosen and provided estimated success rates in a range reported before (Marsischky and LaBaer 2004): ORFs smaller than 2 kb were cloned with ~90% efficiency; ORFs of 2-3 kb were cloned with ~50% efficiency, and ORFs of 3-4 kb with ~30% efficiency.

Although the collections of cloned ORFs are growing steadily, researchers still often do not have access to cloned ORFs for their genes of interest. Now, during the work described here, a collection of ORF entry clones was established for 167 different Hsa21 genes. Of this set, 84% of clones were newly generated during the work described here, while 16% were retrieved from other public sources (RZPD – German Resource Center for Genome Research and HIP – Harvard Institute of Proteomics). This Hsa21 ORF clone set is the largest collection freely available to date and exceeds by far the 99 ORF entry clones currently available for Hsa21 from the largest academic initiative cloning the human ORFeome (Rual *et al.* 2004).

4.3 New subcellular localizations for Hsa21 proteins

A set of 96 Hsa21 full ORFs was used for cell array experiments to perform a rapid and cost-effective analysis of the subcellular localization of these proteins. The project was performed in collaboration with the cell array group headed by Dr. Michal Janitz at the Max Planck Institute for Molecular Genetics, Berlin. This collaboration resulted in a joint publication (Hu *et al.* 2006).

4.3.1 Different small fusion tags resulted in similar localizations

For analysis of the subcellular localizations of 89 Hsa21 proteins in HEK293T cells on transfected cell arrays, proteins were overexpressed with N-terminal hexahistidine fusions. This strategy was chosen to minimize the effects of the affinity tag on the native protein localization. An alternative strategy would be the use of fluorescent proteins, such as GFP and its relatives, for detection without secondary staining methods. However, large fusions have been reported previously to result in protein mislocalization, e.g. abolishing localization to peroxisomes (Brosius *et al.* 2002) or to the secretory pathway (Pouli *et al.* 1998; Simpson *et al.* 2000).

For verification purposes, the strategy of N-terminal addition of a hexahistidine affinity tag was complemented with a test set of 17 C-terminally myc-tagged Hsa21 ORFs to test for differences in subcellular localizations. The localizations of these proteins represented all localization categories that were previously classified using an N-terminal hexahistidine tag. 16 out of 17 tested proteins did not differ in the localization of N- and C-terminally tagged proteins. Apart from one exception, the plasma membrane otassium channel KCNJ6, which, when tagged by C-terminal myc tag, was found only in intracellular vesicles, the small N-terminal hexahistidine tag appeared to be a reliable and efficient fusion tag for determination of subcellular localizations.

However, for those clones that failed to be detected using an N-terminal tag, change of fusion terminus may be helpful to reveal their localization behaviors.

4.3.2 Success rate of subcellular localization experiments

A suitable anti-hexahistidine antibody was identified, and nine different organelle counterstaining procedures were established to enable colocalization analysis of the overexpressed Hsa21 proteins with distinct cellular compartments. Then, localization information could be obtained for 52 out of the 89 different Hsa21 proteins (58%) tested for expression on HEK293T transfected cell arrays. Twenty-eight subcellular localizations could be described for the first time.

The success rate (58% obtained localizations) could not be compared to other resembling studies. A similar study by the German cDNA consortium reported the subcellular localizations of 107 human proteins, but does not point out the starting number of ORFs initially tested to obtain these results (Simpson *et al.* 2000). Also, another similar study by the Alliance for Cellular Signaling (AfCS) does not report success rates for localization experiments (Zavzavadjian *et al.* 2007).

4.3.3 Observed localizations and potential effects of epitope tagging

A large fraction of the tested ORF products localized to either the cytosol (31%) or the nucleus (17%) or both compartments (23%). A significant number of proteins were found associated with the secretory pathway (29%). These numbers are very comparable to those from a previous study reporting the localization of 107 previously unknown human proteins (Simpson *et al.* 2000), except that a larger fraction of proteins localized to the cytosol here than reported before (31% versus 18%).

This larger portion of cytosolic proteins may result from sampling effects, resulting in a coincidentially higher number of cytosolic proteins in the tested set of proteins. But it could also be associated with the fact that no localizations were observed here to mitochondria and peroxisomes. Most nuclear-encoded mitochondrial proteins contain aminoterminal signal peptides rich in basic and hydroxylated amino acids; only few are targeted by carboxyterminal signals (Verner and Schatz 1988). Accordingly, it has been shown that N-terminal protein fusions abolish the function of N-terminal mitochondrial targeting signals (Simpson *et al.* 2000). The targeting to peroxisomes, on the other hand, mostly depends on a C-terminal tripeptide motif, with only few proteins bearing an N-terminal nonapeptide sequence (Verner and Schatz 1988). Since the Hsa21 proteins studied here were fused with an aminoterminal epitope tag,

the targeting to mitochondria may have been disturbed, resulting in a larger number of proteins observed in the cytoplasm.

Interestingly, targeting to the ER and the secretory pathway does not seem to be disturbed by N-terminal epitope fusions, although the signal peptide consists of a aminoterminal basic region, followed by a stretch of at least 7-8 apolar, largely hydrophobic residues (Verner and Schatz 1988). While a previous study found that N-terminal GFP fusions of proteins of the ER fail to localize correctly (Simpson *et al.* 2000), here 10% of the analyzed proteins were found in the ER and 29% in the secretory pathway. These observations clearly show that short N-terminal fusions, in contrast to large fusions like GFP (Huh *et al.* 2003; Wiemann *et al.* 2004), do not abolish correct localization to the endoplasmic reticulum and the secretory pathway.

4.3.4 Known and new localizations in different compartments

An overview of new subcellular localizations identified for Hsa21 proteins is presented in Figure 4-1.

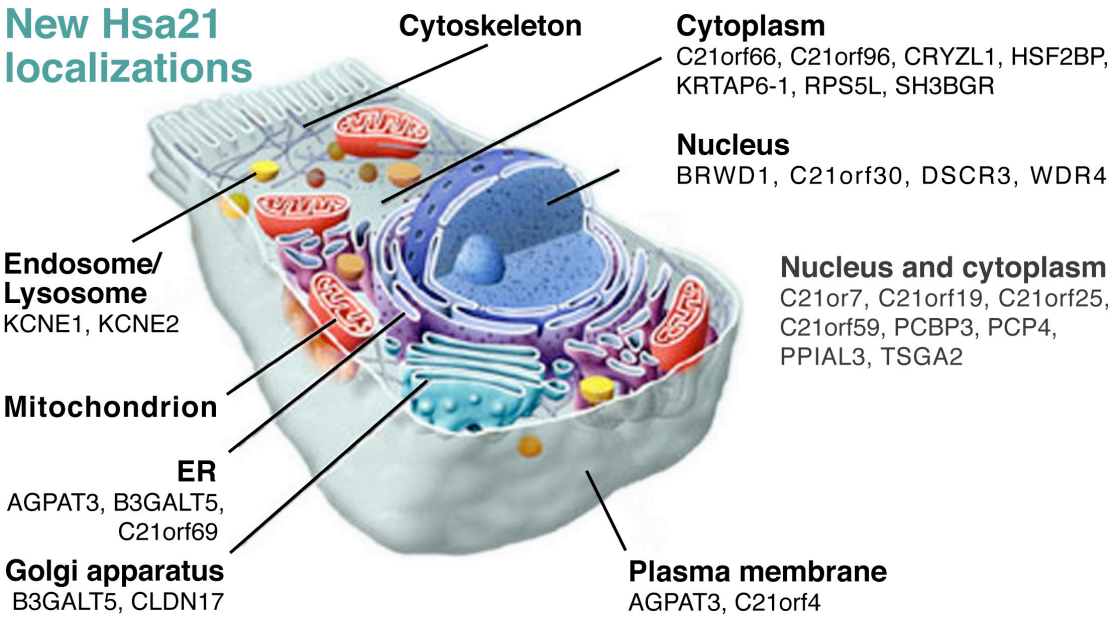

Figure 4-1. New subcellular localizations identified for Hsa21 proteins during the work presented here. Figure modified from 'The Internet Encyclopedia of Science', entry for 'animal cell' [http://www.daviddarling.info/encyclopedia/A/animal_cell.html].

