3

RESULTS

3.1 Annotation of the proteins encoded on Hsa21

3.1.1 Annotation of open reading frames (ORFs)

The sequenced part of human chromosome 21 (NCBI build 36.1) currently consists of 46,944,323 bases of high-quality DNA sequence (Hattori *et al.* 2000). Ongoing efforts to align known transcripts and ESTs and to annotate genes on Hsa21 have resulted in detailed maps of transcript and gene locations on Hsa21. Over 60,000 transcripts can be currently aligned to the Hsa21 genomic sequence, resulting in 335 annotated RefSeq genes, according to the NCBI mRNA reference sequences collection (Pruitt *et al.* 2007), including 57 uncharacterized loci and 55 pseudogenes.

These RefSeq genes are basis for annotating 231 consensus coding sequences (consensus CDS, CCDS, http://www.ncbi.nlm.nih.gov/CCDS/). When the project described in this thesis started, no consensus CDS annotation was yet available, since the CCDS project was launched later in 2005. Therefore, the genes in the chromosome 21 gene catalog published by the Chromosome 21 mapping and sequencing consortium were used as basis for the present study (http://chr21.molgen.mpg.de/chr21_catalogs/chr21_catalog9.html). This catalog contains 284 annotated genes on Hsa21 deduced by alignment of all non-redundant cDNA sequences deposited in GenBank to the Hsa21 genomic sequence using the "est2genome" program (Watanabe *et al.* 2004). Four gene categories were classified: (1) known genes corresponding to a complete mRNA (as annotated in GenBank); (2) gene predictions supported by spliced ESTs but without a complete mRNA; (3) gene signatures supported only by gene prediction programs (e.g GENSCAN, PRED*) or by homology to orthologous/paralogous genes; and (4) pseudogenes. The cDNA sequences annotated for categories 1 to 3 in the Hsa21 gene catalog represented the framework for cloning of the Hsa21 open rading frames (ORFs).

Complete ORFs (>50 bp) were found in 232 out of 284 genes and corresponding transcripts (82%). The remaining transcripts contain only partial ORFs (37

sequences, 13%) or no ORF at all (15 sequences, 5%). In cases where several mRNAs were assigned to one gene, the longest isoform was chosen to include as many protein sequence and domains as possible.

In the set of 232 complete ORFs, 27 mRNAs from two gene clusters on 21q22.1 and 21q22.3 encode paralogous keratin-associated proteins (KRTAP). These proteins are closely related to each other (Shibuya *et al.* 2004) and were excluded from further analysis, except for KRTAP6-1. The remaining set of 206 mRNAs with complete ORFs (73% of all genes) was used for downstream analysis (Hsa21 full ORFs set).

3.1.2 Retrieval of transcript sequences and ORF coordinates

Complete GenBank reports were downloaded using mRNA accession numbers as queries for the longest transcripts of the 206 genes with known full ORFs. Below is an example of the FEATURES part from such a report, in this case GenBank nucleotide accession number X93349, belonging to purkinje cell protein 4 (PCP4):

```
FEATURES             Location/Qualifiers
     Source          1..540
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="21"
                     /map="21q22.1-q22.2"
                     /tissue_type="retina"
     CDS             59..247
                     /note="human homolog of rat PEP-19"
                     /codon_start=1
                     /product="PEP-19"
                     /protein_id="CAA63724.1"
                     /db_xref="GI:1072378"
                     /db_xref="GOA:P48539"
                      /db_xref="UniProtKB/Swiss-Prot:P48539"
                      /translation="MSERQGAGPTNGKDKTSGENDGQKKVQEEFDIDMDAPETERAAVA
                      IQSQFRKFQKKKAGSQS"
ORIGIN
        1 gaattccgag gggtcgctgt gctgagcggc gggactgagc tgttgagtta gagccaacat
       61 gagtgagcga caaggtgctg ggccaaccaa tggaaaagac aagacatctg gtgaaaatga
      121 tggacagaag aaagttcaag aagaatttga cattgacatg gatgcaccag agacagaacg
      181 tgcagcggtg gccattcagt ctcagttcag aaaattccag aagaagaagg ctgggtctca
      241 gtcctagtgg gagaaccccc tcctagtcca cctgaaagca ccaaattcaa ccatcatctg
      301 tcaagaaatt aaaagaacaa caccctagag agaagtcatc cacacacaat ccacacacgc
      361 atagcaaacc tccaatgcat gtacagaaac ctgtgatatt tatacccttg taggaaggta
      421 tagacaatgg aattgtgagt agcttaatct ctatgtttct ctccattttc attcctcctg
      481 caactatttt ccttgatgtt gtaataaaat gaagttacga tgagaaaaaa aaaaaaaaaa
//
```

Nucleotide sequences of the mRNAs were extracted from the ORIGIN field. The FEATURES field was used for extraction of ORF coordinates. Protein sequences were extracted using the corresponding peptide accession numbers listed in the

/protein_id field. This procedure resulted in a list of 206 nucleotide sequences with ORF coordinates and corresponding protein sequences.

