DISSERTATION

Species-Specific Evolving Regions in the Human and Chimpanzee Genomes

angestrebter akademischer Grad
Doktor der Naturwissenschaften (Dr. rer.nat.)

Verfasser: Ricardo de Matos Simões
Matrikel-Nummer: 648262
Dissertationsgebiet (lt. Studienblatt): Molekularbiologie
Betreuer: Univ.-Prof. Dr. Arndt von Haeseler

Wien, im Juni 2010
... ontogeny does not recapitulate phylogeny: it creates it.

Walter Garstang (1922)
Abstract

Comparative analyses of the human and chimpanzee transcriptomes aim at the identification of genes whose expression pattern has changed since both species last shared a common ancestor. This approach complements the search for genes with altered function in compiling the catalogue of genetic changes responsible for the distinct phenotypes of the contemporary species. However, gene expression analyses are ultimately dependent on the availability of tissue samples. Embryonic tissues from chimpanzees are rare, if available at all. Thus, changes in the gene expression pattern in these early stages of ontogenesis, which potentially play a key role in differential development, are likely to be missed. Here I present a bioinformatics approach to identify candidates that may be expressed in a species-specific manner. Specifically, I have searched for genes in the human and chimpanzee genomes of which evolutionary sequence change shows signs of different extents of transcription-coupled-repair in the two species. Human, chimpanzee and rhesus branch-specific substitution matrices were estimated for 12,596 non-overlapping sliding windows 125 Kb in size representing the transcribed fraction in a human-chimpanzee-rhesus genome alignment. Applying a novel test statistic, 717 transcribed regions were identified in which the estimated branch-specific substitution models differ significantly between humans and chimpanzees. More specifically, it is shown that the two species differ mainly in their relative rates of $A \leftrightarrow G$ and $T \leftrightarrow C$ substitutions, and in the rate ratio of the two transition types. This pattern is expected when transcription-coupled-repair acts to different extents on the corresponding genes in the two species and provides initial evidence that these genes may be differentially expressed during early stages of human and chimpanzee development. A subsequent Gene Ontology enrichment analysis of the corresponding genes revealed an enrichment for embryonic developmental processes such as anatomical structure development (e.g., skeleton, spinal cord, brain, gut), neurogenesis, signal transduction and regulation processes for transcription, translation and replication.
Contents

1 Introduction ................................................................. 1

2 Genome alignment and metadata ...................................... 4
   2.1 Mammalian genome alignment .................................. 4
      2.1.1 Genome sequences ......................................... 6
      2.1.2 Sequence quality ......................................... 6
      2.1.3 Genome alignment ......................................... 7
   2.2 Transcribed regions in the human genome ....................... 8
   2.3 Genome alignment window extraction ............................ 8
   2.4 Transcribed and non-transcribed alignment fraction ........... 10
   2.5 Discussion ...................................................... 13

3 Modeling the evolution of DNA sequences ......................... 15
   3.1 Markov substitution models ................................. 16
   3.2 Reversible substitution models .............................. 19
   3.3 Non-reversible substitution models ......................... 21
   3.4 The Maximum likelihood .................................... 22
      3.4.1 The Likelihood-Ratio-Test ......................... 24

4 The influence of transcription on the substitution process ...... 25
   4.1 Testing the strand symmetry in transcribed genomic regions . 29
      4.1.1 The transcribed and non-transcribed alignment fraction . 29
      4.1.2 Likelihood-Ratio-Test using reversible models .......... 30
      4.1.3 Likelihood-Ratio-Test using non-reversible substitution models . 32
      4.1.4 The direction of the strand-specific substitution rates . 35
   4.2 Discussion ...................................................... 37
<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Species-Specific evolving regions in the human and chimpanzee genomes</td>
</tr>
<tr>
<td>5.1 Estimation of branch-specific substitution models</td>
</tr>
<tr>
<td>5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees</td>
</tr>
<tr>
<td>5.2.1 Normalization of $\delta$ estimated from different alignment windows</td>
</tr>
<tr>
<td>5.2.2 Overview of the significant alignment windows</td>
</tr>
<tr>
<td>5.2.3 Characterization of the substitution patterns in significant alignment windows</td>
</tr>
<tr>
<td>5.2.4 Species-specific transition patterns in the human and chimpanzee genomes</td>
</tr>
<tr>
<td>5.3 Discussion</td>
</tr>
<tr>
<td>6 Simulation and evaluation studies</td>
</tr>
<tr>
<td>6.1 Evaluation of the Single-Branch-Test</td>
</tr>
<tr>
<td>6.2 Minimal window size</td>
</tr>
<tr>
<td>6.3 Evaluation of the test statistic $\delta$</td>
</tr>
<tr>
<td>6.3.1 Effects of the alignment length on the estimate of $\delta$</td>
</tr>
<tr>
<td>6.3.2 Branch length effects</td>
</tr>
<tr>
<td>6.3.3 Effects of the outgroup divergence on the estimate of $\delta$</td>
</tr>
<tr>
<td>6.4 Normalization of $\delta$</td>
</tr>
<tr>
<td>6.5 How to decrease the computing time of the Single-Branch-Test</td>
</tr>
<tr>
<td>6.5.1 Estimating the p-values of the Single-Branch-Test from a Gamma distribution</td>
</tr>
<tr>
<td>6.6 Discussion</td>
</tr>
<tr>
<td>7 Functional analysis of candidate genes</td>
</tr>
<tr>
<td>7.1 Gene Ontology enrichment analysis</td>
</tr>
<tr>
<td>7.1.1 Selected candidate and reference genes</td>
</tr>
<tr>
<td>7.1.2 Significantly enriched Gene Ontology terms</td>
</tr>
<tr>
<td>7.1.3 Graphic representation of enriched Gene Ontology terms</td>
</tr>
<tr>
<td>7.2 Enrichment analysis for curated gene sets</td>
</tr>
<tr>
<td>7.3 Discussion</td>
</tr>
<tr>
<td>8 Outlook</td>
</tr>
</tbody>
</table>
# Bibliography

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>101</td>
</tr>
</tbody>
</table>

## A Gene Ontology Enrichment Analysis - Supplement

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1 Gene Ontology Biological Process</td>
<td>113</td>
</tr>
<tr>
<td>A.2 Gene Ontology Molecular Function</td>
<td>117</td>
</tr>
<tr>
<td>A.3 Gene Ontology Cellular Component</td>
<td>118</td>
</tr>
</tbody>
</table>

## B C2 Dataset Enrichment Analysis - Supplement

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>119</td>
</tr>
</tbody>
</table>

## C Zusammenfassung

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
</tr>
</tbody>
</table>

## D Curriculum Vitae

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>124</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Humans and their closest living relatives, the chimpanzees, separated from their last common ancestor only about 5 million years ago. Since then substitutions have driven the divergence of the two genomes such that now on average 1.2% of compared bases between humans and chimpanzees differ (CSAC, 2005). The small genetic difference of the two contemporary species is contrasted by the markedly distinct phenotypes. It was therefore hoped that those genetic changes are identifiable that account for the major phenotypic changes during the evolution of humans and chimpanzees. However, despite numerous attempts (e.g., Varki and Altheide, 2005), no major breakthrough has yet been achieved. Only a handful of genetic changes in protein coding sequences have been detected that are suspected of having a major impact on the evolution of the species phenotype (e.g., Vallender and Lahn, 2004). Accordingly, the question was raised as to whether it is indeed substitutions in the protein sequences that played a major role during the evolution of the species. Alternatively, changes in the pattern of gene expression may be the driving force of phenotypic evolution (King and Wilson, 1975). Meanwhile, a number of comparative studies have been performed on the transcriptomes of humans and chimpanzees and identified a catalogue of genes that are differentially expressed in the two species (Enard et al., 2002; Caceres et al., 2003; Uddin et al., 2004; Khaitovich et al., 2005). Genome-wide transcriptome comparisons between humans and chimpanzees using microarrays have been performed in a variety of studies (Enard et al., 2002; Khaitovich et al., 2005; Gilad et al., 2006; Marvanova et al., 2003; Caceres et al., 2003; Karaman et al., 2003; Uddin et al., 2004). The first published study compared human and chimpanzee gene expression in liver and brain tissues (Enard et al., 2002). Among the differentially expressed genes in brain tissues a higher fraction of genes were
upregulated in humans compared to chimpanzees. In the other tissues such a trend was not observed. The study of Khaitovich et al. (2005) reported that in testis the largest number of differentially expressed genes were found compared to other tissues. Brain tissues showed the least gene expression differences between humans and chimpanzees. Gene expression differences were observed to be correlated with sequence divergence of the encoded proteins. Tissue-specific genes seem to show more expression divergence than genes that have no tissue specific expression patterns.

The measurement of gene expression using microarrays is limited by high noise levels due to technical difficulties, e.g., in the hybridization procedures and also biological fluctuations of the observed gene expression levels. For example the measurement of individual gene expression levels are not constant and can differ depending on variables such as gender (Reinius et al., 2008), tissue types and different environments throughout various developmental stages, where the individual tissue and specific cell types emerge (Khaitovich et al., 2006).

Comparative gene expression analysis between humans and chimpanzees are restricted to adult tissues. Gene expression changes during early embryogenesis that may be responsible for the observed differences of the phenotypes between humans and chimpanzees have therefore not been considered. Due to ethical reasons gene expression analyses from embryonic tissue samples of humans and chimpanzees are not feasible. An experimental approach may therefore miss differences of genes active early in embryonic development since the required samples are not available; moreover, subtle changes in the gene expression level may be overlooked due to too stringent cut-offs.