Proteins exclusively localized to the nucleus

Seven of the tested Hsa21 proteins were were found to localize exclusively to the nucleus. Among these, there are three known nuclear proteins, namely the v-ets oncogene homolog 2 (ETS2), the high-mobility group nucleosome binding domain 1 (HMGN1) and the ribosomal RNA processing protein 1 (RRP1, alias novel nuclear protein 1, Nnp1).

Distinct localization patterns could be observed inside this compartment. For example, RRP1 localized exclusively to the nucleoli, which is in accordance with its role in the synthesis of 28S rRNA (Savino *et al.* 1999). In contrast, the non-histone chromosomal protein HMGN1 was found mainly in the nucleoplasm outside of the nucleoli, correlating with its function in unfolding higher-order chromatin structure to facilitate transcriptional activation of mammalian genes (Ding *et al.* 1997).

For the WD repeat-containing putative signal transduction protein BRWD1 and the putative tRNA modification protein WDR4 (Alexandrov *et al.* 2002), nuclear localization was shown here for the first time. Also, nuclear localization was discovered here for the uncharacterized Down syndrome critical region gene 3 (DSCR3) and for the uncharacterized protein C21orf30. Their nuclear localizations render these four genes interesting candidates as regulators of gene expression.

Proteins localized to both nucleus and cytoplasm

This dual localization category usually delivers the least information about protein function (Simpson *et al.* 2000). Fourteen Hsa21 proteins were found to localize to both the nucleus and the cytoplasm. The nucleus/cytoplasm distribution ratio for particular proteins, however, often varied from cell to cell. In some cells, proteins with dual localization could be found either only in the nucleus, or only in the cytoplasm, suggesting a continuous translocation activity (see chapter 4.3.5, ‘Protein translocation‘).

There are six Hsa21 proteins for which a dual localization in nucleus and cytoplasm has been shown before:

(1) The minichromosome maintenance complex component 3 associated protein (MCM3AP) was found in both nucleus and cytoplasm, which is an important feature for integrity of the cellular replication cycle (Takei *et al.* 2002).

(2-5) For four Hsa21 transcriptional regulators, a cell-cycle dependent translocation between nucleus and cytoplasm was observed. These transcriptional regulators are BTB and CNC homology 1 (BACH1), chromatin assembly factor 1B (CHAF1B), DNA (cytosine-5-)-methyltransferase 3-like (DNMT3L) and PBX/knotted 1 homeobox 1 (PKNOX1). All of these dual localizations have been reported before (Marheineke and Krude 1998; Berthelsen *et al.* 1999; Suzuki *et al.* 2004; Nimura *et al.* 2006).

(6) Another protein where a nuclear localization has been assumed before is the ribosomal RNA processing 1 homolog B (RRP1B, alias KIAA0179).

Apart from the six proteins with known dual localization, there are eight proteins showing nuclear and cytoplasmic localization for the first time:

(1-4) The uncharacterized proteins C21orf7, C21orf19, C21orf25 and C21orf59 were found localized to nucleus and cytoplasm.

(5) The poly(rC)-binding protein 3 (PCBP3) was found to be remarkably versatile, ranging from nucleus-only over both nucleus and cytoplasm to cytoplasm-only, in contrast to a previous study (Chkheidze and Liebhaber 2003), where PCBP3 was found only in the cytosplasm. Its complex localization pattern is presumably associated with the functions of PCBP3 in a broad spectrum of post-transcriptional events, including regulation of mRNA stability and translation. In line with the nuclear localization, it was recently reported that PCBP3 can also function as a transcriptional repressor of the mu opioid receptor, dependent on a poly(C) sequence in the promoter of this gene (Choi *et al.* 2007).

(6) Purkinje cell protein 4 (PCP4), a neuronally expressed polypeptide with the ability to alter the calcium-binding dynamics of calmodulin (Putkey *et al.* 2003), was found in both nucleus and cytoplasm.

(7) Peptidylprolyl isomerase A-like protein 3 (PPIAL3), a putative molecular chaperone involved in correct folding of proteins, was observed in nucleus and cytoplasm. Interestingly, a nuclear function for the closely related PPIA protein was recently discovered, where it participates in the nuclear translocation of apoptosis-inducing factor in neurons after cerebral hypoxia-ischemia (Zhu *et al.* 2007), and

(8) Testis-specific gene A2 (TSGA2, recently renamed to radial spoke head 1 homolog, RSPH1), has been reported before to be localized to the radial spokes of the axonemes of both sperm and cilia (Shetty *et al.* 2007), but is also expressed in

brain, thyroid, trachea and lung in addition to testis, and was seen in nucleus and cytoplasm here.

Proteins localized to the cytoplasm

Sixteen Hsa21 proteins were found localized exclusively to the cytoplasm. In this, proteins were either distributed evenly throughout the entire cytosol (12/16 proteins), or formed punctuate patterns (4/16 proteins).

An evenly distributed cytoplasmic localization was known before for six proteins:

(1-3) Three metabolic enzymes have been known before to reside throughout the cytoplasm, namely cystathionine-beta-synthase (CBS), holocarboxylase synthetase (HLCS) and pyridoxal kinase (PDXK).

(4-5) An even distribution was also known for the chaperonin containing TCP1 subunit 8 (CCT8), a subunit of a cytosolic heterooligomeric molecular chaperone assisting in the folding of proteins in eukaryotic cytosol (Kubota *et al.* 1995). Moreover, cytoplasmic localization was known previously for ubiquitin associated and SH3 domain containing protein A (UBASH3A), where a similar localization pattern has been reported in a study where overexpression of UBASH3A inhibited clathrin-dependent endocytosis at the plasma membrane (Bertelsen *et al.* 2007).

(6) An interesting case is ubiquitin-conjugating enzyme E2G 2 (UBE2G2), which has been found previously in the cytosol, but also in a diffuse perinuclear expression pattern that is indicative of an ER membrane-associated localization. As described in chapter 3.5.1 (‘Retrieval of interaction data for Hsa21 orthologous proteins’), this dual localization of UBE2G2 in the cytoplasm and around the nucleus was also observed during the work described here in COS-1 cells transfected with a construct encoding a hemagglutinin-tagged protein.

New cytoplasmic localization properties with even distribution were found for six proteins without previous annotation:

(1) Crystalline zeta-like 1 protein (CRYZL1) was found in the cytoplasm, supporting a previous finding that zeta-crystallins, such as CRYZL1, specifically bind to adenine-uracil rich elements in RNA and may have a role as trans-acting factors in the turnover of certain mRNAs in the cytoplasm (Fernandez *et al.* 2007).