3.1.3 Primary structure of Hsa21 protein sequences

The length of ORFs encoded by the Hsa21 genes varied over a broad range of sizes. The shortest ORF (111 bp) was annotated for C21orf9, an unknown gene found expressed only in connective tissue, heart and uterus, whereas the longest ORF (10,011 bp) encodes pericentrin (PCNT), a large protein providing sites for microtubule nucleation in the mammalian centrosome.

For each of the 206 protein sequences, the theoretical isoelectric point (pI) and relative molecular mass (Mr) was calculated using the sequences upload feature of the “Calculate pI/Mw tool” on the ExPASy Proteomics Server (Gasteiger *et al.* 2003). Isoelectric points vary from as low as pI=4.1 for the acidic protein SH3BGR (SH3 domain binding glutamic acid-rich protein) up to a very basic pI=12.4 for C21orf119, an unknown gene expressed predominantly in fetal tissues. The median as well as the average pI was found to be at a physiological acidity of pH 7.2 and 7.4, respectively. The molecular masses of the Hsa21 proteins range from 4.2 to 378 kDa, with an average mass of 49.7 kDa and a median mass of 33.2 kDa. We expected that shorter ORFs, encoding smaller proteins, could be cloned more easily, which was indeed the case, as reported below.

3.1.4 Topology predictions for Hsa21 proteins

Further insight into the function of proteins via analysis of their primary sequences could be provided by prediction of membrane domains, signaling sequences and subcellular localizations. Nowadays, a multitude of programs exist which can be used for these purposes. For analysis of the Hsa21 protein sequences, the following standard programs available online were chosen:

– Prediction of signal peptides and cleavage sites: SignalP 3.0 (Bendtsen *et al.* 2004)

– Prediction of transmembrane helices: TMHMM 2.0 (Sonnhammer *et al.* 1998)

– Prediction of subcellular localizations: WoLF PSORT (Horton *et al.* 2007) and ProtComp 4 (http://www.softberry.com/)

A FASTA file containing all 206 Hsa21 protein sequences was submitted for batch analysis by these programs, except for the ProtComp program, which was a commercial service allowing only 25 executions per day per academic domain at that time.

Signal peptides and cleavage sites

46 proteins (22%) were predicted to contain a signal peptide for cotranslational translocation into the endoplasmic reticulum. The cleavage sites for signal peptidase range from amino acid 15 to 41. Among these 46 proteins, 30 are predicted to contain one or more transmembrane helices (see below), while the sixteen others are predicted to be soluble proteins, presumambly residing in compartments of the secretory pathway.

Transmembrane helices

45 proteins (22%) are predicted to contain transmembrane helices. 25 of these proteins span the membrane only once, while the others have multiple membrane domains, with up to 14 hydrophobic helices in the sodium/myo-inositol cotransporter (SLC5A3). The N-terminus of 29 proteins (64%) is predicted to reside in the outer compartment, while for the other 16 proteins, the C-terminus resides in the outer compartment.

Subcellular localizations

Two programs were used for prediction of subcellular localizations. Protein sequences were submitted in multiple FASTA format to the online prediction algorithms WoLF PSORT (Horton *et al.* 2007) and ProtComp 4 (http://www.softberry.com/), resulting in predicted localizations for all 206 protein sequences (Table 3-1). The total number of proteins for each compartment was calculated as the sum of all proteins predicted by at least one algorithm to reside in this compartment. The overlap is calculated as the fraction of proteins with the same prediction by both algorithms in relation to the total number of proteins for this compartment. The overlap in predictions ranged from only 19% for mitochondrial proteins up to 56% for plasma membrane proteins, which are more easily predicted by the combination of signal peptide and transmembrane helices.

Table 3-1. Predictions of subcellular localizations for the proteins encoded by 206 Hsa21 ORFs

Predicted localization	ProtComp 4	WoLF PSORT	Total Proteins	Overlap
Nucleus	72	64	91	49.5% (45/91)
Cytoplasm	41	40	59	37.3% (22/59)
Plasma membrane	39	36	48	56.3% (27/48)
Extracellular (secreted)	34	50	62	35.5% (22/62)
Mitochondria	13	19	27	18.5% (5/27)
Endoplasmic reticulum	5	–	5	–
Golgi apparatus	1	–	1	–
Lysosomes	1	–	1	–
Peroxisomes	–	2	2	–

The corresponding Supplemental Table S0 containing all Hsa21 proteins (including accession numbers) together with their predicted subcellular localizations can be found online at http://chr21.molgen.mpg.de/21ppi/S0_predicted_prot_loc.xls.

3.2 ORF cloning and protein expression

3.2.1 Sources of ORF primers

The set of 206 annotated full ORFs were used as a basis for PCR amplification and cloning of all open reading frames for subsequent functional analyses. Five pre-cloned full ORF Gateway clones could be obtained from the German Resource Center for Genome Research (RZPD). Additional information on these clones can be found in the Appendix, chapter 6.1.4 ('ORFs obtained from public Gateway clones').