Here, a bioinformatics approach is proposed to predict genes whose expression is likely to have changed during human and chimpanzee evolution. To this end, we utilize the proposed influence of transcription-coupled-repair on the substitution pattern in transcribed sequences (Green, 2003). Genes that are differentially expressed in the two species should be subject to different extents of transcription-coupled-repair, and thus should differ in their rates and patterns of DNA sequence change. In general, DNA sequences of humans and chimpanzees are believed to evolve with the same evolutionary rates and the same substitution model (Ebersberger et al., 2002; Webster et al., 2003). Exceptions to this assumption have been described (Ebersberger and Meyer, 2005), but a genome-wide analysis of this aspect has not been conducted.

The study that was conducted in this thesis aims to identify differences in the sub-
stitution model between humans and chimpanzees that are exclusively derived from a different extend of the $A \rightarrow G$ substitution rate and are presumed to result from species-specific differences in the transcription-coupled repair on the respective genomic regions. The thesis will describe a pipeline how to extract alignment fractions from a multiple genome alignment and outline the datasets that were used for the analyses.

In the second part the extent of the influence of transcriptional processes in the human genome is analyzed. In detail, a Likelihood-Ratio-Test is conducted to test for the strand symmetry between substitution rates between a transcribed and a non-transcribed alignment fraction. The strand-specific substitution rates are then compared between both datasets. The third part describes the Single-Branch-Test to identify genomic regions in the human and chimpanzee genomes, where the homogeneity between the human and chimpanzee substitution models is rejected. The substitution patterns in the identified genomic regions are then further studied for differences in strand-specific substitution patterns. In the fourth part a Gene Ontology enrichment analysis is conducted on the set of candidate genes located within the identified transcribed genomic regions. The final part describes evaluation study for the Single-Branch-Test on simulated data.
Chapter 2

Genome alignment and metadata

In this chapter I describe details about the human-chimpanzee-rhesus genome alignment that was used in this study. The concepts how publicly available genome alignments are built are briefly introduced. Finally, the pipeline I developed to handle such data for genome-wide studies is described.

2.1 Mammalian genome alignment

The prerequisite for a genome-wide phylogenetic or comparative analysis between species is an accurate multiple sequence alignment of their entire genomes. A genomic multiple sequence alignment describes the lineup of corresponding nucleotides between three or more homologous sequences. In the case of comparing entire genomes between species, the collinearity of the sequences is usually not given, due to rearrangements such as inversions and translocations of genomic segments. Homologous segments need to be identified and individually aligned by a local alignment. The strategies that are currently developed for this purpose are based on a local or combined local-global alignment approach described in the following.

The UCSC Genome Browser [Karolchik et al., 2007] provides a multiZ [Blanchette et al., 2004] multiple genome alignment from several vertebrate species that were generated by a local alignment approach [Miller et al., 2007]. All local pairwise alignments between the genomes are identified using BlastZ [Schwartz et al., 2003]. In a subsequent filtering step, segments of collinear local alignments are constructed for each pair of genomes [Kent et al., 2003]. MultiZ is based on a progressive alignment strategy, where
sequences are sequentially aligned to form a multiple sequence alignment according to a predefined guide tree. In each step pairwise or multiple alignments are combined, where the corresponding alignment sites are matched by a reference sequence that is present in both alignments. The resulting multiple sequence alignment is projected along a defined reference genome and represents any nucleotide from a genome in at most one alignment block.

The local-global approach starts with the identification of local homologous sequence segments between the individual genomes. Subsequently, the segments are aligned using a global alignment strategy. More precisely, all local alignments between two genomes are identified first. The best scoring subset of pairwise local alignments is then selected to create a rearrangement map, where syntenic regions are identified and aligned using a global alignment method (Green et al., 2003). The strategy is based on a progressive alignment according to a predefined phylogenetic tree. At each step the genomes are progressively aligned to an ancestral genome sequence that is estimated for each node in the tree. The resulting alignments are then joined using the ancestral genome as reference. Thus, this method does not depend on the choice of a certain reference genome. This strategy is for example implemented in the VISTA Genome Pipeline (Dubchak et al., 2009) that provides a multiple genome alignment between vertebrates and other genomes.

Another reference free genome alignment strategy forms the basis of the Ensembl Genome Browser (Flicek et al., 2008). The procedure described in Paten et al. (2008) first identifies local pairwise alignments between all genomes. In the next step a graph-based approach is used to order the alignments into consecutive multiple sequence segments. These segments are subsequently globally aligned. The advantage of this method is that neither the underlying phylogenetic relations of the genomes nor a predefined reference genome is required.

For the genome-wide comparative study between the human and chimpanzee genomes a pipeline was developed to efficiently extract fractions from genome alignments. Here, the pipeline is described using the UCSC multiZ genome alignment (Miller et al., 2007) containing the genomes of human, chimpanzee and rhesus. However, any other of the introduced publicly available alignments can be processed with this pipeline as well.
2.1 Mammalian genome alignment

2.1.1 Genome sequences

The genome sequences of human (hg18, finished), chimpanzee (pt2, draft, 6x coverage) and rhesus (rheMac2, draft, 5.1x coverage) were obtained from the UCSC Genome Browser database \cite{Karolchik2007}. We compressed and indexed the genome sequences from each chromosome using ”faToNib” \cite{Kent2003} to decrease storage. From the NIB-encoded sequences, segments are extracted using ”nibFrag” \cite{Kent2003}.

2.1.2 Sequence quality

Most algorithms to call nucleotides from Sanger DNA sequence traces assign an error probability to a called nucleotide. The definition of a quality value $q$ is given by \cite{Ewing1998}:

$$q = -10 \cdot \log_{10}(p)$$  \hspace{1cm} (2.1)

The quality value $q$ is computed for a given error probability $p$. For example the quality value $q$ of 40 denotes a base-call with an error probability $p$ of $1/10000$ (0.01%).

The genome sequences from rhesus and chimpanzee are in draft status, and thus errors due to low sequence quality must be taken into account. The quality files corresponding to the rhesus genome\footnote{ftp://hgdownload.cse.ucsc.edu/goldenPath/rheMac2/bigZips/rheMac2.qual.qv.gz} and the quality files corresponding to the chimpanzee genome\footnote{ftp://hgdownload.cse.ucsc.edu/goldenPath/panTro2/bigZips/chromQuals.zip} were therefore retrieved. The quality sequence file from chimpanzee chromosome 1 contains about 249 million quality values and uses about 670 Mb of disk-space stored as plain-text. The size and format of the file is inconvenient for retrieving quality sequence segments from a given genomic region of interest. To facilitate an instant access to the quality information for any genomic region of interest the quality files were indexed and compressed.

Quality information is stored as values varying from 0 to 99. Storing a quality value in ASCII format (text format) requires at most 3 bytes; an one-digit or two-digit integer and a space character as delimiter. However, a single byte can store an unsigned number ranging from 0 to 255 and is thus sufficient to encode a given quality value. Removing
also the space delimiter, the quality file can be reduced to a storage volume $\sim 30\%$ of the original size. The quality files for each of the chimpanzee and rhesus chromosomes were compressed and indexed individually (Figure 2.1 A). The compression and extraction of the quality information was done with the perl functions pack() and unpack(), respectively. The program to compress and index the quality files "qualToqib.pl" and for the extraction of a quality value sequence "qFrag.pl" were kindly provided by Sascha Strauss.

### 2.1.3 Genome alignment

I retrieved a multiple genome alignment of 28 vertebrate species including human, chimpanzee and rhesus from the UCSC Genome Browser. The genome alignment is referenced to each human chromosome and is provided in a MAF format (multiple alignment format). Multiple aligned sequence segments are called alignment "blockset". Every blockset contains the start and end coordinates, strand orientation and the corresponding aligned sequence for each particular species. From this blocksets other species than human, chimpanzee and rhesus were discarded (Figure 2.1 B).

I wanted to eliminate effects caused by sequencing errors in the chimpanzee and rhesus draft genomes in subsequent analyses. Therefore only nucleotides were considered with a quality value of at least 40. The threshold was chosen since it resembles almost the quality of the human genome. Each human chromosome MAF file is split into chunks and masked in parallel to accelerate the procedure (Figure 2.1 B). In the next step (Figure 2.1 C) the quality masked MAF files are rejoined.

The program "maf2fasta" was used to format the quality masked alignments into a fasta format for each human chromosome. Here, the list of blocksets are formatted into a linear multiple alignment along the entire human chromosome sequence string. Subsequently, all columns with a gap position in the aligned human sequence were removed from the alignments using "deleterefgap.c" (Figure 2.1 C). The resulting human genome projected sequences were separately compressed and indexed for each human chromosome. The indexed position of an alignment column for the three aligned sequences corresponds to the indexed position of a nucleotide in a human chromosome.

\[\text{ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way/}\]
\[\text{http://www.bx.psu.edu/miller_lab/}\]
2.2 Transcribed regions in the human genome

The transcribed sequence positions in the human genome sequence were identified using the information provided by Ensembl (build 50). The Ensembl (Hubbard et al. 2007) annotation of human genome transcript coordinates were retrieved using the biomaRt package (Durinck et al. 2005). The information of genomic transcripts were reformatted and compressed into a linear representation. This facilitated the identification whether an individual sequence position is transcribed.

A sequence string was defined that assigns the annotation of human genomic transcripts along the entire human chromosome sequences. The annotation of a transcript is described by the start position in the genome, genomic end position and the orientation of the coding strand. We define "0" for a human sequence position which has no annotation to a transcript, "1" and "-1" for a transcribed position on the "+" or the complementary "-" strand respectively. Nucleotide positions transcribed on both strands were represented by a "0". The string is created by the the program "transARRAY.c", which takes the genomic coordinates and chromosome sizes of each chromosome as input. Further, this sequence is indexed and compressed into a 1-byte char using the perl pack() function and extracted with the perl unpack() function by the perl scripts "traARRAY2tip.pl" and "tipFrag.pl" (Figure 2.2 B).