(2) Heat shock transcription factor 2 binding protein (HSF2BP), a cofactor that can bind HSF2, potentially resulting in activation and nuclear translocation of this transcription factor (Yoshima *et al.* 1998), was found in the cytoplasm.

(3-6) Other cytoplasmic localizations observed for the first time were those for the uncharacterized ribosomal protein S5-like (RPS5L), keratin associated protein 6-1 (KRTAP6-1), C21orf66 and SH3 domain binding glutamic acid-rich protein (SH3BGR), the founding member of a novel class of thioredoxin fold proteins (Yin *et al.* 2005).

Four proteins appeared in punctuate patterns in the cytosol. Among these, the enzyme phosphofructokinase liver type (PFKL) was reported before to be organized in complex multimeric structures together with a large number of effectors (Barcena *et al.* 2007). Also found organized in punctuate patterns were myxovirus (influenza virus) resistance protein 1 (MX1), phosphodiesterase 9A (PDE9A) and the uncharacterized protein C21orf96. It is known that some PDE9A splice variants are targeted to membrane ruffles and cellular vesicles, while other variants appear to be cytoplasmic (Rentero and Puigdomenech 2006). MX1, a large GTPase with ability to form high molecular weight oligomers, is known to self-assemble into highly ordered dynamin-like complexes (Kochs *et al.* 2002). These rings can tubulate lipids *in vitro* and associate with the smooth endoplasmic reticulum (Accola *et al.* 2002).

For the unknown protein C21orf96, a punctuate cytosolic pattern was shown for the first time.

Proteins localized to the ER and the Golgi apparatus

Eleven Hsa21 proteins were found localized to the first stations of the secretory pathway, namely the endoplasmic reticulum (ER) and the Golgi apparatus.

Three of these proteins were resident only in the ER and/or Golgi, showing that they are actively retained to escape the default pathway. The three proteins fulfilling the criterion of active retention are TMPRSS3, B3GALT5 and C21orf69. For the transmembrane serine protease TMPRSS3, localization to the ER has been shown before (Guipponi *et al.* 2002), whereas the ER/Golgi localization of the galactosyltransferase B3GALT5 had only be inferred from its enzymatic function, and the unknown protein C21orf69 was found in the ER for the first time here.

Four Hsa21 proteins were found both in the ER and the plasma membrane. Three of them were known to reside here, namely the claudins CLDN8 and CLDN14, which are tetraspan transmembrane proteins of tight junctions (Krause *et al.* 2007), and the interferon gamma receptor IFNGR2. The ER/PM localization was shown for the first time here for the acyltransferase AGPAT3, which is supposed to contribute to glycerolipid synthesis and to play an important role in regulating lipid metabolism (Lu *et al.* 2005).

The last four proteins in this group were found in both Golgi apparatus and plasma membrane. Interestingly, another claudin, namely CLDN17, showed this localization property, differing from CLDN8 and CLDN14, as described above. However, the potassium channels KCNJ6 and KCNJ15 as well as the ATP-binding cassette (ABC) transporter ABCG1, which is involved in cellular lipid homeostasis, were found in the plasma membrane, as expected (Reimann and Ashcroft 1999; Kennedy *et al.* 2001).

Proteins localized to the plasma membrane and lysosomes

Four Hsa21 proteins were found localizing only to the plasma membrane and/or lysosomes.

While the potassium voltage-gated channels KCNE1 and KCNE2 correctly localized to the plasma membrane, where they are responsible for the delayed-rectifier potassium current in cardiac myocytes, they were also found in the lysosomal compartment. Indeed, a previous study showed that KCNE1 can be sensitive to disturbances in the lysosomal pathway (Knipper *et al.* 2006).

In contrast, the two proteins CXADR and C21orf4 were found exclusively localized to the plasma membrane. For the coxsackie virus and adenovirus receptor (CXADR), a membrane protein that may affect cell migration through its interaction with microtubules (Fok *et al.* 2007), this localization was known and expected. The unknown protein C21orf4, on the other hand, has not yet been functionally characterized. Further analysis of this protein is needed, since interestingly, C21orf4 has been reported to be one of six genes that can be used to differentiate between benign and malignant thyroid tumors with high sensitivity and specificity (Rosen *et al.* 2005).

4.3.5 Protein translocation

Many proteins change their cellular localization depending on their state of activation. Well-studied examples for this kind of translocation are steroid receptors, which move to the nucleus in hormone-stimulated cells (Hager *et al.* 2000), and NFκB, a transcriptional regulator that translocates from the cytosol to the nucleus in response to cellular stress (Baeuerle and Henkel 1994). Other kinds of protein translocation include changes of localization from the cytosol to the plasma membrane and to other internal membranes, trafficking along the secretory pathway and movement of membrane proteins in and out of the plasma membrane.

Here, six proteins were found that show subcellular translocation. A cell-cycle dependent translocation between nucleus and cytoplasm was observed for four Hsa21 transcriptional regulators. These transcriptional regulators are BTB and CNC homology 1 (BACH1), chromatin assembly factor 1B (CHAF1B), DNA (cytosine-5-)-methyltransferase 3-like (DNMT3L) and PBX/knotted 1 homeobox 1 (PKNOX1). All of these dual localizations have been reported before (see chapter 4.3.4, 'Proteins localized to both nucleus and cytoplasm').

Also, the minichromosome maintenance complex component 3-associated protein (MCM3AP) was found in both nucleus and cytoplasm. This dual localization is in line with the role of MCM3AP in the acetylation and translocation of minichromosome maintenance complex component 3 (MCM3) from the cytosol into the nucleus. MCM3 is one of the MCM proteins essential for the initiation of DNA replication. Hence, MCM3AP is a potent natural inhibitor of the initiation of DNA replication whose action is mediated by interaction with MCM3 (Takei *et al.* 2002), and dual localization of MCM3AP is an important feature for integrity of the cellular replication cycle.

Another example is the distribution of the poly(rC)-binding protein 3 (PCBP3), which was found to be remarkably versatile, ranging from nucleus-only over both nucleus and cytoplasm to cytoplasm-only, in contrast to a previous study (Chkheidze and Liebhaber 2003), where PCBP3 was found only in the cytoplasm. The observed complex localization pattern is presumably associated with the functions of PCBP3 in a broad spectrum of posttranscriptional events, including regulation of mRNA stability and translation. In line with the nuclear localization, it was recently reported that PCBP3 can also function as a transcriptional repressor of the mu opioid receptor, dependent on a poly(C) sequence in the promoter of this gene (Choi *et al.* 2007).

4.3.6 Comparison with computational predictions

An expanding number of computational algorithms are becoming available for the prediction of subcellular localization of proteins. To assess the reliability of these predictions in regard to experimentally determined localizations, the recent versions of two established algorithms (WoLF PSORT and ProtComp 4) were used to make localization predictions for the 52 proteins with observed localizations in HEK293T cells.

The results show that the localization of cytoplasmic proteins was better predicted by WoLF PSORT than by ProtComp 4 (81% vs. 69% correct), while for plasma membrane proteins, ProtComp 4 scored 100% correct, but WoLF PSORT reached only 83%. Cellular localization restricted to the nucleus was correctly predicted by both algorithms for 71% of nuclear proteins. For all correct predictions, the accordance between both algorithms ranged from 83% for proteins in nucleus and cytoplasm up to 100% for proteins restricted to the nucleus.