Also, 58 pre-synthesized ORF primer pairs with Gateway extensions could be obtained from the RZPD. These primers had been designed for amplification of full ORFs using RefSeq transcript sequences. Additional information on these primers can be found in the Appendix, chapter 6.2.1 ('Primer for ORF cloning').

The remaining 143 ORF sequences were used as input for the GenomePRIDE primer design software (Haas *et al.* 2003), which resulted in the design of primer pairs for 139 ORFs. Additional information on these primers can be found in the Appendix, chapter 6.2.1 ('Primer for ORF cloning').

In total, 197 primer pairs were available for PCR amplification and cloning of full ORF Gateway entry clones.

3.2.2 Amplification and cloning of Hsa21 ORFs

The entire workflow for ORF amplification and cloning is illustrated in Figure 3-1. Briefly, ORF-specific primers were used in a first PCR round (Fig. 3-1A), followed by adapter PCRs to install the full Gateway recombination sites (Fig. 3-1B). PCR products were purified by PEG precipitation to remove primer dimers that would interfere with recombinatorial cloning. Then, amplified ORFs were cloned into Gateway entry vector pDONR201 (Fig. 3-1C). After transformation into *E. coli* strain DH5α, clones with correct inserts were identified through colony PCR using vector primers flanking the inserts. Plasmids with correct insert sizes were purified from bacterial mini cultures and used for verification of ORF identity and reading frame through 5' and 3' sequencing (Fig. 3-1D).

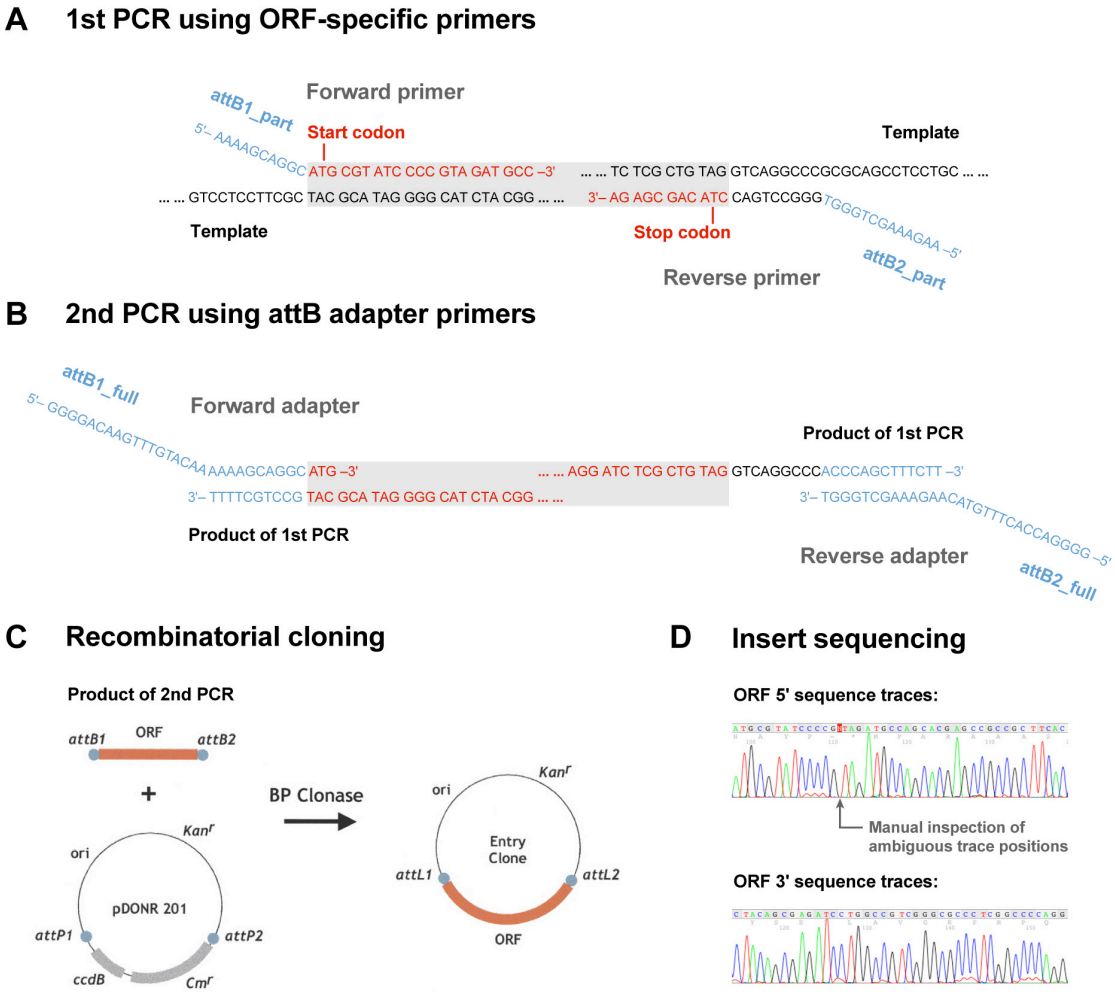

Figure 3-1. Schematic illustration of ORF amplification, cloning and verification. (A) In a first PCR round, ORF-specific primers, containing partial attB extensions (attB_part) for Gateway cloning, were used for specific ORF amplification. (B) The full attB1 and attB2 recombination sites were attached during a second PCR round with universal atttB adapter (attB_full) primers. (C) Purified PCR products were used for recombinatorial cloning (BP cloning) into the Gateway entry vector pDONR201. (D) Purified plasmid DNA was used for generation of 5' and 3' ORF sequence traces using universal sequencing primers matching vector sequences upstream and downstream of the cloned ORFs.