2.3 Genome alignment window extraction

In this section I describe the procedure to obtain alignment fractions from the processed human referenced multiple genome alignment. The transcribed fractions of 125 Kb alignment windows were extracted and oriented such that the coding strand is in '+'

Figure 2.1 (following page): Compression and indexing of a human-chimpanzee-rhesus genome alignment projected along the human genome sequence. A) The genome sequence quality files of chimpanzee and rhesus are compressed and indexed for each individual chromosome. B) Human, chimpanzee and rhesus aligned blocksets in the MAF files are extracted from the 28-way alignment and split into chunks for parallelized quality masking of the chimpanzee and rhesus sequences. C) The MAF files for each chromosome are rejoined and reformatted into a linear multiple sequence alignment along the human chromosome sequences. All alignment columns with a gap in the human sequence are removed. The aligned sequences of human, chimpanzee and rhesus projected along a human chromosome are separately compressed.
2.3 Genome alignment window extraction

**Genome Alignment Processing, Compression, and Quality Masking**

**Legend:**
- shell-pipeline
- input/output files
- custom programs
- external scripts/programs

**Diagram A:**
- Compressed chimpanzee chr1.qib
- Compressed rhesus chr1.qib
- QUAL Files (UCSC) chimpanzee chr1.qual
- QUAL Files (UCSC) rhesus chr1.qual
- qFrag
- qual2qib.pl

**Diagram B:**
- MAF (chr...)
- human-chimp-rhesus split
- getmafqual.pl

**Diagram C:**
- Concatenate MAF
- MAF Masked chr1 chr2 ...
- mal2fasta
- delete refgap.c
- FaToNib
- split fasta compress

Human chromosome projected compressed sequences:
- Human (nib)
- Chimpanzee (nib)
- Rhesus (nib)
orientation. First, an input file was created with human referenced coordinates for each alignment window with the script "hg_cutter.pl". Then the input file is split to extract the alignment windows in a parallel manner using the script "getTRFwindow.pl". The string defining the transcript annotation along the chromosomes is extracted to identify the non-transcribed sequence positions and the orientation of a transcribed sequence position. Alignment columns corresponding to the non-transcribed human nucleotides were discarded (Figure 2.2 A).

Further, the alignment fraction of non-overlapping windows of 125 KB, 250 KB, 500 KB and 1000 KB in size were extracted. An input file was created with human referenced coordinates for each alignment window according to a window size with the script "hg_cutter.pl". The input file is split to extract the alignment windows in a parallel manner using the script "getwindow.pl". (Figure 2.2 C).

2.4 Transcribed and non-transcribed alignment fraction

The extraction of the datasets corresponding to the alignment fraction of individual genes and further subsets of the transcribed regions such as from exon, intron, repeats, intergenic flanking regions of genes and a dataset with excluded CpG related transitions are described in the following. The exon coordinates of genes in the human genome were obtained from Ensembl (build 50). In total, 24,890 human-chimpanzee-rhesus alignments according to the genomic transcript start and end location of human genes were extracted using the described pipeline in Figure 2.3.

A subset of alignments was defined that accounted for genes that are expressed in

Figure 2.2 (following page): Extraction procedures for the transcribed fraction (A, B) and entire fraction (C) of windows from the human projected human-chimpanzee-rhesus genome alignment. A) The genomic transcript annotation for each human chromosome is used to construct a compressed sequence, which assigns whether a nucleotide is transcribed or non-transcribed in the human genome. B) For the 125 Kb windows the corresponding transcribed fractions are extracted in coding strand orientation and alignment columns corresponding to non-transcribed genomic regions are discarded. C) The entire fraction of non-overlapping alignment windows of sizes 125 Kb, 250 Kb, 500 Kb and 1 Mb are extracted from the compressed aligned human, chimpanzee and rhesus sequences.
the germline. The germline is defined by the line of cells starting from the zygote and leading to mature sperm and egg cells. As gene expression along the germline is not well studied in human, the identification of such genes was realized by obtaining EST sequencing data from tissues that include germline cells. Unigene (Wheeler et al., 2003) provides clustered and annotated collections of transcribed sequences derived from different tissues and species, where the human testis tissue transcriptome data was used to define genes expressed in the germline. The mapping of ensembl gene identifiers to unigene cluster identifiers were retrieved from Biomart (Smedley et al., 2009). In total, 11,285 Human Unigene clusters from testis tissue corresponded to 10,962 Ensembl genes having a primary transcript length \( \leq 500 \text{ Kb} \). Concatenating the corresponding alignment fractions into a superalignment resulted in 634 Mb gap free alignment columns with unambiguous nucleotides. Then, the alignment fraction corresponding to exons and introns of the germline expressed gene set was extracted. Here, the intron coordinates were computed along the genomic transcript boundaries for each gene in respect to the exon coordinates. The concatenated intron alignment fractions resulted in a total of 588 Mb and the concatenated exon alignment fraction in 46 Mb.

From the defined transcribed regions, the fraction corresponding to 4 categories of interspersed repeats (Smit, 1996) were extracted as well: LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), DNA transposons and LTRs (long terminal repeats). The coordinates of interspersed repeats in the human genome were obtained from the RepeatMasker (Smit et al., 1996-2004) repository.

The alignment fractions corresponding to the 4 categories of interspersed repeats were extracted from the genome alignment and merged into a superalignment for each category. Only the subset of repeat alignment fractions corresponding the germline gene set were considered. The concatenated alignment fractions comprised a total of 75 Mb (LINE), 64 Mb (SINE), 24 Mb (LTR) and 15 Mb (DNA Transposons) gapfree alignment columns with unambiguous nucleotides.

Additionally, a dataset was constructed which represented a non transcribed fraction of the genome alignment. Here, it was assumed that the genomic flanking regions of genes are not transcribed. The 50 Kb flanking regions of the 5’ and 3’ ends of the genes in the human genome were determined according to the genomic transcript start and end coordinates. The flanking region was extracted from genes with no adjacent gene

\text{http://www.repeatmasker.org/genomes/hg18/hg18.fa.out.gz}
2.5 Discussion

The genome-wide analysis of the substitution processes affecting the human and chimpanzee genomes was based on the multiZ genome alignment provided by the UCSC. First, a versatile tool was developed to extract human, chimpanzee and rhesus alignment fractions corresponding to the coordinates of the human genome that provided the infrastructure for this analysis. Then, the tool was used to extract a number of different datasets for the analysis. The procedure begins with a quality masking of the draft genome sequences of chimpanzee and rhesus. For each human chromosome the corresponding aligned sequences were extracted and stored in an indexed compressed format. The information whether an alignment position corresponds to a human transcript and its direction of transcription was additional stored in an indexed form. This information is extracted to exclude alignment columns corresponding to the non-transcribed alignment fraction.

The presented pipeline can be adapted to any of the publicly available genome alignments. Here, a reference genome needs to be defined, where each nucleotide of the reference genome is aligned to at most one nucleotide in the other genome. The main advantage of the UCSC multiZ genome alignments is that the one to one relation of each human nucleotide to a nucleotide in the other genomes is provided. The sequence quality cut-off was restricted to $\geq 40$ for the draft genomes. However, for the sequence quality masking step any other desired cut-off can be defined. For the analysis described in the thesis only the genome of human, chimpanzee and rhesus was considered. The pipeline can also be extended to any number of aligned sequences that are projected along a defined reference genome.

The pipeline was used to extract alignment windows and a variety of subsets from the transcribed fraction or the entire genome alignment. The extracted datasets correspond to (i) individual windows along the genome (sizes ranging from 125 Kb to 1000 Kb), (ii) the transcribed fraction of the 125 Kb windows and (iii) individual genes. Further, the transcribed fraction from introns, exons, the four classes of transposable elements (LINEs, SINEs, DNA transposon and LTRs) were extracted.
### Dataset Description

1) **Window alignment fractions:**

<table>
<thead>
<tr>
<th>Window Size</th>
<th>Number of Windows</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Kb</td>
<td>&gt; 742 Kb</td>
<td>1.79 Gb</td>
</tr>
<tr>
<td>500 Kb</td>
<td>&gt; 370 Kb</td>
<td>1.79 Gb</td>
</tr>
<tr>
<td>250 Kb</td>
<td>&gt; 184 Kb</td>
<td>1.79 Gb</td>
</tr>
<tr>
<td>125 Kb</td>
<td>&gt; 742 Kb</td>
<td>1.8 Gb</td>
</tr>
</tbody>
</table>

125 Kb transcribed fraction ($TRF_{125Kb}$)

- 12,596 windows (> 23 Kb)  
  - 926 Mb

**Gene Transcripts ($TRF_{Transcript}$)**

- 7,292 transcripts (> 23 Kb)  
  - 654 Mb

2) **Concatenated alignment fractions:**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigene (Testis)</td>
<td>10,962 Ensembl genes (≤ 500 Kb)</td>
</tr>
<tr>
<td>all unigene transcripts</td>
<td>634 Mb</td>
</tr>
<tr>
<td>CpG transitions excluded</td>
<td>574 Mb</td>
</tr>
<tr>
<td>Exon</td>
<td>46 Mb</td>
</tr>
<tr>
<td>Intron</td>
<td>588 Mb</td>
</tr>
</tbody>
</table>

Interspersed repeats:

- LINE | 75 Mb |
- SINE | 64 Mb |
- DNA | 15 Mb |
- LTR | 24 Mb |

3) **Non-transcribed**

- 50 Kb intergenic flanking region of genes with no adjacent gene in 200 Kb  
  - 7,499 regions  
  - 295 Mb

---

**Table 2.1:** Overview of the extracted window, transcribed and non-transcribed alignment fractions from the human, chimpanzee and rhesus genome alignment. 1) The sizes shown in brackets denote the defined minimal size for an extracted alignment window. 2) The transcribed fraction corresponds to human germ line expressed genes with a maximal primary transcript size of ≤ 500 Kb. For this purpose the genomic locations of Ensembl genes of testes expressed EST sequences represented as Unigene Clusters are used. 3) The non-transcribed fraction corresponds to 50 Kb directly flanking up and downstream regions of genes with no adjacent gene in a distance of 200 KB.
Chapter 3

Modeling the evolution of DNA sequences

This chapter introduces substitution models that are used to describe the process of single nucleotide substitutions in biological sequences.