More problematic cases are proteins in the secretory pathway (ER and Golgi), where only ProtComp 4 succeeded in predicting localization in the endoplasmic reticulum (9% correct), and lysosomal localization, where no correct prediction was obtained. This result shows that current prediction algorithms still cannot correctly foresee all possibilities arising from the intracellular trafficking machinery.

Also, neither program succeeded in correct prediction of the localization for two nuclear proteins (WDR4, DSCR3), three cytoplasmic proteins (PDE9A, UBASH3A, KRTAP6-1) and two proteins residing in the ER (C21orf69, TMPRSS3). And for nine other proteins, at least one algorithm failed, making it difficult to evaluate the specificity of the predictions.

While computational predictions of subcellular localizations can greatly help in initial characterization of unknown protein sequences, no computational method has a sufficient level of sensitivity across the entire protein set that would enable reliable annotation of unknown proteins (Sprenger *et al.* 2006). Nuclear and membrane proteins were predicted here with relatively high sensitivity, whereas proteins localized to the secretory pathway were the most difficult to predict. The complexity of cellular protein sorting still limits the application of computational predictions in determination of subcellular localization. Experimental validation by localization analysis in living cells remains an important step in functional annotation of proteins.

4.4 New protein interactions for Hsa21 proteins

4.4.1 On the identification of new PPIs by mating array-based Y2H screening

To contribute to the identification of new PPIs involving Hsa21 proteins, a set of 62 Hsa21 proteins available as full ORF clones in Gateway entry vectors and devoid of transmembrane domains was cloned as baits to perform a Y2H screening. The Y2H mating array used for this work contained 5,632 preys and has already been used to identify 3,186 mostly novel interactions among 1,705 proteins, building a large human protein-protein interaction network (Stelzl *et al.* 2005).

A total of 53 non-autoactivating baits were obtained which were subsequently screened in a Y2H mating array procedure, which has the advantages of being both efficient and reproducible (Stelzl and Wanker 2006). In this, arrays of defined bait and prey proteins are systematically tested for interaction against each other to ensure that each interaction has the same probability of being identified. As result, 13 out of 53 Hsa21 baits (25%) identified 29 PPIs, which were all novel, whereas the remaining 40 baits failed to identify interactions in these experimental conditions.

Among the interactions that can be recovered from a Y2H screen, there are those of biological relevance, and the false-positives. The latter fall into two main categories. The first category represents artifacts of the Y2H assay itself: transcriptional activity can occur independently of any PPI, when a protein fused to the DNA-binding or activation domain can activate transcription on its own. Also, plasmid rearrangements or copy number changes can generate such auto-activators, as well as alterations at the reporter genes can result in constitutive expression. These false positives may be highly reproducible, but are generally eliminated when small-scale experiments follow the large-scale Y2H screen (Fields 2005). The second category of false positives consists of those proteins that do indeed bind to the bait protein in the context of the Y2H assay, but not in the normal *in vivo* context. For example, such proteins might be members of a family with similar protein domains, but not the specific member that recognizes the bait protein in its normal cellular environment. This kind of false positives can be eliminated by using cell biological assays for further verification, such as immunofluorescence colocalization.

Seemingly modest (25% of baits with hits), the result from the Hsa21 Y2H screen is nevertheless comparable with that of previous mating array-based Y2H screens. In the large CCSB Y2H screen (Rual *et al.* 2005), ~7,200 ORFs were screened as baits against the identical collection of preys, resulting in interactions for 1,549 proteins (21% of proteins tested). Comparably, in the large MDC Y2H screen (Stelzl *et al.* 2005), 4,456 bait ORFs were screened against 5,632 preys, resulting in interaction information for 1,064 baits (24% of baits tested) and 1,065 preys (19% of preys tested).

Several factors should be taken into account that might explain these results. First, the collections of preys in the mating arrays represent only 20-25% of the estimated number of human genes, and an even much smaller fraction of the possible isoforms in the human proteome. Second, the interactions are tested in inherently constrained defined conditions that might be very different from the optimal microenvironment in the biological context (cell type, cell compartment, developmental time point, etc.). Third, macromolecule assemblies can be constitutive or dynamic and are dependent on the physiological context, e.g. interactions could be modulated by post-translational protein modifications and ligand-induced structural modifications, which cannot be reproduced in the Y2H procedure. Future experiments designed for screening extended collections of preys and baits should enable future expansion of the Hsa21 PPI data set described here.

Also, binary protein-protein interactions can only provide a partial and static snap-shot of a cellular process at a given time and in a particular context. Whereas each Y2H screen only identifies a subset of the interactome, it is noteworthy that those interactions are mostly novel. This reflects the fact that, as reported by others (von Mering *et al.* 2002; Stelzl and Wanker 2006; Chaurasia *et al.* 2007), we are still far from reaching saturation in terms of human PPIs. We probably captured only a minute fraction of the relevant biological networks, and several iterations of screens using complementary techniques will be necessary to approach a more comprehensive view of the human interactome (Stelzl and Wanker 2006). For instance, high-throughput mass spectrometry-based protein interaction data are becoming available at least for the budding yeast (Gavin *et al.* 2006; Krogan *et al.* 2006) and now also for human (Ewing *et al.* 2007).

4.4.2 On the verification of protein-protein interactions

The validation of the PPIs identified by Y2H assays represents a non-trivial issue. In the work described here, 29 newly identified PPIs from the mating-array Y2H screen were tested, as well as another 27 interactions of Hsa21 proteins identified in a previous Y2H screen (Stelzl *et al.* 2005). An independent Y2H co-transformation setup was used to exclude experimental artifacts (Lehner *et al.* 2004). From a total of 41 PPIs for which sequence-verified constructs could be obtained, 23 interactions (56%) could be confirmed. This shows that the specific experimental conditions, such as diploid versus haploid yeast cells and mating versus co-transformation assay, strongly influence the experimental outcome. From the validated interactions, 16 PPIs were chosen for confirmation by cellular colocalization and pull-down assays. For ten of these protein pairs, it was not possible to obtain soluble protein expression for one or both of the interaction partners either in COS-1 or in bacterial cells, pointing out to a common problem that hinders PPI verifications on a large scale. Of the remaining six PPI pairs, five interactions could be confirmed (83%).

Whether the PPIs that either failed to validate or have not yet been tested represent faithful interactions remains an open question. On the one hand, the capture of false-positive interactions cannot be ruled out, as the interactions are tested in the constrained environment of a yeast cell nucleus. On the other hand, Y2H screening also allows the capture of interactions that are difficult to reproduce in other *in vitro* systems, such as proteins forming highly dynamic complexes in a particular microenvironment. More experiments will be needed to assess the biological relevance of the novel interactions reported.

An interesting study challenges the dogma that only specific protein-protein interactions can be biologically functional (Shi *et al.* 2006). It was reported that across-species interactomes have significant differences that reflect the strengths of the protein-protein interactions. The results identify a global evolutionary shift: more evolved organisms have weaker binary protein-protein binding, a result which is consistent with the evolution of increased protein unfoldedness. Thus, the study of the human protein-protein interactome may also be hindered by a relatively large number of interactions that are not specific enough to be reliably detected by current experimental methods.