General strategy for template selection

Different templates for ORF cloning were tried successively, starting with inexpensive material (genomic DNA), continuing with medium-cost templates (public cDNA clones) and continuing for missing ORFs with more expensive template material (cDNA pools and cDNA from single tissues). This strategy allowed for the most cost-efficient ORF cloning process.

Four different kinds of DNA templates were used for PCR amplification and cloning of the Hsa21 ORFs: (1) Genes without introns were amplified from genomic DNA; (2) ORFs for which a full ORF cDNA clone was publicly available were cloned from the

corresponding plasmid DNA; for ORFs with no available cDNA clones, PCR was performed (3) from cDNA pools made from RNA isolated from 20 different human tissues or, if amplification from the pool did not work, (4) from cDNA made from RNA isolated out of single human tissues that were annotated to express the corresponding genes. An overview of the progressive steps of ORF cloning can be found in Figure 3-2.

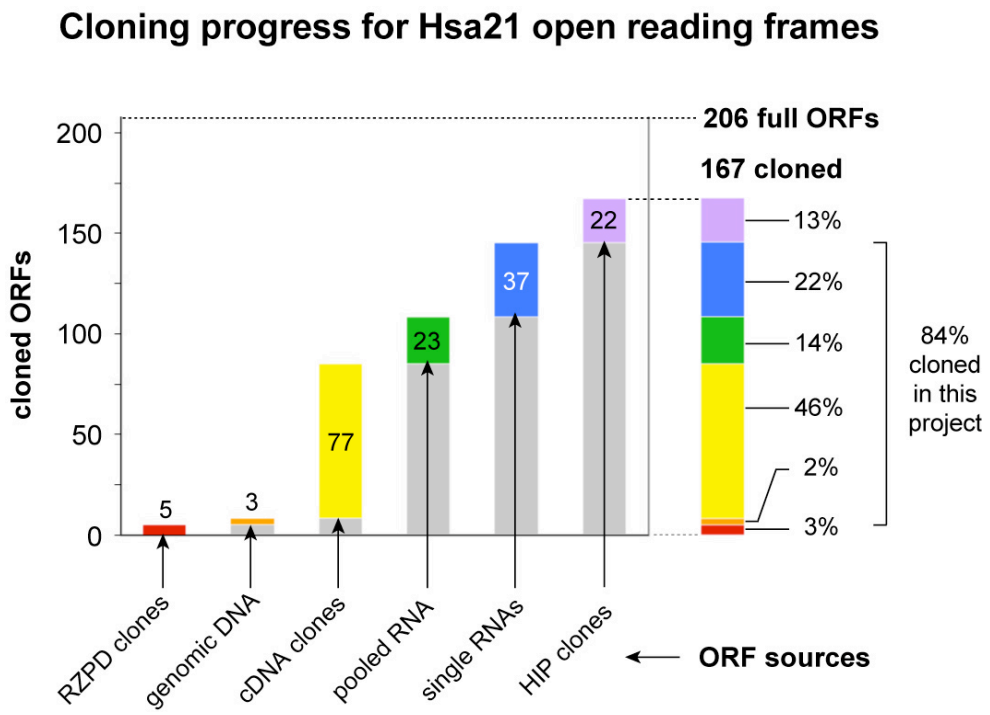

Figure 3-2. Schematic illustration of the progressive steps of ORF cloning during this thesis. Five Gateway ORF clones were available from the German Resource Center for Genome Research (RZPD clones, red bar). 3 ORFs were cloned from HeLa genomic DNA (orange bar). 77 ORFs were cloned from publicly available cDNA clones (yellow bar). 23 ORFs were cloned from pools of RNA derived from various tissues (pooled RNA, green bar). 37 ORFs were cloned from RNA derived from single tissues (single RNAs, blue bar). And 22 ORFs were obtained from the human ORFeome 3.1 entry clone collection (HIP clones, magenta bar). As result from the work described here combined with other efforts, 167 of 206 annotated Hsa21 ORFs are now available as Gateway entry clones. Of these, 84% were cloned in the framework of the thesis presented here.

Amplification from genomic DNA

Three ORFs from genes without introns were amplified and cloned from genomic DNA isolated from human HeLa cells. These ORFs encode the histone H2B family member S (H2BFS), the keratin associated protein KRTAP6-1 and the unknown protein C21orf71 (Figure 3-2, orange bar). Additional information on these ORFs can be found in the Appendix, chapter 6.1.3 (‘ORFs cloned from genomic DNA’).