Mutations drive the molecular evolution of biological sequences such as DNA, RNA and protein sequences. Causes of mutations are replication errors in a dividing cell, chemical mutagens, radiation or viruses and transposons. These events occur in all cells of an organism, but only the mutations that were inherited by germline cells generation after generation are accumulated in contemporary sequences.

The mutations of single nucleotides (point mutations) are distinguished from mutations, that affect whole sections of genomic regions. The mutations of whole genomic sections that are e.g., derived from chromosomal rearrangements, where several consecutive nucleotides are deleted, inserted, translocated (from another genomic region) or duplicated. Point mutations are deletions or insertions of single nucleotides or substitutions of a single nucleotide by another nucleotide.

Sequence evolution is in general investigated by analyzing single nucleotide substitutions inferred from an alignment between sequences. This chapter will introduce Markov models to describe such a substitution process in a stochastic framework. The formulas shown in this chapter are based on Salemi and Vandamme (2003), Felsenstein (2004) and Yang (1994b).
3.1 Markov substitution models

For a given alignment, e.g., between two homologous sequences, the substitution process is in general investigated from the observed substitutions (Figure 3.1).

![Figure 3.1: Point mutations inferred from a pairwise sequence alignment. Mismatch nucleotides are called substitutions. An insertion/deletion of a single nucleotide is represented by a gap character.](image)

The observable number of substitutions is limited due to a saturation of the substitutions (Figure 3.2). Therefore the number of observed substitutions generally underestimates the actual number of substitutions that have occurred. The saturation of substitutions can be explained by the following scenarios: 1) Backmutation: a substitution of a nucleotide is not observed when a second substitution recovers the initial state of the nucleotide; 2) Multiple substitutions: only one substitution is observed when multiple substitutions occur at the same site; 3) Parallel substitutions: when comparing two homologous sequences a substitution is not observed when the same substitution occurs at the same site in both sequences. To overcome such problems a substitution model is used to describe the evolution of aligned sequences.

Suppose that the rate of changes in DNA sequences from one nucleotide into another follows a Markov process with four states, where each state represents one of the four nucleotides (A, G, C, T). In general, the substitutions described by a Markov model are assumed to occur independently for each site in the sequence. Assume the substitution of a nucleotide to another nucleotide (Figure 3.3) within a DNA sequence $X$ from a time point $t_0$ to $t$: 
3.1 Markov substitution models

Figure 3.2: The number of observed differences and expected number of substitutions per site (branch length) between two DNA sequences. The observed differences are defined by the mismatch characters per site observed between two homologous sequences and the expected number of substitution per site (branch length) is estimated. The saturation of observed substitutions between two homologous aligned sequences are shown. The observed number of substitutions (black line) underestimates the expected number of substitutions (dotted line), e.g., due to parallel, multiple and backmutations as explained in the text.

\[ X_{t_0} \rightarrow X_t \]

The substitution model describes the probability \( P_{ij}(t) \) to observe nucleotide \( j \) at time \( t \), if nucleotide \( i \) is present at time \( t_0 \). The probability is solely dependent on the initial state without considering the putative changes that occurred within a given time period. The transition probabilities follow an exponential distribution, where the probability for a substitution increases with time \( t \). In general, the transition probabilities from one nucleotide into another are described in a \( 4 \times 4 \) matrix \( P(t) \):

\[
P(t) = \exp(Q(t))
\]  

(3.1)

The transition probabilities in \( P(t) \) are a function of time \( t \) and can differ at each sequence position. The rate matrix \( Q \) describes the process in an infinitesimal time
interval and therefore does not depend on time. The matrix $Q$ has off-diagonal entries $q_{ij} > 0$ that give the instantaneous substitution rates per unit time at which nucleotide $j$ is replaced by nucleotide $i$ ($i \neq j \in \{A, C, G, T\}$). The diagonal elements of $Q$ are given by

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

such that rows sum up to zero. Let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be the equilibrium frequency of nucleotide $i$ ($i \in \{A, C, G, T\}$). The rate matrix $Q$ is typically normalized such that the total rates of changes is 1 per unit time by scaling the negative sum of the diagonal components (trace) to 1.

$$-\sum_i \pi_i q_{ii} = 1$$

The normalization allows to measure time in expected numbers of substitutions per site.

The transition probability matrix $P(t)$ is computed from the matrix exponential of $Q$ (equation 3.1). There are several approaches to compute the matrix exponential (Moler and Charles 2003). The most popular method to compute the exponential is the eigenvalue decomposition method (Bailey et al. 1990). Assume that the transition probability matrix $P(t)$ has a spectral decomposition:

$$P(t) = U \cdot \text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t}) \cdot U^{-1}$$

The matrix $P(t)$ is described by the product of left 4 eigenvectors in the matrix $U$, the diagonal matrix of the 4 eigenvalues $\lambda_i$ and the 4 right eigenvectors in the matrix $U^{-1}$ of $Q$. The eigenvectors describe the properties of a matrix and are equal between $P(t)$ and $Q$. The exponential of the matrix is obtained by replacing the eigenvalues $\lambda_i$ by $e^{\lambda_i t}$. The rate matrix $Q$ can be expressed from equation 3.4 as

$$Q = U \cdot \text{diag}(\lambda_1 t, \lambda_2 t, \lambda_3 t, \lambda_4 t) \cdot U^{-1}.$$ 

The substitution process is generalized in a so-called Markov process which describes the relative rates of change of the nucleotides along a sequence defined in the rate matrix $Q$. The substitution model defined in $Q$ is in general restricted to the following
3.2 Reversible substitution models

The most commonly used substitution models assume the reversibility of the substitution process. In a reversible model the substitution rate between any nucleotide $i$ to $j$ is the same as the reverse substitution rate from nucleotide $j$ to $i$:

$$q_{ij} = q_{ji} \quad (3.6)$$

This class of substitution models are called time-reversible substitution models that are introduced in the following.

The general time-reversible model (GTR) describes the evolutionary process by the
following instantaneous rate matrix $Q$ (Lanave et al., 1984):

$$Q = \begin{pmatrix}
-\alpha\pi_C & \beta\pi_G & \gamma\pi_T \\
\alpha\pi_A & - & \delta\pi_G & \varepsilon\pi_T \\
\beta\pi_A & \delta\pi_C & - & \eta\pi_T \\
\gamma\pi_A & \varepsilon\pi_C & \eta\pi_G & -
\end{pmatrix}$$

(3.7)

The components $q_{ij}$ of matrix $Q$ are defined as parameters, where the rate parameters ($\alpha$, $\beta$, $\gamma$, $\delta$, $\varepsilon$, $\eta$) and the parameters for the equilibrium nucleotide frequencies $\pi_i$ for $i \in \{A, C, G, T\}$ are defined. The off-diagonal components of the matrix $Q$ quantify the flow of a nucleotide exchange by another nucleotide. The diagonal elements in $Q$ are defined such that a row in the matrix sum to zero and represent the total rate of an exchange from a nucleotide $i$ to any other nucleotide $j$.

The number of free parameters of a substitution model depend on the defined symmetry assumptions and the scaling of the parameters in the substitution model. For example in the GTR model the number of free parameters are defined in the following way: The sum of the nucleotide frequencies is one,

$$\sum_{i=A,C,G,T} \pi_i = 1$$

(3.8)

thus, 3 free parameters are sufficient to describe the nucleotide frequencies. Assuming a reversible process (equation 3.6) and scaling the rate matrix $Q$ to one (equation 3.3), the substitution rates can be described by 5 free rate parameters. In total, the GTR model has therefore 8 free parameters.

Within the GTR model (8 free parameters) a variety of reversible submodels can be defined, where the number of parameters are reduced by introducing additional restrictions. In the following a few of these submodels are introduced. The TN93 model (Tamura and Nei, 1993) describes a parameter for the ratio between transitions and transversions ($\kappa$) and another parameter the ratio between purine and pyrimidine transitions $\gamma$. The TN93 model has 5 free parameters with 2 free rate parameters and 3 parameters for the nucleotide frequencies. The HKY85 model (Hasegawa et al., 1985) describes only the parameter for the ratio of transition and transversion rate $\kappa$. The HKY85 model has 4 free parameters, 1 free rate parameter and 3 free parameters for
the nucleotide frequencies. The F81 describes one parameter for the substitution rate and remains with 3 free parameters for the nucleotide frequencies. By assuming uniform nucleotide frequencies ($\pi_i=0.25$) and a parameter for the ratio between purine and pyrimidine substitutions describes a K80 model (Kimura 1980) that has only 1 free parameter. The simplest model of DNA evolution is the JC model (Jukes and Cantor 1969), where the rate of change from one nucleotide $i$ to another nucleotide $j$ is the same for all substitution rates and the equilibrium nucleotide frequencies are uniform ($\pi_i=0.25$). The JC model has no free parameters.