The power of the interolog approach was recognized in the identification of new human PPIs, and significant efforts have been involved in establishing interolog

maps (Lehner and Fraser 2004) and databases, such as OPHID (Brown and Jurisica 2005) and HomoMINT (Persico *et al.* 2005). Two out of five interologs tested here on human proteins were confirmed, and it is difficult to gauge whether the remaining interactions are either not transferable to human PPIs, or whether those take place in a different context not reproduced in the conditions tested here. Although raw Y2H data cannot be taken as definitive, the study reported here has produced an additional informative coverage of PPIs, allowing the assignment of novel functions for several Hsa21 proteins.

4.5 On current protein-protein interaction data sets

All available PPI information involving Hsa21 proteins was retrieved from the two large literature-curated databases BIND and HPRD, from the MDC and CCSB large Y2H screens and from systematic literature searches using PubMed. The resulting compilation represents the most comprehensive Hsa21-specific PPI collection to date, with 684 pair-wise interactions linking 108 Hsa21 proteins to 547 interactors.

4.5.1 Retrieval of PPI data from databases and large Y2H screens

Despite the presence of a number of common Hsa21 proteins in the different data sources, there is little overlap in PPIs, as reported for other PPI data sets (Futschik *et al.* 2007). In BIND and HPRD combined, there are 431 PPIs connecting 78 Hsa21 proteins with 375 interactors. Computational comparison showed that each of these PPI databases display their own unique features with a large variation in the type and depth of annotations (Mathivanan *et al.* 2006), reflecting their different curation methodologies.

Combining information from different PPI databases is a non-trivial task requiring manual curation steps. Major problems reside in the use of different nomenclature, ambiguous names, multiple aliases and the difficulty to map proteins to each other. There is virtually no overlap between BIND and HPRD in regard to interactions of Hsa21 proteins; e.g. only four PPIs from CCSB in HPRD or BIND, and none from the MDC Y2H set. General awareness of the crucial need for a federated database

integrating various PPI data sources has prompted recent efforts in this direction (Chaurasia *et al.* 2007).

4.5.2 Retrieval of PPI data from literature records

Because of the relatively small number of proteins encoded on Hsa21, it was possible to perform manual searches querying all protein names and their aliases in PubMed. This resource yielded a set of 303 PPIs involving 71 Hsa21 proteins. The fraction of literature-extracted interactions also reported in either HPRD or BIND was 44%. When excluding overlaps with interologs and the CCSB Y2H set, the systematic literature search permitted the retrieval of 168 PPIs not reported anywhere else, showing that a significant fraction of interaction data can be solely found in individual literature records. However, gene-by-gene searches are cumbersome and not yet manageable on a large-scale. Emerging systems for automated text mining (Hirschman *et al.* 2002; Rzhetsky *et al.* 2004; Ramani *et al.* 2005) would represent an attractive and powerful alternative to manual curation of the literature. However, these systems are still in their early stages of development, and are currently hampered by the lack of complete common dictionaries for protein names as well as numerous obstacles in defining appropriate vocabulary terms and semantics. Comparison with the Hsa21 data showed that automatically extracted PPIs suffer of a high rate of false positives (data not shown).

There is crucial need to integrate interaction data in common repositories, and to promote standardization of protein-related data, as initiated by the Proteomics Standards Initiative (Hermjakob *et al.* 2004). The challenge is to collapse all PPI data sets, based on a common nomenclature, as initiated with the UniHI database (Chaurasia *et al.* 2007). The next step will be to establish a single system for PPI data submission and annotation, allowing also batch queries of PPI information.

4.5.3 Characteristics of proteins without PPI information

Out of 206 Hsa21 full ORFs, 108 have been mapped in the interaction networks, whereas no PPI could be associated for 98 of those. As expected, there is a strong bias towards unknown proteins in the set of proteins lacking interaction information. Only 27% of the proteins without interactions (26/98) have a GO annotation in 'biological process', in contrast to 81% of the proteins with interactions (88/108). And there are 51% uncharacterized ORFs among the proteins lacking interaction

information (50/98), but only 10% of uncharacterized ORFs among the proteins with known interactions (11/108). Regarding disease genes, 18/19 Hsa21 proteins with known association to a disease also have known interactions. The only exception is claudin 14, a transmembrane protein associated with autosomal recessive deafness type 29.

4.6 On the analysis of protein interaction networks

4.6.1 Higher-order protein networks

Although Y2H assays identify only binary PPIs, most of these are involved in larger protein complexes. Ideally, the next steps will be to study all isoforms, and put all binary interactions into a cellular context in space and time, in order to understand the topological and dynamic properties of this network. The molecular cartography capturing all macromolecular complexes is still in its early infancy. In our current state of knowledge, we present PPI networks as a static view of all identified nodes and edges. We need to understand in which biological context these interactions occur and to decipher the highly dynamic partners of multifunctional complexes.

In order to approach these complexes, a so-called interactome walking (Cusick *et al.* 2005) was performed here, using direct interaction partners of Hsa21 proteins as bait in an additional electronic screen of PPI databases, hence identifying indirect interactors within putative macromolecular complexes. This strategy yielded a network composed of 5,660 proteins linked by 14,982 connections, which was deposited in the '21ppi' database (http://chr21.molgen.mpg.de/21ppi). However, there is inherently more noise in results from interactome walking than in the original PPI data set. Although data should be considered as only putative at this stage, results can be insightful in identifying regulatory networks in which Hsa21 proteins are involved, allowing the establishment of testable hypotheses.

On a functional level, it will be interesting to integrate data on protein complexes and networks with information pertaining to transcriptome analysis in trisomy 21. For instance, analysis of gene expression variation in a mouse model for Down syndrome allowed to prioritize candidate genes that are tightly regulated and more likely to behave as dosage-sensitive genes (Sultan *et al.* 2007). Interestingly, the top scoring gene in this study, APP, encodes a hub protein for which at least 60 interactions have been found.

4.6.2 Mapping protein interactions to cellular processes

The next step in the analysis of molecular networks is to assign functional roles to protein complexes within cellular processes. The 'guilt by association' principle hooks newly identified interactors to given processes via protein partners exhibiting previously identified functions. However, assigning functions to proteins within biological pathways is a challenging task, since many proteins participate in multiple complexes, each of them exerting a particular function in a specific time and space.

Networks of transcriptional regulators

Thirteen Hsa21 transcription factors were found linked to 119 direct interactors, 2/3 of which are transcription factors themselves. Eight Hsa21 transcriptional regulators were found interconnected within a large network via ten other transcription factors. On a whole-network scale, the 13 Hsa21 transcription factors are linked to 437 proteins with transcriptional activity via 899 interactions. This information will be particularly relevant for studying potential perturbations of gene regulation networks in the context of aneuploidy.

Recently, a study has been carried out at the Max Planck Institute for Molecular Genetics in Berlin, where all transcription factors encoded on human chromosome 21 were knocked down in HEK293T cells one-by-one using RNA interference (RNAi) experiments. Subsequently, changes in the transcriptome of the cells were measured through mRNA hybridization on DNA microarrays. Thus, direct and indirect target genes of the transcription factors could be elucidated. Although these results are not yet published, preliminary analysis shows that Hsa21 transcriptional regulators that are interconnected via other transcription factors can indeed also share target genes. For example, GABPA and BACH1, connected via their PPIs with ATF1 and MAF, regulate the expression of a common set of target genes, as identified in the knock-down experiments (I. Piccini, manuscript in preparation).