Amplification from cDNA clones

For 85 ORFs, a cDNA clone containing the full ORF was available at the German Resource Center for Genome Research (RZPD). These clones were ordered as bacterial strains and cultured, the plasmids purified by plasmid mini preparation and used as template for PCR amplification. Out of 85 PCR reactions, 77 correct products were obtained and cloned (91% success rate; Figure 3-2, yellow bar), raising the total number of Gateway entry clones for Hsa21 to 85 clones. Additional information on these ORFs can be found in the Appendix, chapter 6.1.1 (‘ORFs cloned from public cDNA clones’).

Amplification from cDNA pools

For the missing ORFs, PCR reactions were performed using a pool of cDNAs generated by reverse transcription of RNAs derived from 20 human tissues as template (RNA from Human Total RNA Master Panel II, BD Biosciences). As result of these PCR reactions, 23 ORFs could be amplified and cloned (Figure 3-2, green bar), raising the number of generated Gateway entry clones to 108 clones. Additional information on these ORFs can be found in the Appendix, chapter 6.1.2 (‘ORFs cloned from cDNA’).

Amplification from cDNA from single tissues

For the remaining ORFs, PCR reactions were performed using cDNAs from 16 single human tissues as template. The cDNAs were as follows: (i) HeLa cDNA and (ii) Human Multiple Tissue cDNA Panels I and II (Clontech), each containing cDNA from eight different tissues. Appropriate tissues were chosen for each ORF according to available EST data supporting expression of the corresponding mRNAs in these tissues. Altogether, 37 additional ORFs could be amplified and cloned from cDNAs derived from the following sources (sorted by decreasing number of ORFs cloned):

Adult normal brain (10), HeLa cancer cell line (7), adult normal testis (4), adult normal placenta (3), adult normal pancreas (3), adult normal leukocyte (2), adult normal lung (2), adult normal small intestine (2), adult normal ovary (1), adult normal heart (1), adult normal thymus (1) and adult normal liver (1). Additional information on these ORFs can be found in the Appendix, chapter 6.1.2 (‘ORFs cloned from cDNA’).

As result, the number of generated Gateway entry clones was raised to 145 clones (Figure 3-2, blue bar), with 57 ORFs not amplifyable from any of the sources used here.

Retrieval of missing ORFs from the ORFeome 3.1 collection

After the various cloning steps described above, a total number of 145 Hsa21 full ORF entry clones had been generated, corresponding to 70% of the 206 annotated ORFs. Only 61 ORFs were still missing. At a later stage of the project, a genome-wide effort to generate human full ORF clones was reported (Lamesch *et al.* 2007). This so-called 'ORFeome' collection contains 12,212 ORF entry clones representing 10,214 human genes. Among these, there are 99 ORFs from Hsa21 available as entry clones, 22 of which were not yet cloned here (Figure 3-2, magenta bar). These ORF clones were obtained from the ORFeome collection. Additional information on these ORFs can be found in the Appendix, chapter 6.1.4 ('ORFs obtained from public Gateway clones').

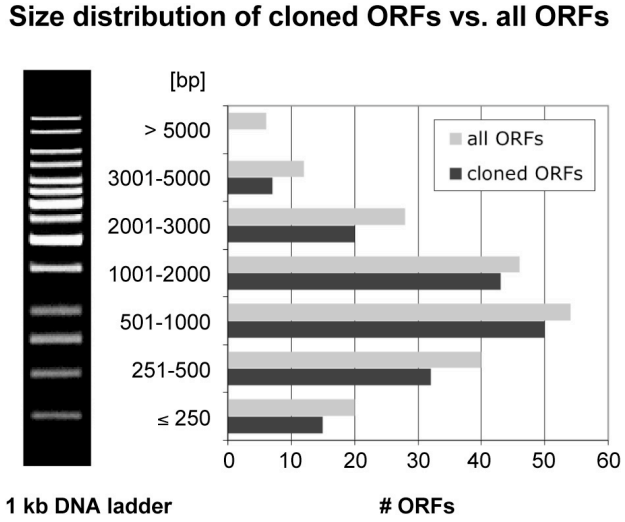

Figure 3-3. Size distribution of the 167 cloned Hsa21 ORFs versus all 206 annotated ORFs. ORF sizes were categorized into representative size bins, shown here next to a corresponding DNA molecular weight marker. The numbers of all annotated ORFs (grey bars) and all cloned ORFs (black bars) are shown next to each size bin.