### 3.3 Non-reversible substitution models

The most general Markov model is shown in Figure 3.4, where the rates are defined by the product of the rate of a substitution from one nucleotide $i$ to the nucleotide $j$ and the nucleotide frequency $\pi_j$ (Figure 3.4). The generalization to describe the substitution process leads to a Markov model that does not assume reversibility. The non-reversible substitution describes the instantaneous rate matrix $Q$ as following

$$Q = \begin{pmatrix}
-a\pi_C + b\pi_G + e\pi_T & a\pi_C & b\pi_G & e\pi_T \\
-g\pi_A + d\pi_G + e\pi_T & -g\pi_A + d\pi_G + e\pi_T & d\pi_G & e\pi_T \\
-h\pi_A + i\pi_C + f\pi_T & i\pi_C & -h\pi_A + i\pi_C + f\pi_T & f\pi_T \\
-j\pi_A + k\pi_C + l\pi_G & k\pi_C & l\pi_G & -(j\pi_A + k\pi_C + l\pi_G)
\end{pmatrix}$$

**Figure 3.4:** The most general instantaneous rate matrix $Q$. The diagonal elements are defined such that the rows sum up to zero. The rate parameters $a, b, c, d, e, f, g, h, i, j, k, l$ define the relative rates of the substitution from one nucleotide to any other nucleotide. The parameters $\pi_A, \pi_C, \pi_G, \pi_T$ define the nucleotide frequencies.

The so called unrestricted model (Yang 1994b) has 12 parameters (Greek letters), one for each substitution type. Note, in the unrestricted model the nucleotide frequencies are...
a function of the parameters. The reason to this constraint is that there might not exist
an equilibrium nucleotide frequency for the substitution model (the largest eigenvalue
of \( Q \) is not 0). In the unrestricted model the actual base frequencies of the sequences
are used as the equilibrium nucleotide frequencies. The total number of substitutions
are standardized to one, the model has in total 11 free parameters.

A variety of additional constraints can be defined to simplify the unrestricted model.
In general, substitutions can not be distinguished on which strand they occur. A sub-
stitution that occurred on one strand will result in the complementary substitution on
the other strand. In [Sueoka (1995)] a strand symmetric model was introduced, where
the substitution rates between complementary substitutions get the same parameter
assigned.

\[
Q = \begin{pmatrix}
A & G & C & T \\
A & - & \alpha & \beta & \gamma \\
G & \lambda & - & \delta & \varepsilon \\
C & \beta & \gamma & - & \alpha \\
T & \delta & \varepsilon & \lambda & -
\end{pmatrix}
\] (3.10)

In the matrix \( Q \), for example, \( A \rightarrow G \) is strand complementary to \( T \rightarrow C \) and hence are
assigned to the same parameter \( \alpha \). The model describes parameters for 6 complementary
substitution pairs and has 5 free parameters.

### 3.4 The Maximum likelihood

The substitution models are used to compute the distances and for the estimation of
the substitution process among the sequences observed in an alignment that are repre-
sented by a phylogenetic tree (Figure 3.5). The substitution model describes the process
along the branches that most likely generated the observed sequences. In general, the
parameters of a specified substitution model \( Q \) are estimated to compute a likelihood
for the model and a given tree. Let \( T \) denote a phylogenetic tree and \( M \) a substitu-
tion model for a sequence alignment \( D \). The likelihood of the data given a substitution
model is computed as follows. Suppose \( D_s \) be a site pattern or column in the alignment
(Figure 3.1). For this site pattern the product of the transition probabilities in \( P(t) \) for
The likelihood for the site pattern is defined by the sum of the products for all possible nucleotide states at the internal nodes. The likelihood to observe $D_s$ is denoted by:

$$P(D_s|T, M)$$ (3.11)

The likelihood function $L(T, M)$ computes the total likelihood over all sites in an alignment by the sum of the log likelihoods over all sites:

$$\log L(T, M) = \sum_{s=1}^{m} \log P(D_s|T, M)$$ (3.12)

The log scale is used to avoid arithmetic underflow, because the probabilities of the site patterns are usually close to zero. The parameters of the substitution model are separately estimated by maximizing the likelihood function $L$. In general, iterative global optimization procedures are used for this purpose (Swofford et al., 1996). The likelihood method allows to estimate the parameters for a given model to describe the substitution process leading to the observed sequences in an alignment. A method to infer the most appropriate substitution model or number of parameters for a given alignment is described in the following section.
3.4.1 The Likelihood-Ratio-Test

A substitution model with several parameters such as the GTR model or the unrestricted model described in the previous sections can be simplified when further symmetries are assumed between the parameters. When the number of parameters are reduced within a substitution model, the simplified model is called nested within the more complex model. Usually, the more complex substitution model provides a better fit to the data (higher likelihood) than the model with fewer parameters. However, the more complex model will also have a larger variance in the estimation that might lead to an overfit of the data.

The standard method to compare the likelihood between two nested substitution models to select the most appropriate is a Likelihood-Ratio-Test (LRT) (Felsenstein, 2004). The statistic of the test is described by:

$$ \Delta = 2(\log_e L_1 - \log_e L_0) $$ (3.13)

The log likelihood of the parameter rich model is denoted by $L_1$ and the null model with fewer parameters by $L_0$. In this particular case the $\Delta$ statistic follows approximately a $\chi^2$ distribution with the degree of freedom according to the differing numbers of free parameters between the null and alternative model.
Chapter 4

The influence of transcription on the substitution process

In this chapter the role of strand-specific substitution patterns in transcribed genomic regions are described. The extent of the $A \rightarrow G$ has been suggested to be specifically increased over the $T \rightarrow C$ in transcribed genomic regions and was explained to result from transcription-coupled-repair. A genome-wide study is conducted for the estimation of the strand-specific substitution process from the human-chimpanzee-rhesus genome alignment. Using a Likelihood-Ratio-Test, it is shown that the unrestricted nonreversible substitution model gives a significantly better fit to the data from a transcribed alignment fraction than a simpler model in which $A \rightarrow G$ and $T \rightarrow C$ are assigned to the same rate parameter. Furthermore, the extent of the differences between the substitution patterns in transcribed and non-transcribed genomic regions are estimated and compared.

In the course of his work, Erwin Chargaff has established several rules describing the relative frequencies of the four nucleotides A, G, C, and T in DNA sequences. His first parity rule states that A occurs with the same frequency as T, and G occurs with the same frequency as C in all DNA molecules. This rule contributed substantially to the understanding that native DNA occurs as a double helix where two antiparallel DNA strands are paired via H-bond formation between the complementary bases A and T, and G and C, respectively. While the general applicability of the first parity rule is undisputed, Chargaffs second rule of intrastrand parity is not. According to this rule, the complementary bases should occur with the same frequency also along a single DNA strand. However, this intrastrand parity rule is only approximately followed on a genome-wide level and substantial deviations are observed on a local level. [Touchon and]
This raises the question how the relative frequency of the four nucleotides along a DNA sequence is locally modulated. The base composition of contemporary DNA sequences is determined by the substitution process, which changes DNA sequences over time (c.f., Chapter 3). Thus, one needs to hypothesize that genomic regions with different base compositions differ also in their substitution processes. In the past years several biological factors have been proposed to influence the substitution process on region specific scale and, thus the base composition of DNA sequences: mutation bias (e.g., Wolfe et al. 1989), natural selection (e.g., Eyre-Walker and Hurst 2001), biased gene conversion (BGC, e.g., Duret and Arndt 2008). Common to all these factors is that they affect both DNA strands to the same extent.

More recently, the focus has been extended to strand-specific processes that affect the substitution process of the two DNA strands to a different extent (Touchon and Rocha, 2008). In the literature a variety of terms are used to distinguish between the two DNA strands. In this chapter the two DNA strands are distinguished by the terms leading and lagging strand when the process of DNA replication is referred. For the process of DNA transcription the non-coding strand is defined as the template for the RNA Polymerase to copy the mRNA strand and the coding strand is defined as the opposite strand. As a unifying principle it is suggested that single stranded DNA is more vulnerable to nucleotide changes than either double stranded DNA or single stranded DNA protected by large protein complexes. For example, it has been shown in prokaryotes that deviations from the intrastrand parity rule can be attributed to processes related to DNA replication. The lagging strand, which remains temporarily single-stranded during replication, is distinguished from the leading strand by having compositional skews of G &gt; C, and, although less prominent, of T &gt; A (Lobry 1996). This would correspond to T → C and to a lesser extend A → G transitions occurring more frequently on the lagging than on the leading strand. The correlation of this compositional skew and replication process is so marked, that origins of replication could be reliably predicted by analyzing compositional skews in bacterial genomes (Lobry 1996). In eukaryotes a similar but less pronounced influence of replication on compositional skews has been observed (Huvet et al. 2007). Transcription is the second biological process that leaves one DNA strand transiently single stranded and has been associated to compositional skews. After separation of the template and the coding strands by the transcription machinery, only the template strand is protected by the protein complex of the RNA
polymerase. In both bacteria and eukaryotes it has been shown that this difference between the two strands results in a different mutation pattern. In bacteria \( C \rightarrow T \) transitions have been found to occur more frequently on the coding strand than on the transcribed strand (Francino and Ochman 2001). In mammals, it was postulated that the rate of the \( A \leftrightarrow G \) transition is higher on the coding than on the transcribed strand (Green 2003; Mugal et al. 2009). As a result a different compositional skew is abound for the transcribed than for the non-transcribed fraction of the genome. One biological model that is used to explain these rate differences between the complementary strands is provided by the process of transcription coupled repair (TCR; Figure 4.1). According to this model, an \( A \rightarrow G \) mutation occurs more frequently on the coding strand that results in G:T mismatches along the genomic transcribed fraction of a gene. In vivo and in vitro assays have shown that the \( \text{MutS}^\alpha \) repair enzyme binds efficiently to such G:T mismatches (Jiricny et al. 1988; Lamers et al. 2000). It was therefore postulated that the binding of \( \text{MutS}^\alpha \) to such a bulky mismatch is able to stall the transcription machinery and induce a specific repair mechanism.