Multifunctional proteins and hub proteins

Analysis of Hsa21 protein interactions revealed several multifunctional proteins, such as the moonlighting proteins MCM3AP and C21orf127, as well as hub proteins, e.g. APP, NRIP1, ITGB2, RUNX1 and S100B, some of which have pivotal roles in multiple processes, thus representing attractive candidates for the complex DS phenotype. MCM3AP was placed at an interesting functional position including

initiation of DNA replication, cell proliferation, axonal transport and axonal outgrowth. C21orf127 was found to be involved in protein biosynthesis and to be connected to a number of other processes via conserved proteins in a potentially ancient interaction network related to metabolism of proteins and transcripts. And the second-“busiest” hub on Hsa21 is NRIP1 with 41 interactions, a protein for which a potential new role in transcriptional regulation via both acetylation and deacetylation of histones could be described.

Far from being complete, an extensive Hsa21 PPI data set was generated here. This data set offers a framework that can be directly exploited for predicting gene function and for studying Down syndrome relevant perturbations. Additional PPIs pertaining to Hsa21 will be identified which will enrich the present PPI database. Integration with other qualitative and quantitative data types, such as protein localizations, pattern of co-expression (synexpression groups), RNAi information and genetic data are necessary to construct models of protein networks that reflect physiological situations.

4.7 Interaction networks and signaling pathways

The results of the Hsa21 pathway connection analysis can expand existing hypotheses on known pathway perturbations in Down syndrome. Moreover, new hypotheses can be derived from this data, predicting an effect on associated pathways for perturbation of the expression and/or activity of Hsa21 proteins, e.g. in cell culture or mouse model systems.

After the identification of pathway connections that seem prominent candidates for association with pathological changes observed in Down syndrome patients, connections were seeked between these pathways and the developmental delays in cognitive and motor skills as well as medical problems, including heart abnormalities, risk of hearing and vision defects, infection, leukaemia, thyroid disorders, and the development of Alzheimer-type dementia later in life.

Currently, several pathways or processes, including MAP kinase signaling, calcineurin signaling, mitochondrial function, RNA processing, and protein modification, have already been shown to be impacted by multiple chromosome 21 proteins, making them strong candidates for relevance to the DS phenotype

(Gardiner *et al.* 2004). These data suggest that studying pathways, not single genes, will be necessary for making phenotypic correlations.

As shown in the study described here, the Hsa21 protein-protein interactions have a significant potential to influence twelve important regulatory pathways, as shown by analysis of representation of Hsa21 proteins and direct interactors in these pathways. Some of these candidate pathways, such as the interferon pathway and the p38 pathway, have been previously implicated in the pathogenesis of Down syndrome or models for DS. Other pathways appear in this context for the first time and thus present interesting prospects for future analyses, e.g. growth factor signaling and the p73 pathway. Hypotheses generated on the basis of this analysis can now be tested by pathway analysis *in vivo*, for example by overexpression or RNAi-mediated knock-down of Hsa21 proteins and/or their direct interactors, followed by analysis of pathway activity in the model system used in the experimental set-up. Of course, this list of potentially affected pathways is not comprehensive, mainly due to the limitations in the number of known protein-protein interactions. Recently, for example, a defective cerebellar response to mitogenic Hedgehog signaling has been described in the Ts65Dn mouse model for Down syndrome (Roper *et al.* 2006), adding another pathway to the candidate list for the multitude of symptoms associated with this congenital chromosome aberration defect.

4.7.1 Previously known pathway connections of Hsa21 proteins

Hsa21 proteins involved in interferon signaling

The interferon (IFN) pathway is perhaps the best-studied pathway for trisomy 21. The IFN-alpha/beta/omega receptor chains IFNAR1 and IFNAR2 as well as the IFN-gamma receptor beta chain IFNGR2 are located on Hsa21. Other Hsa21 proteins connected to this pathway are IL10RB, ITGB2, PPIA, UBE2G2 and UBASH3A. Altogether, eight Hsa21 proteins could be linked to interferon signaling via direct interactions, including five embrane receptors.

As early as in 1980, a hypothesis relating interferon action and the chromosome 21 trisomy genotype and phenotype was presented (Maroun 1980). Several years later, it could be shown that anti-interferon immunoglobulins can improve the phenotype of mice with trisomy 16, one model for human trisomy 21 (Maroun 1995). Also, partial

IFN-alpha/beta and IFN-gamma receptor knockout trisomy 16 mouse fetuses show improved growth and cultured neuron viability (Maroun *et al.* 2000).

Aberrant expression of interferon-related protein RACK1 in fetal Down syndrome brain has been shown by 2D-gel and MALDI analysis (Peyrl *et al.* 2002). Moreover, IFN-gamma plays a crucial role in the induction of Abeta 40 and Abeta 42 production from the Alzheimer precursor protein APP, another protein encoded on Hsa21 (Blasko *et al.* 1999). A more recently reported cross-talk to the Notch pathway in Down syndrome by activation of Notch and Hes1 through APP signaling might affect brain development, since the Notch pathway plays a pivotal role in neuron-glia differentiation (Fischer *et al.* 2005).

In the DS thymus, overexpression of IFN-gamma and TNF-alpha has been demonstrated and been linked to abnormal thymic anatomy (Murphy *et al.* 1995). Taken together, these findings support a connection between interferon signaling and several aspects of the DS phenotype. Therefore, the identification of the interferon pathway by the pathway connection analysis performed here is a positive control for the applicability of pathway analysis for deregulation by aberrant Hsa21 gene expression.

Hsa21 proteins involved in p38 MAP kinase signaling

Another regulatory pathway previously implicated in the pathogenesis of DS as well as Alzheimer's disease (AD) is the p38 mitogen-activated protein kinase (MAP kinase) pathway. The Hsa21 proteins found connected to this pathway are the transcription factors NRIP1, ERG, BACH1, PKNOX1 and GABPA, and the signal transducers PCP4, S100B, HRMT1L1, IFNAR1, IFNAR2 and RIPK4. Altogether, eleven Hsa21 proteins could be linked to p38 MAP kinase signaling via direct interactions. It has been proposed that neuronal apoptosis, as well as the formation of Alzheimer-type pathology, in Down syndrome is due to an aberrant re-entry of neurons into the cell division cycle (Nagy 1999). A protein coded on Hsa21, the Down syndrome cell adhesion molecule (DSCAM), interacts with and activates PAK1, and also activates both JNK and p38 MAP kinases, potentially serving a function in axon guidance (Li and Guan 2004). Increased activity of mitogen-activated protein kinases (MAPKs) (including ERK1/2, SAPKs and p38) was observed in post mortem AD and DS brains, which could not be accounted for by expression changes. The authors conclude that enhanced MAPK activity, which has

an established role in regulating neuronal plasticity and survival, could account for irregular tau phosphorylation, a feature of AD (Swatton *et al.* 2004).

Another report states that Hsa21-coded DYRK1A overexpression potentiates nerve growth factor (NGF)-mediated PC12 neuronal differentiation by up-regulating the Ras/MAP kinase signaling pathway independently of its kinase activity. Furthermore, DYRK1A prolongs the kinetics of ERK activation by interacting with Ras, B-Raf, and MEK1 to facilitate the formation of a Ras/B-Raf/MEK1 multiprotein complex (Kelly and Rahmani 2005). As a conclusion, the multitude of p38 pathway-connected proteins encoded on Hsa21 which are involved in transcriptional and regulatory activity has the potential to disturb this important MAP kinase pathway in several aspects.