The resulting Hsa21 ORF collection

The 167 ORFs available in the resulting Hsa21 entry clone collection represent 81% of the 206 ORFs annotated on Hsa21. As a comparison, the ORFeome collection of 10,214 genes (ORFeome 3.1) represents only 35% of the 28,961 genes annotated in NCBI Build 36.2 of the human genome, highlighting the 2.3x higher coverage (81%) of Hsa21 ORFs reached during the work presented here. Figure 3-3 shows the size

distribution of all 206 annotated ORFs versus all 167 cloned ORFs. The sizes of the clones Hsa21 ORFs vary from 111 bp for the unknown protein C21orf9 up to 3,663 bp for intersectin 1 (ITSN1), a cytoplasmic membrane-associated scaffold protein that icoordinates endocytic membrane traffic with the actin assembly machinery. The mean size of the cloned ORFs is 1,086 bp, which is significantly smaller than the average size of 1,623 bp among all 206 annotated ORFs. It is clearly visible that above 2,000 bp, the fraction of cloned ORFs drops significantly below the 81% reached for the whole ORF set, and no ORF longer than that of ITSN1 has been cloned yet.

3.2.3 Subcloning of ORFs

For further functional analyses, ORFs were transferred by recombinatorial cloning into expression vectors for protein expression in bacteria, yeast and mammalian cells. After transformation of *E. coli* strain DH5α with the recombined expression vectors, positive clones were identified through colony PCR of bacterial transformants, with the sizes of the PCR products indicating successful recombination of the corresponding ORFs. For each ORF, one correct expression construct was chosen for further experiments. An example set of 48 subcloned and insert-verified ORFs in Gateway destination vector pDEST17 is shown in Figure 3-4.

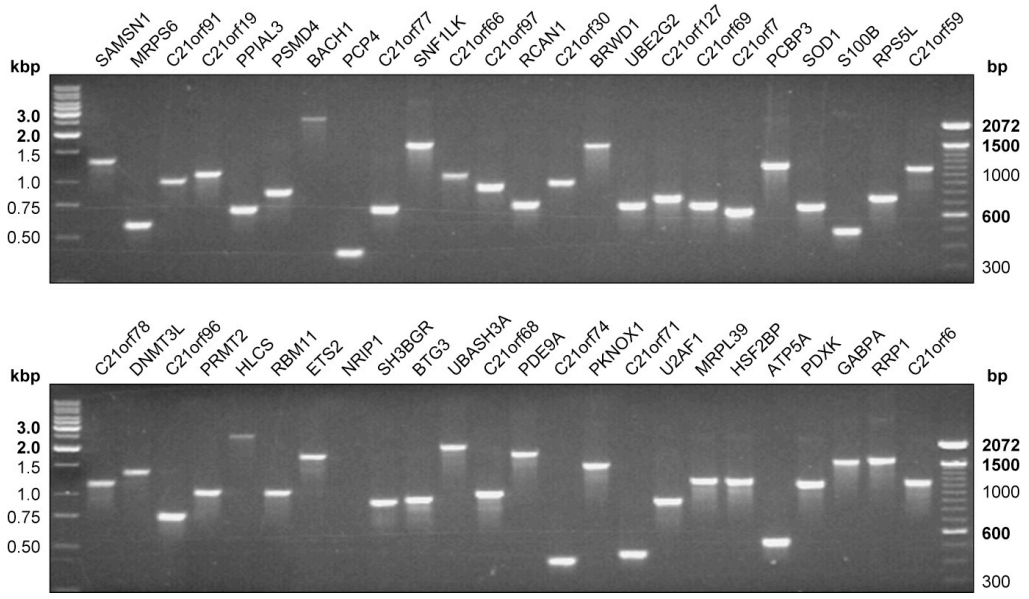

Figure 3-4. Example of 48 cloned Hsa21 ORFs in Gateway expression vector pDEST17. Cloned ORFs were amplified by PCR using primers flanking the vector inserts. PCR products were analyzed by agarose gel electrophoresis, ethidium bromide staining and detection of fluorescent bands. The observed band sizes were compared with the expected insert sizes to ensure correct ORF identities.

As described in detail in the following chapters, sets of cloned ORFs were transferred by recombinatorial subcloning into constructs for bacterial, yeast and mammalian protein expression. The success rates for the subcloning reactions were 92-93%, as shown in Table 3-2.

Table 3-2. Success rates of recombinatorial ORF subcloning for different expression vectors.

Experimental aim	Protein fusion	Vector	Success rate	Success rate [%]
Bacterial overexpression	6xHis	pDEST17	110/119	92
Eukaryotic overexpression	6xHis	pDEST26	89/96	93
Yeast two-hybrid screening	LexA	pBTM116-D9	62/67	93

3.2.4 Protein expression in bacterial cells

To test whether the cloned ORFs can serve as templates for the production of functional proteins and for producing binders (antibodies and single chain antibody fragments) for all Hsa21 proteins, the cloned Hsa21 ORFs were transferred into Gateway destination vector pDEST17 for bacterial expression with N-terminal hexahistidine fusion, also termed $(His)_6$ or 6xHis tag (Figure 3-5).