In this chapter a study is conducted to describe strand-specific substitution patterns in the human genome that is estimated from a transcribed and a non-transcribed alignment fraction. Bielawski and Gold (2002) analyzed replication associated strand asymmetry between the leading and lagging strand in mitochondrial DNA sequences. For this purpose they developed a Likelihood-Ratio-Test (c.f. chapter 3) to infer, whether a more complex substitution model gives a significant better fit to the data than a simpler strand symmetric model. In the strand symmetric model, a single rate parameter is assigned for each pair of complementary substitution rates (Sueoka, 1995, see chapter 3, substitution model defined in 3.10). However, the more complex non strand symmetric model with one additional parameter allows to distinguish between complementary substitution rates. Bielawski and Gold (2002) used three more complex substitution models

**Figure 4.1 (following page):** Transcription-coupled-repair in mammalian genomes as proposed by Green 2003. a) During transcription the complementary DNA strands are transiently single stranded. While the transcribed strand (non-coding strand) is protected by the transcription machinery the coding strand is more exposed to DNA damage. b) MutS\(^\alpha\) recognizes and binds efficiently G:T mismatch pairs c) A DNA repair machinery is activated at RNA polymerase (RNAP) stalled lesions. d) Preferential repair of non-coding strand that favors the incorporation of \( G \) that leads to an increased fixation of \( A \rightarrow G \) mutations on the coding strand.
Chapter 4: The Influence of Transcription on the Substitution Process

a) Transition A → G
b) RNAP

MutSα

Stalled RNAP

MutSα

c) Fixation of A → G on coding strand

Non-coding Strand
that assigned individual rate parameters for $G \rightarrow A \neq C \rightarrow T$, $A \rightarrow G \neq T \rightarrow C$ and $C \rightarrow A \neq G \rightarrow T$, respectively.

A genome-wide analysis using the Likelihood-Ratio-Test of [Bielawski and Gold (2002)] for the analysis of the strand asymmetry of substitutions associated to transcription has not been performed to date. The method allows to answer the question whether a more complex non strand symmetric substitution model would be more appropriate to describe the substitution patterns in the corresponding transcribed genomic regions from the human-chimpanzee-rhesus genome alignment. In the following a similar Likelihood-Ratio-Test will be described to test for the strand symmetry between pairs of substitution rates estimated from sampled alignments corresponding to human transcribed and non-transcribed genomic regions. The direction of a particular strand-specific substitution rate is quantified as ratio to the corresponding strand complementary substitution rate. Finally, the difference of substitution rates between transcribed and non-transcribed alignment fraction are discussed.

4.1 Testing the strand symmetry in transcribed genomic regions

4.1.1 The transcribed and non-transcribed alignment fraction

The transcribed genomic regions from the human genome were selected from genomic transcript coordinates corresponding to genes that are also expressed in the germline. A set of non-transcribed genomic regions were defined by the intergenic left and right flanking regions of genomic transcripts, where the flanking region was not in the direct neighborhood of another gene. The alignments corresponding to the defined genomic regions were extracted from the human-chimpanzee-rhesus genome alignment (see chapter 2, Table 2.1). In the following, the alignment set corresponding to the transcribed regions will be termed transcribed alignment fraction and the alignment set corresponding to the non-transcribed genomic regions will be termed non-transcribed alignment fraction. Additional subsets from the transcribed alignment fraction corresponded to exon, intron, interspersed repeats and a subset where CpG related transitions were excluded.

The datasets were sampled from the concatenated alignments fractions corresponding
to a particular dataset. The alignment set of the transcribed fraction and the alignment set of the non-transcribed fraction were each concatenated (see section [2.4]). The transcribed alignment fraction comprised a total of 634 Mb and the non-transcribed alignment fraction a total of 295 Mb excluding all columns with gap characters, ambiguous or masked nucleotides. From both datasets 1000 bootstrap alignments, each consisting of 100 Kb in size were sampled.

4.1.2 Likelihood-Ratio-Test using reversible models

Two reversible null models with 4 rate parameters are defined assuming strand symmetry of transversion rates $M_1^0$ ($A \leftrightarrow C=T \leftrightarrow G$) and the strand symmetry of transition rates $M_2^0$ ($A \leftrightarrow C=T \leftrightarrow G$). The alternative model is defined by the general reversible model (GTR) with 5 free rate parameters (Figure 4.2).

![Reversible substitution models](image)

Figure 4.2: Reversible substitution models used for the Likelihood-Ratio-Test. The null models have 4 free rate parameters, each, and the alternative GTR model has 5 free rate parameter.

The log likelihood of the alternative model $M_{GTR}$ and the strand symmetric models $M_{01}^0$ were estimated using paml (Yang 2007). The substitution models $M_{01}^0$ differ by having one parameter less than the alternative model $M_{GTR}$. The statistic for the Likelihood-Ratio-Test follows therefore approximately a $\chi^2$ distribution with one degree of freedom. A confidence limit of 5% was chosen for the tests.

The fraction of Likelihood-Ratio-Tests rejecting the null hypothesis of a strand symmetry between a pair of substitution rates was computed and compared between the sampled alignments corresponding to human transcribed and non-transcribed genomic
regions. The null hypothesis for strand symmetry was rejected, when >95% of the performed tests for a particular dataset rejected the simpler strand symmetric substitution model. The strand symmetry assumption of transition rates was rejected in the transcribed fraction and could not be rejected in the non-transcribed fraction (Figure 4.3). In contrast, for transversion rates the strand symmetry assumption could not be rejected in both the transcribed and non-transcribed fraction.

![Figure 4.3: Testing for strand symmetry of substitution rates using reversible substitution models. The null hypothesis for the strand symmetry between transitions ($\alpha = A \leftrightarrow G = T \leftrightarrow C$) and between transversions ($\beta = A \leftrightarrow C = T \leftrightarrow G$) was tested in the transcribed alignment fraction and non-transcribed alignment fraction. The fraction of tests that reject $M_0$ in the LRT are shown. In total, 1000 sampled alignments of the transcribed fraction and 1000 sampled alignments of the non-transcribed fraction (Flanking) were used for the tests. The strand-symmetry for a particular dataset was rejected when at least 95% of the corresponding alignments rejected the null-hypothesis. In the transcribed fraction the transition strand symmetry assumption was rejected in all tested alignments, whereas the transversion strand symmetry assumption was rejected in 11% of the tested alignments. The non-transcribed fraction the strand symmetry assumption for transitions and transversions was rejected in \(\sim 5\%\) of the performed tests.](image-url)
4.1.3 Likelihood-Ratio-Test using non-reversible substitution models

When using reversible models it is not possible to infer, whether the rejection of the strand symmetry assumption between the transitions \( A \rightarrow G \) and \( T \rightarrow C \), \( G \rightarrow A \) and \( C \rightarrow T \) or both contributed to the preference for the GTR model. Therefore, a second analysis was performed using an unrestricted non-reversible substitution model.

In total, 10 strand symmetric models were defined for the strand-symmetry of each of the six complementary substitution rate pairs, and for the 4 remaining substitution rates that are assumed reversible in the GTR model (Figure 4.4). The alternative model was the unrestricted model (GTR) with 11 free rate parameters (Figure 4.4). The log likelihoods of the substitution models were estimated using \textit{paml} [Yang, 2007]. The individual substitution models \( M_0 \) differ by having one parameter less than the alternative model \( M_A \). The statistic for the Likelihood-Ratio-Test follows therefore approximately a \( \chi^2 \) distribution with one degree of freedom. Again, a confidence limit of 5% was chosen for the tests. The fraction of Likelihood-Ratio-Tests rejecting the null hypothesis of strand symmetry was computed and compared between the sampled bootstrap alignments corresponding to the transcribed and non-transcribed regions in the human genome. In addition, the transcribed fraction was divided in subsets corresponding to exons and introns, a subset where CpG related substitutions were excluded, and the fraction corresponding to the 4 categories of interspersed repeats. The null hypothesis for strand symmetry was rejected, when \( >95\% \) of the performed tests for a particular subset rejected the simpler strand symmetric substitution model. The strand symmetry for the \( A \rightarrow G \) and \( T \rightarrow C \) transitions was rejected in all subsets, except for the non-transcribed and exon alignment fraction. In contrast, the strand symmetry for the complementary pair \( G \rightarrow A \) and \( C \rightarrow T \) was not rejected in any subset (Figure 4.1). The other performed LRTs for symmetry did not discriminate between the transcribed and non-transcribed alignment fraction.