Hsa21 proteins involved in growth factor signaling

Another interesting pathway potentially affected by deregulation of Hsa21 gene expression is the growth factor signaling cascade, exemplified here by epidermal growth factor (EGF). Altogether, 15 Hsa21 proteins could be linked to EGF signaling via direct interactions, including three transcription factors. The EGF receptor (EGFR) family of receptor tyrosine kinases represents both key regulators of normal cellular development as well as critical players in a variety of pathophysiological phenomena (Prenzel *et al.* 2001). Brain derived neurotrophic factor (BDNF) and nerve growth factor (NGF) constitute a family of growth factors that have been established to play critical roles in vertebrate nervous system and cardiovascular development and function (Tessarollo 1998; Huang and Reichardt 2001; Chao 2003). The actions of neurotrophins are dictated by two classes of cell surface receptors, namely the Trk receptor tyrosine kinase and the p75 neurotrophin receptor. In this study, a number of Hsa21 proteins and direct interactors have been identified to be involved in the growth factor signaling cascade, as shown in a visualization of network connections (Figure 4-2).

It is clearly visible from Figure 4-2 that proteins from chromosome 21 interact with many parts of the EGF signaling pathway on all levels of signal transduction, starting at the plasma membrane, continuing with various cytosolic activating as well as inhibitory signal transducers and going right down to the effector level of transcriptional regulation. The same is true for other signaling pathways identified here to be connected to Hsa21 proteins. The figure also shows that examination of the contribution of all deregulated Hsa21 proteins in trisomy 21 to perturbations of the

EGF pathway is a task that will require an integrated approach, taking into account other data sets, such as information on gene expression, activation states of proteins, and others. In this, the available protein-protein interaction data set can serve as basis for modeling the pathway in different physiological states.

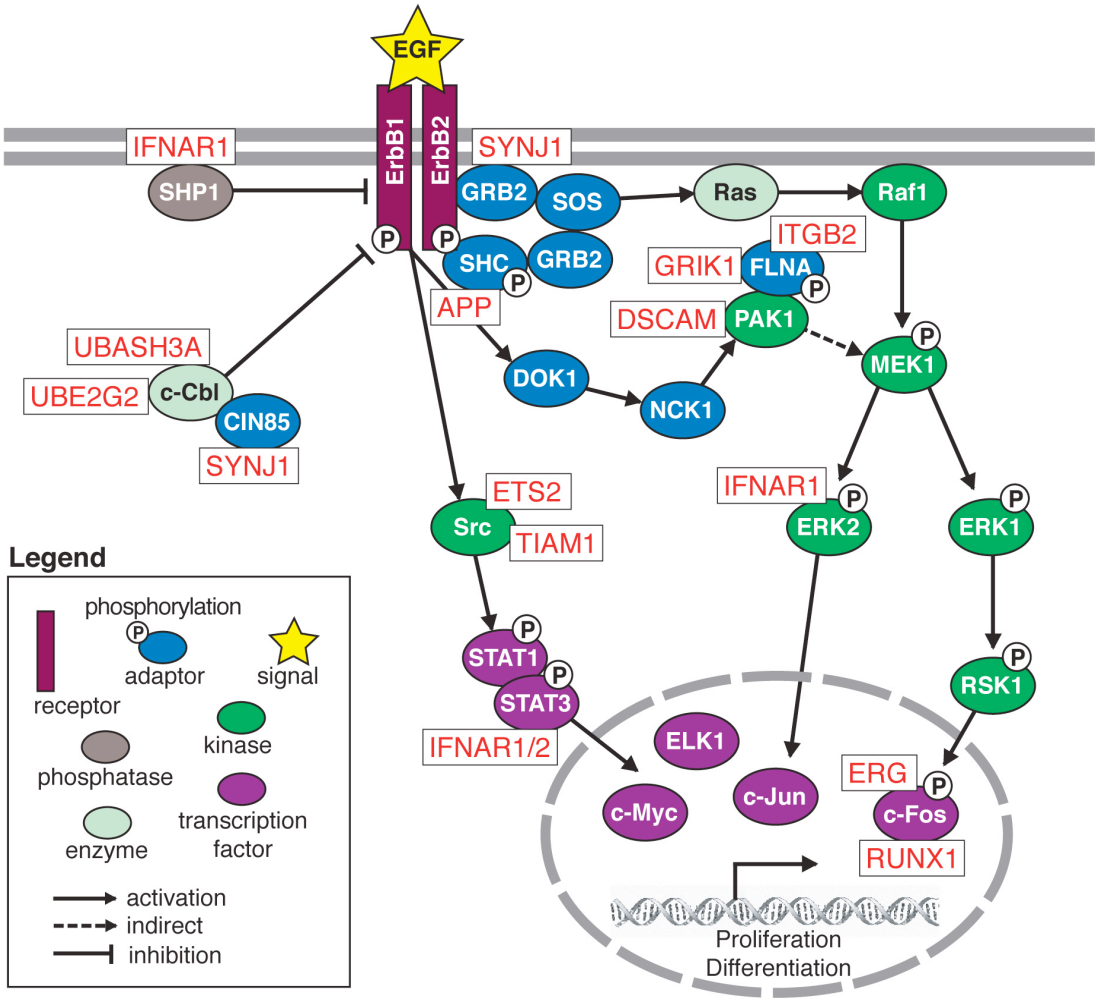

Figure 4-2. Diagram of the epidermal growth factor (EGF) signaling pathway, with interactions to proteins of chromosome 21. Figure modified from the TransPath database, pathway accession number CH000000722, map for 'EGF network'. Gene symbols of chromosome 21 proteins with known direct protein-protein interactions to proteins in the pathway have been colored in red and placed in a box next to their direct interactor.

4.7.2 Newly identified pathway connections of Hsa21 proteins

By Hsa21 pathway analysis, two regulatory pathways were identified here which to date have only poorly or not at all been associated with the pathogenesis of DS. In the following paragraphs, possible connections of these pathways to phenotypic aspects of Down syndrome will be proposed.

Hsa21 proteins involved in p73 signaling

The p73 protein, a paralog of the p53 tumor suppressor, is necessary for survival and long-term maintenance of CNS neurons in mice, including postnatal cortical neurons (Pozniak *et al.* 2002). Postnatal p73-/- mice show a mild hypoplasia of the rostral cortex and a severely disrupted architecture of the posterior telencephalon. In the developing p73-/- hippocampus, the most striking abnormality is the absence of the hippocampal fissure, suggesting a role of p73 in cortical folding and patterning (Meyer *et al.* 2004). It has been reported that p73 can associate with HIV-1 Tat, and that this association modulates Tat transcriptional and apoptotic activities in human astrocytes (Saunders *et al.* 2005). p73 is also involved in the regulation of apoptosis by its interaction with c-Abl. It was shown in rat that c-Abl activation is involved in cell signals that regulate neuronal death response to Abeta fibrils in hippocampal neurons (Alvarez *et al.* 2004). p73 in subcapslar and medullary human thymic epithelium was suggested to play a role in the regulation of the production of GM-CSF and G-CSF, which might stimulate other stromal cells such as dendritic cells, macrophages and endothelial cells (Kikuchi *et al.* 2004).

Altogether, eleven Hsa21 proteins could be linked to p73 signaling via direct interactions, including three transcription factors. Further studies are now necessary to elucidate possible contributions of each of these Hsa21 proteins, and others whose interactions are yet to be discovered, in the hippocampal malformations observed in Down syndrome.