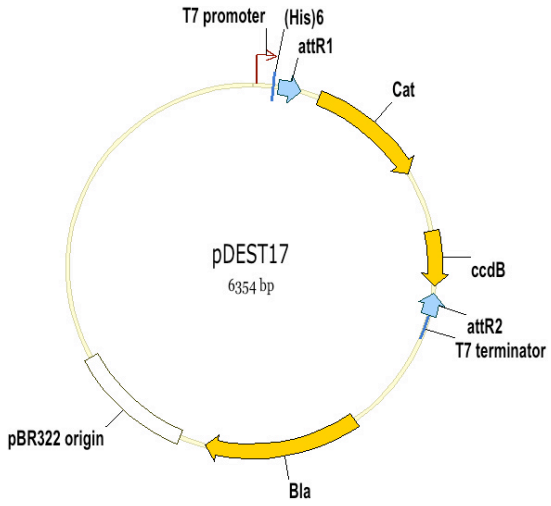

Figure 3-5. Vector map of pDEST17 used for bacterial protein expression with N-terminal hexahistidine fusion tag. Abbreviations: T7 promoter – used for IPTG-inducible transcription of the fusion ORF in *E. coli* strain Rosetta(DE3)pLysS; (His)6 – hexahistidine fusion tag; attR1/2 – Gateway recombination sites for integration of ORF; Cat – Chloramphenicol acetyltransferase used for vector propagation in *E. coli* DB3.1 strain; ccdB – controller of cell division or death B gene, inhibits bacterial gyrase; T7 terminator – terminator of transcription from T7 phage; Bla – β-lactamase, used for Ampicillin selection and propagation of ORF-containing constructs (pEXP17-ORFs) in *E. coli*; pBR322 origin – origin of replication derived from plasmid pBR322.

As bacterial protein expression strain, a BL21 derivative was chosen that is designed to enhance the expression of eukaryotic proteins which contain codons rarely used in *E. coli*. The strain called Rosetta(DE3) supplies tRNAs for the codons AUA, AGG, AGA, CUA, CCC, GGA on a compatible chloramphenicol-resistant plasmid (pRARE). Thus, the Rosetta strain provides for 'universal' translation which is otherwise limited by the codon usage of *E. coli*. The tRNA genes are driven by their native promoters. In the Rosetta derivative Rosetta(DE3)pLysS, the rare tRNA genes are present on a plasmid that additionally carries the T7 lysozyme gene to avoid leaky expression of the fusion protein in non-induced bacteria. Additional information on on this bacterial strain can be found in the Appendix, chapter 6.3.1 ('*Escherichia coli* strains').

110 ORFs were subcloned in pDEST17 and tested for expression in *E. coli* Rosetta(DE3)pLysS. Bacterial expression cultures were harvested, and expression of the transgene fusion proteins was determined by SDS-PAGE analysis after fractionation of bacteria into soluble and insoluble protein fractions. As example, Figure 3-6 shows two Coomassie-stained SDS-PAGE gels with the soluble and insoluble protein fractions from the expression of 48 Hsa21 proteins in *E. coli* cells.