Figure 4.4 (following page): Non-reversible substitution models used for the LRT. A Likelihood-Ratio-Test is used to estimate whether a null model that assumes strand symmetry between rates corresponding to complementary substitutions or assume reversibility between substitution rates describes the data better than an unrestricted model. The null models have 10 free parameters and the alternative unrestricted model has 11 free parameters.
4.1 Testing the strand symmetry in transcribed genomic regions

Nullmodels: Transition-Strand-Symmetry 10 free parameters

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \beta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^3 & \eta = G \rightarrow A = C \rightarrow T
\end{align*}
\]

Nullmodels: Transversion-Strand-Symmetry 10 free parameters

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \beta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^4 & \gamma = T \rightarrow A = A \rightarrow T
\end{align*}
\]

Nullmodels: Reversibility 10 free parameters

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \theta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^5 & \theta = A \rightarrow G = G \rightarrow A
\end{align*}
\]

 Alternative Model 11 free parameters

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \beta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^6 & \beta = A \rightarrow G = T \rightarrow C
\end{align*}
\]

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \beta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^7 & \mu = A \rightarrow C = T \rightarrow G
\end{align*}
\]

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \beta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^8 & \epsilon = C \rightarrow G = G \rightarrow C
\end{align*}
\]

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \theta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^9 & \lambda = C \rightarrow T = T \rightarrow C
\end{align*}
\]

\[
\begin{align*}
\mathbf{A} & \begin{bmatrix} -\alpha & \theta & \gamma \\ \delta & -\epsilon & \eta \\ \theta & \vartheta & -\delta \\ \kappa & \lambda & \mu \end{bmatrix} \\
\mathbf{C} & \begin{bmatrix} \delta \\ \theta \\ \kappa \end{bmatrix} \\
\mathbf{G} & \begin{bmatrix} -\epsilon & \eta \\ \vartheta & \mu \end{bmatrix} \\
\mathbf{T} & \begin{bmatrix} -\beta & \gamma \\ \eta & \delta \\ \lambda \end{bmatrix} \\
\mathbf{M}_0^{10} & \alpha = A \rightarrow C = C \rightarrow A
\end{align*}
\]
### Table 4.1: Testing for strand symmetry of substitution rates using non-reversible substitution models.

The null hypothesis for the strand-symmetry of rates between strand complementary substitutions and for reversibility assumption of substitutions was tested in the transcribed and non-transcribed alignment fraction. A variety of additional datasets have been extracted from the transcribed dataset such as an intron and exon subset, a subset where CpG related substitutions were excluded and a subset corresponding to the 4 categories of interspersed repeats (LINE, SINE, LTR and DNA transposons). The fraction of tests rejecting the strand symmetric model between complementary substitution rates and for reversibility of substitution rates are shown. In total, the null-hypothesis was tested in 1000 alignments 100 Kb in length that were sampled from the individual datasets. The strand symmetry for a dataset was rejected when at least 95% of the corresponding alignments rejected the null-hypothesis (datasets exceeding 95% shown in bold).

<table>
<thead>
<tr>
<th></th>
<th>Transcribed</th>
<th>Flanking</th>
<th>Intron</th>
<th>Exon</th>
<th>non-CpG</th>
<th>LINE</th>
<th>SINE</th>
<th>LTR</th>
<th>DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>transition sym.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M^1_{\eta}$  $G \rightarrow A = C \rightarrow T$</td>
<td>0.604</td>
<td>0.050</td>
<td>0.644</td>
<td>0.936</td>
<td>0.682</td>
<td>0.814</td>
<td>0.513</td>
<td>0.760</td>
<td>0.932</td>
</tr>
<tr>
<td>$M^2_{\beta}$ $A \rightarrow G = T \rightarrow C$</td>
<td><strong>1.000</strong></td>
<td>0.103</td>
<td><strong>1.000</strong></td>
<td>0.932</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.991</strong></td>
<td><strong>0.998</strong></td>
<td><strong>1.000</strong></td>
</tr>
<tr>
<td><strong>transversion sym.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M^3_{\delta}$ $C \rightarrow A = G \rightarrow T$</td>
<td>0.284</td>
<td>0.123</td>
<td>0.313</td>
<td>0.555</td>
<td>0.238</td>
<td>0.298</td>
<td>0.488</td>
<td>0.209</td>
<td>0.345</td>
</tr>
<tr>
<td>$M^4_{\gamma}$ $A \rightarrow T = T \rightarrow A$</td>
<td>0.187</td>
<td>0.086</td>
<td>0.193</td>
<td><strong>0.956</strong></td>
<td>0.110</td>
<td>0.207</td>
<td>0.196</td>
<td>0.111</td>
<td>0.187</td>
</tr>
<tr>
<td>$M^5_{\mu}$   $A \rightarrow C = T \rightarrow G$</td>
<td>0.073</td>
<td>0.088</td>
<td>0.102</td>
<td>0.724</td>
<td>0.061</td>
<td>0.160</td>
<td>0.362</td>
<td>0.091</td>
<td>0.184</td>
</tr>
<tr>
<td>$M^6_{\varepsilon}$ $C \rightarrow G = G \rightarrow C$</td>
<td>0.217</td>
<td>0.109</td>
<td>0.238</td>
<td><strong>0.958</strong></td>
<td>0.072</td>
<td>0.207</td>
<td>0.172</td>
<td>0.192</td>
<td>0.297</td>
</tr>
<tr>
<td><strong>reversible sym.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M^7_{\theta}$ $A \rightarrow G = G \rightarrow A$</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td>0.910</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td>0.934</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
</tr>
<tr>
<td>$M^8_{\lambda}$ $C \rightarrow T = T \rightarrow C$</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td>0.919</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.994</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
</tr>
<tr>
<td>$M^9_{\alpha}$ $A \rightarrow C = C \rightarrow A$</td>
<td>0.034</td>
<td>0.735</td>
<td>0.052</td>
<td>0.802</td>
<td>0.059</td>
<td>0.226</td>
<td>0.677</td>
<td>0.045</td>
<td>0.154</td>
</tr>
<tr>
<td>$M^{10}_{\iota}$ $T \rightarrow G = G \rightarrow T$</td>
<td>0.191</td>
<td>0.711</td>
<td>0.292</td>
<td>0.783</td>
<td>0.129</td>
<td>0.746</td>
<td>0.159</td>
<td>0.155</td>
<td>0.519</td>
</tr>
</tbody>
</table>
4.1.4 The direction of the strand-specific substitution rates

A more detailed view on the substitution strand symmetry is obtained by comparing the ratios of complementary substitution rates estimated from the transcribed and non-transcribed alignment fractions (Figure 4.5). The substitution rate parameters were estimated from the unrestricted non-reversible substitution model using paml \(^{1994a}\). In the transcribed fraction the ratio between complementary transitions \(A \rightarrow G\) and \(T \rightarrow C\) was in average 1.26. The ratio between other strand complementary substitution rates showed only an average ratio of 1.1. An exception are the transversions \(A \rightarrow C\) and \(T \rightarrow G\) that showed strand symmetry in the transcribed and non-transcribed dataset. In the non-transcribed alignment fraction all performed comparisons showed a strand symmetry.

The transition rates \(G \rightarrow A\) and \(C \rightarrow T\) have the highest rates followed by the \(A \rightarrow G\) and \(T \rightarrow C\) transitions. Transversion rates were about 4 fold lower than transition rates. The estimated substitution rates were then compared between the transcribed and non-transcribed fraction. Two transition rates \((G \rightarrow A, A \rightarrow G)\) and three transversion rates \((C \rightarrow G, A \rightarrow C, T \rightarrow G)\) are significantly increased in the transcribed fraction (t-test, \(p \leq 0.05\)). The remaining 7 substitution rates showed a significant decreased rate in the transcribed fraction (t-test, \(p \leq 0.05\)). The distributions of the rates for the individual substitution types are shown in Figure 4.6 A. To judge the extent of the substitution rate differences between the transcribed and non-transcribed fraction the difference of the mean substitution rates estimated from both alignment fractions are shown (Figure 4.6 B). Although a significant difference was observed between all substitution rates of the transcribed and non-transcribed alignment fraction, a large difference is more meaningful than a small or a slight change. The \(A \rightarrow G\) transition is remarkably increased in the transcribed fraction, whereas other substitution rates are

---

**Figure 4.5 (following page):** Ratio between strand complementary substitution rates. The individual distributions and mean of the substitution rate ratios between strand complementary substitution rates that were estimated from 1000 bootstrap sampled alignments in the transcribed and non-transcribed fraction using an unrestricted non-reversible substitution model are shown. In the transcribed fraction the ratio of the transitions \(A \rightarrow G\) and \(T \rightarrow C\) was in average 1.26 and largest in comparison to the other ratios with an average of 1.1 in the transcribed fraction. The transversions \(A \rightarrow C\) and \(T \rightarrow G\) showed an average ratio of 1 in the transcribed fraction. All ratios from strand complementary substitution rates estimated from the non-transcribed fraction had also an average ratio of 1.
4.1 Testing the strand symmetry in transcribed genomic regions

![Graphs showing comparisons between transcribed and non-transcribed genomic regions.](image)
4.2 Discussion

Transcription-coupled repair has been suggested to specifically increase the rate of $A \rightarrow G$ transitions relative to the rate of the complementary change, $T \rightarrow C$ (Green, 2003). Yet, a genome-wide statistical analysis of this proposed transition asymmetry in transcribed regions has not been confirmed. Here, I have separated the human genome into two fractions. The first part consists of genes for which evidence exist that they are transcribed in the human germ line. The second part, corresponding to the bona-fide non-transcribed fraction of the human genome, is comprised by genomic regions located outside of known genes.

A Likelihood-Ratio-Test (LRT) revealed that in the transcribed dataset a more complex non-reversible substitution model (11 parameters) gives a significantly better fit to the data than the simpler model in which the same rate was assigned to the complementary $A \rightarrow G$ and $T \rightarrow C$ transitions (10 parameters). Notably, the strand symmetric null models for the 5 remaining pairs of complementary substitution pairs could not be rejected. For the non-transcribed alignment fraction none of the strand symmetric models could be rejected. Thus, I provide evidence that an increased rate of $A \rightarrow G$ over $T \rightarrow C$ substitutions is specific to transcribed sequences in the human genome. It may therefore be indeed an effect of transcription-coupled repair.