Hsa21 proteins and the PPAR pathway

Signal transduction via peroxisome proliferator activated receptors (PPARs) shares some overlap with the p73 pathway. PPARs form heterodimers with retinoid X receptors (RXRs), and these heterodimers regulate transcription of various genes. It is well established that the PPARs are essential regulators of adipogenesis and modulator of fat cell function (Lowell 1999). Only recently, other functions could be

elucidated. PPAR-gamma is a novel target of the nerve growth factor signaling pathway in PC12 cells used to study neuronal signaling (Fuenzalida *et al.* 2005). Also, neurite extension of embryonic midbrain cells was increased after agonist-mediated upregulation of PPAR-gamma via the JNK-dependent pathway (Park *et al.* 2004). It has been proposed that PPAR-gamma may be involved in activation and development of several cell types during hematopoiesis and immune response (Greene *et al.* 2000). A peroxisome proliferator induced severe thymic and splenic atrophy in mice (Yang *et al.* 2001b).

Altogether, eight Hsa21 proteins are connected to signaling through peroxisome proliferator activated receptors via direct interactions, including five transcription factors. Therefore, pertubation of the PPAR pathway by aberrant expression of Hsa21 proteins might influence the pathogenesis of the neuronal and thymic abnormalities observed in Down syndrome.

4.8 Conclusion and outlook

In the work described here, several new data sets and resources involving the genes and proteins of human chromosome 21 were generated and analyzed. A general overview of these results can be found in Figure 4-3.

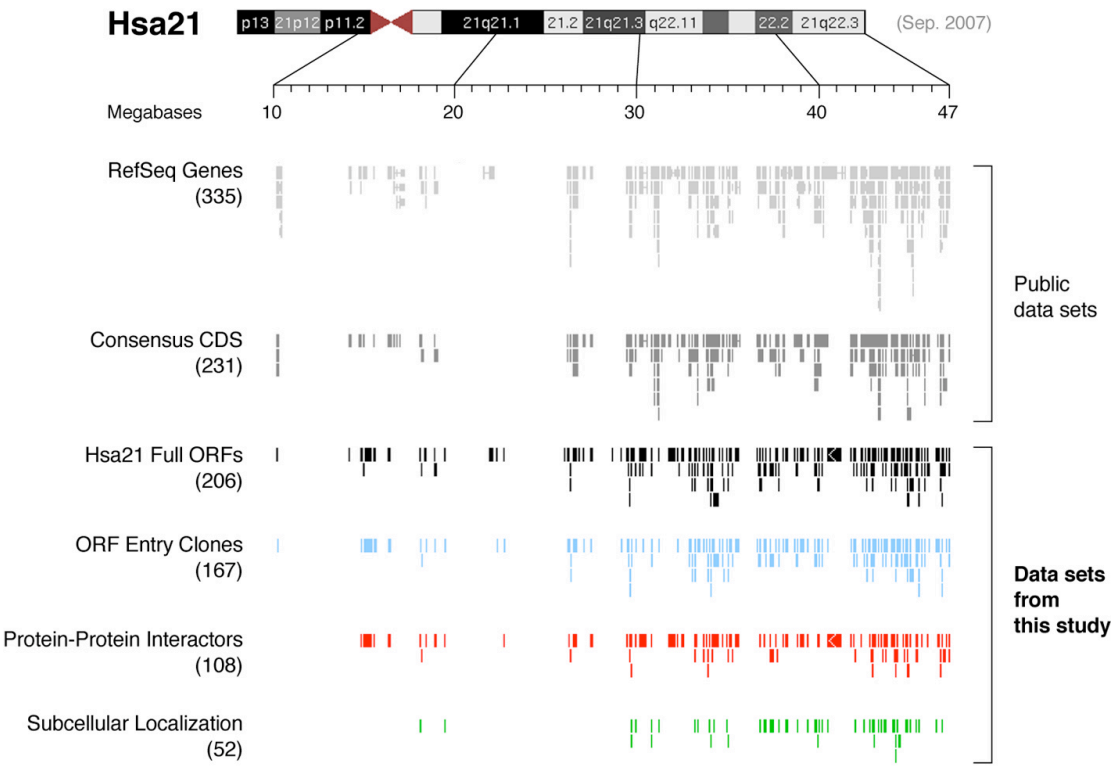

Figure 4-3. Data sets from public databases and from this project associated with the sequenced part of human chromosome 21. In GenBank, 335 **RefSeq genes** are currently basis for annotation of 231 consensus coding sequences (**consensus CDS**). Due to lack of comprehensive CDS annotation at the beginning of this project, the 284 genes in the chromosome 21 gene catalog (Watanabe et al. 2004) were basis for manual annotation of 206 full open reading frames (**Hsa21 full ORFs**). A high similarity between the different annotations can be observed (295 RefSeq genes and 205 consensus CDS overlap with the manually annotated Hsa21 full ORFs). Based on the 206 Hsa21 full ORFs, a collection of 167 **ORF entry clones** was generated. Using the clone collection, a yeast two-hybrid screen for new protein-protein interactions (PPIs), combined with the retrieval of all previously known PPIs, resulted in **protein-protein interactors** for 108 chromosome 21 proteins. Also based on the clone collection, the **subcellular localizations** of 52 Hsa21 proteins could be determined using transfected cell arrays. [Data sets viewed in UCSC Genome Browser (v165) on the genomic sequence of the human assembly of March 2006 (NCBI Build 36.1]

Future clone-based functional studies

Advances were made regarding functional annotation of chromosome 21 proteins during the work described here. To enable researchers to conduct future low- and high-throughput studies using clone-based approaches, all cloned Hsa21 ORFs will

be transferred to the German Cancer Research Center (DKFZ), which is part of the 'ORFeome Collaboration' for an unrestricted source of sequence-validated human full ORF clones (see http://www.orfeomecollaboration.org). Thus, the cloned chromosome 21 ORFs will become part of the ORFeome collection for future functional genomics studies. Also, the cloned ORFs will be included in future projects of the German cDNA Consortium.

Future affinity-based functional studies

Another important task is the generation of affinity binders for all chromosome 21 proteins to enable future 'affinity-based' proteomic analyses. In a project accompanying the work presented here, polyclonal antibodies have already been generated for 53 chromosome 21 proteins, out of 110 proteins tried. These antibodies make it now possible to perform protein expression profiling in normal versus trisomic tissues and to analyze endogenous subcellular protein localizations.

To expand the collection of affinity binders, all 167 cloned chromosome 21 ORFs have already been transferred to the 'Antibody Factory' in the framework of the European 'ProteomeBinders' initiative (see http://www.proteomebinders.org) for protein expression and generation of single-chain antibodies. The expected large set of affinity binders will further expand the molecular toolbox for functional characterization of all chromosome 21 proteins.

Data integration for network modeling

There is still a long way to go until a complete picture will emerge on all chromosome 21-associated gene functions, signal transduction pathways and gene regulatory networks that play a role in the formation of the phenotypes observed in Down syndrome. It has to be expected that only integration of different qualitative and quantitative data types, such as genetic data, gene expression profiles, information on non-coding RNAs, protein localizations, protein-protein and protein-ligand interactions will allow to construct network models that reflect physiological situations. In the case of chromosome 21, a large body of this data is already there or is becoming available, opening a large field of investigation in the systems biology of Down syndrome.