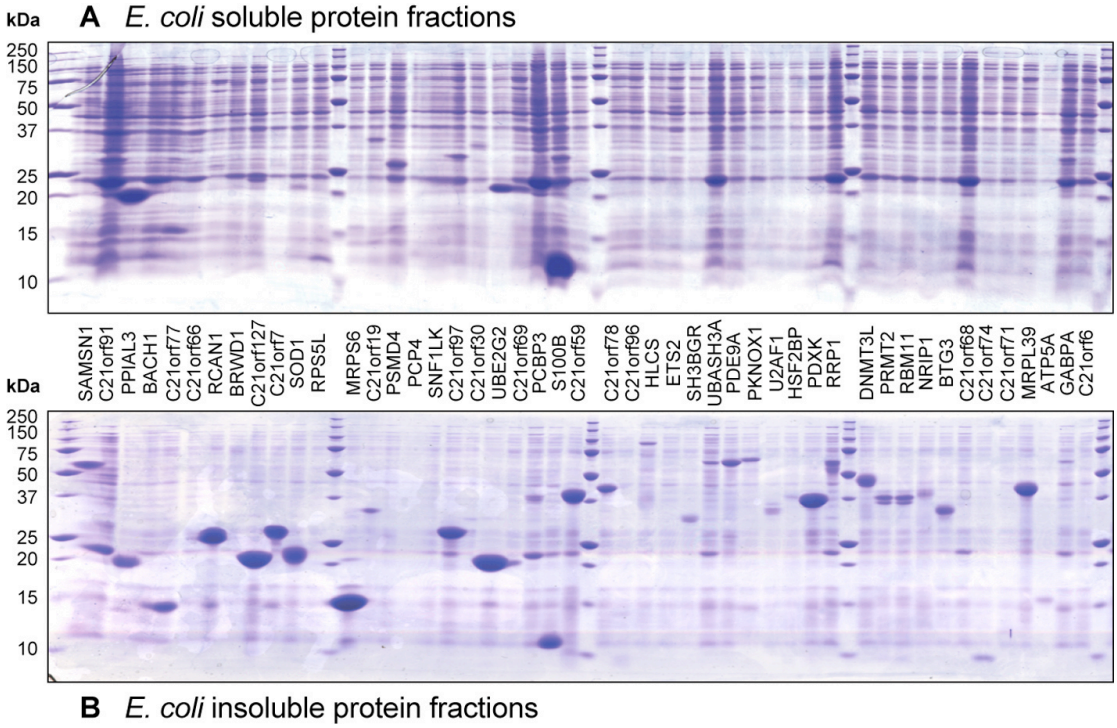

Figure 3-6. Example for the production of 48 Hsa21 proteins in *E. coli* cells. Cloned ORFs were transferred by recombinatorial LR reaction from the entry vectors into Gateway expression vector pDEST17 encoding a N-terminal 6xHis fusion. After transformation into *E. coli* Rosetta(DE3)pLysS, bacteria were grown to OD600=0.6-0.7, then fusion protein expression was induced by addition of IPTG to a final concentration of 0.1 mM. During expression for six hours, the cultures were incubated at 30°C with 200 rpm shaking. Bacterial cells were disrupted in native lysis buffer, the insoluble fraction was

collected by centrifugation, and the supernatant contained the soluble fraction. Proteins in both fractions were analyzed by SDS-PAGE and Coomassie staining. A: Proteins in the soluble supernatant after cell lysis. B: Proteins in the insoluble pellet after cell lysis.

Expression analysis for 110 Hsa21 ORFs showed that 86 proteins (78%) were expressed in the used *E. coli* strain at a level detectable by analysis on SDS-PAGE gels. The other 24 proteins (22%) were expressed either below the level of detection of the gels or not at all. Of the 86 expressed proteins, 19 proteins (22%) were only weakly expressed, six proteins (7%) showed a band of intermediate strength, and 61 proteins (71%) were strongly expressed in the *E. coli* cells.

For 17 proteins of the 24 non-expressed proteins (71%), it was expected that they were problematic cases to produce in *E. coli*. Eleven of these proteins were predicted to be secreted in eukaryotic cells by the WoLF PSORT localization prediction program, and six other proteins were potential plasma membrane proteins.

Regarding the solubility of the 86 proteins expressed in *E. coli*, analysis showed that 39 proteins (45%) were at least party soluble in the bacterial cells. No significant correlation could be detected between expression levels and solubility. The 39 soluble proteins displayed a distribution of expression levels similar to that of all expressed proteins (20.5% low, 2.5% intermediate and 77% high expression).

3.2.5 Protein expression in yeast cells

For high-throughput detection of protein-protein interactions using a yeast two-hybrid mating array set-up, 62 Hsa21 ORFs were transferred into Gateway expression vector pBTM116-D9 for expression with N-terminal fusion of the LexA DNA-binding domain (Y2H bait constructs) and into Gateway expression vector pGAD426-D3 for expression with N-terminal fusion of the GAL4 transcriptional activation domain (Y2H prey constructs). More information can be found in chapter 3.4.1 (‘Yeast two-hybrid constructs for interaction screening’) and in chapter 3.4.3 (‘Interaction confirmation by cotransformation Y2H’).

3.2.6 Protein expression in mammalian cells

For detection of subcellular protein localizations, the systems HEK293T and COS-1 were used, respectively. For this, ORFs were transferred into Gateway expression vector pDEST26 for expression with N-terminal hexahistidine tag and into Gateway expression vector pDEST474 for expression with C-terminal myc fusion tag. Also,

ORFs were transferred into Gateway expression vector pDEST475 for expression with N-terminal hemagglutinin (HA) fusion tag and into Gateway expression vector pDEST515 for expression with N-terminal FLAG fusion tag.

Experiments are detailed in chapter 3.3.1 (‘Cloning of mammalian expression constructs’), in chapter 3.4.4 (‘Confirmation by cellular colocalization assays’) and in chapter 3.4.5 (‘Confirmation by pull-down assays’).

3.2.7 Comparison of protein expression in different cellular systems

There are 71 proteins that have been tested here for expression both in mammalian cells and in bacterial cells. For these, it is possible to compare the resulting expression. 58% (41 of 71 proteins) are expressed in HEK293T cells, and 86% are expressed in bacterial cells (61 of 71). Comparison of the expression in mammalian cells with that in bacterial cells shows that 36 of the 41 proteins expressed in HEK293T cells (88%) also showed a protein band in *E. coli* lysate, whereas the other five proteins (12%) were not expressed in bacteria. Three of these non-expressed proteins are membrane proteins (CLDN8, CLDN17 and KCNE1), while the other two are extracellular or nuclear proteins (KRTAP6-1 and C21orf25, respectively).

Thirty of the 71 proteins (42%) could be expressed in *E. coli*, but were not detectable in HEK293T cells. These proteins are predicted to reside in various subcellular compartments, only one trend can be observed, namely that the list is depleted of membrane proteins, because these normally cannot be expressed in bacterial cells.

Five proteins could be expressed in neither system. These proteins are either membrane proteins (chondrolectin CHODL), membrane-associated proteins (DSCR2), secreted proteins (trefoil factors TFF1 and TFF2) or very short peptides (unknown polypeptide C21orf9).