To analyze the impact of other components of the transcribed fraction to the results

Figure 4.6 (following page):  A) Distribution of substitution rates estimated from the transcribed and non-transcribed dataset. Shown are the substitution rates estimated using the unrestricted substitution model (Yang, 1994a) from 1000 alignments sampled from the transcribed (red) and non-transcribed (blue) dataset. The distribution of the $A \rightarrow G$ transition rates shows the most prominent increase in the transcribed dataset compared to the non-transcribed dataset. Strand complementary substitution rates are placed next to each other in the diagram. The median of transitions are about 4-fold higher than the median of transversions. The $C \rightarrow T$ and $G \rightarrow A$ transition rates have the highest median rate followed by the $A \rightarrow G$ and $T \rightarrow C$ transition rates. B) Differences between the average substitution rates estimated from 1000 sampled alignments of the transcribed and non-transcribed dataset. Whereas the average $A \rightarrow G$ rate is prominently increased in the transcribed dataset the average of the other substitution rates are decreased or only slightly changed.
Flanking Substitution rate

A) Transcribed
B) Non Transcribed

Transversions

Transitions

Substitution rate

0.2 0.4 0.6 0.8

0.2 0.4 0.6 0.8

Discussion
obtained from the LRTs, the analysis was additionally performed in the protein-coding exons, in introns, a fraction where substitutions related CpG dinucleotides were excluded, and four fractions corresponding to interspersed repeats. Except for the exon fraction all other alignment fractions corresponding to the transcribed fraction rejected the simpler model for $A \rightarrow G = T \rightarrow C$. However, notably for the exon fraction is, that almost all symmetries introduced into the substitution models lead to LRT results that are at the border of significance (c.f. Table 4.1). This may be attributed to the particular mode of evolution in protein coding exons. The information is encoded in base triplets making the assumption of the neighbor independent substitution process inherent in the models used in this analysis inappropriate. In addition, the substitution rates are highly decreased in exonic sequences due to selective constraints. However, exons make up only for a small part of the transcribed fraction and have no influence on the results for the entire transcribed fraction. To examine a putative impact of CpG related transitions to the observed rejection of the strand symmetry an additional dataset was examined, where CpG related substitutions were excluded. However, as for the entire transcribed alignment fraction the simpler model $A \rightarrow G = T \rightarrow C$ was rejected. This indicates that an increased $A \rightarrow G$ rate is not an effect of CpG-related substitutions.

The substitution model parameters were estimated from sampled alignments of the concatenated alignment fractions for the different datasets, where the estimates represent a global average. It is therefore not expected that results change for other gene subsets that define the transcribed alignment fraction. The study shows that the observed increase of the $A \rightarrow G$ substitution rate in the transcribed fraction of the human genome has a significant impact to distinguish the substitution model between the transcribed and non-transcribed alignment fraction.
Chapter 5

Species-Specific evolving regions in the human and chimpanzee genomes

In the previous chapter it was shown that the substitution rates $A \rightarrow G \neq T \rightarrow C$ in transcribed genomic regions. This can be explained by transcription-coupled-repair (Green, 2003). The extent of the $A \rightarrow G$ signature for a gene could be therefore used as an indicator for the transcription activity in the genome of a particular species. Thus, genes that are differentially expressed in humans and chimpanzees are subject to different extents of transcription coupled repair, and thus should differ in their rates and patterns of DNA sequence evolution. In this chapter, the homogeneity of the substitution model between human and chimpanzee is tested to identify transcribed genomic regions with species-specific substitution patterns.

A substitution model provides a stochastic framework to model the substitution process between DNA sequences in an alignment. As introduced in chapter 3, a markov process is used for this purpose that assumes homogeneity, stationarity and reversibility. The reversibility of the process assumes that a substitution and the corresponding reverse substitution are not distinguished. When the homogeneity of the substitution model is assumed the substitution patterns are the same and independent from time for all branches in the phylogenetic tree. In such a homogeneous substitution model an equilibrium nucleotide composition exists, and is the same for all taxa considered, to which the sequences evolve and approximate in time infinity. The stationarity of the substitution model assumes that the sequences have reached already the equilibrium nucleotide composition which will then remain unchanged during the process. However,
these assumptions are commonly applied simplifications to model the substitution process and are not always valid when biological sequence data are analyzed. Species that do not evolve alike evolve towards a different equilibrium nucleotide composition and can therefore vary extensively in the genomes GC content. For example the genome of the protozoan parasite *Plasmodium falciparum* has a very low GC content of \( \sim 20\% \) \cite{Gardner2002} in comparison to *Plasmodium vivax* that has a GC content of \( \sim 45\% \) \cite{Carlton2001}. However, when the homogeneity assumption is not valid a more complex substitution model may be more appropriate to describe the evolutionary process. Such a more complex ”non-homogeneous substitution model” describes the evolutionary process individually for each branch in the tree \cite{Barry1987}.

Different strategies are introduced for statistical comparisons of the substitution patterns between species. A simple approach was introduced by the disparity index \cite{Kumar2001} that measures the observed compositional differences between a pair of sequences to infer whether the homogeneity of the substitution patterns do not differ more than expected from the divergence from the sequences. However, for comparisons between sequences that have the same nucleotide composition such as, e.g., human and chimpanzee, such a test is not appropriate. Therefore, a more sensitive method that estimates the substitution process between the sequences in an alignment is required. A parametric bootstrap approach to test the homogeneity assumption of the substitution model in a phylogenetic tree has been described by the Model-Homogeneity-Test (MHT) in \cite{Weiss2003}. In the MHT, the statistic \( \Delta \) measures the deviations between the substitution models that are estimated from sequence pairs in an multiple sequence alignment. From all pairwise substitution rate matrices a covariance matrix is computed and decomposed. The test statistic is then defined by the sum of eigenvalues of the decomposed covariance matrix. It is necessary to determine whether the differences between the pairwise estimated substitution models, captured in \( \Delta \) are significant or whether they arise due to the stochastic nature of the substitution process and the finite amount of data. Even if sequences evolve according to the same substitution model, their substitution patterns can vary by chance such that the estimated pairwise substitution models differ from each other. The Model-Homogeneity-Test uses a simulation procedure to infer the significance of \( \Delta \) to reject the homogeneity of the substitution model. For this purpose, a model for the null-hypothesis that the substitution process
is homogeneous over the entire tree is required. To arrive at such a model, the average of the pairwise substitution rate matrices is computed and used as a null-model. The variation of $\Delta$ that is expected by chance is inferred from sequence alignments that are generated under the homogeneous null-model. The procedure generates 1000 simulated alignments to compute the distribution of $\Delta_{\text{sim}}$. The error probability when rejecting the homogeneity of the substitution model in the tree is then assumed from the fraction of $\Delta_{\text{sim}}$ values $\geq \Delta_{\text{obs}}$.

The Model-Homogeneity-Test has been designed to test the homogeneity of the substitution model for an entire phylogenetic tree. For a comparison of the substitution processes between individual branches more complex non-homogeneous substitution models are required. A non-homogeneous model allows to distinguish the substitution process between individual species. An extension to the MHT for testing the homogeneity of the substitution model between individual branches in a tree, called the Single-Branch-Test (SBT), was developed by Gunter Weiss and Arndt von Haeseler (unpublished). The substitution rate matrices are estimated for each branch in the tree from pairwise sequence comparisons in an alignment. The statistic is defined as a contrast between a set of two or more branches, where the variations between the models are measured as squared distance between the corresponding substitution rate matrices. The null-distribution of the defined contrast is again inferred by parametric bootstrap, where sequence alignments are simulated under the homogeneity assumption. The Single-Branch-Test is described in detail in the next section.

5.1 Estimation of branch-specific substitution models

The transition probability matrix $P(t)$ describes the substitution process of a lineage from the ancestral state to the observed present-day sequence. As the transition probabilities are not directly observable from sequence data, it is shown in the following how to estimate $P(t)$ from pairwise sequence comparisons. In a pairwise sequence comparison the time units $t$ leading to the sequence $X_1$ and $X_2$ from the ancestral sequence $X_0$ is unknown: Let $h_{(i,j)}^{(r,s)}$ denote the frequency of having nucleotides $i$ and $j$ ($i,j \in \{A,C,G,T\}$) in sequences $r$ and $s$, respectively. These probabilities are gathered in the so-called divergence matrix $H^{(r,s)}$.

The definition of $Q$ allows to estimate the evolutionary distance and the substitution
model from the enumerated substitutions in a pair of sequences when a reversible process is assumed (Baake, 1998).

Assuming stationarity and reversibility of the substitution process the divergence matrix can be expressed as

\[ H^{(r,s)} = \text{diag}(\pi_A, \pi_C, \pi_G, \pi_T) \cdot P^{(r,s)}. \]  \hspace{1cm} (5.1)

The transition probability matrix for the path from sequence \( x \) to \( y \) is then given by

\[ P^{(r,s)} = \text{diag}\left(\frac{1}{\pi_A}, \frac{1}{\pi_C}, \frac{1}{\pi_G}, \frac{1}{\pi_T}\right) \cdot H^{(r,s)}. \]  \hspace{1cm} (5.2)

From the pairwise leaf-to-leaf transition matrices branch specific transition matrices can be estimated (Baake, 1998) by the following: Let \( i, j, k \) denote leaves in an unrooted three taxon tree and \( m \) the internal node (Figure 5.1). The return-trip transition matrix \( M^{(i,m)} \) for an external branch connecting leaf \( i \) and node \( m \) can be expressed as:

\[ M^{(i,m)} = P^{(i,m)} \cdot P^{(m,i)} = P^{(i,j)} \cdot (P^{(k,j)})^{-1} \cdot P^{(k,i)}. \]  \hspace{1cm} (5.3)

The equation shown in (5.3) requires only leaf-to-leaf transition matrices. Thus, the return-trip matrices \( M^{(i,m)} \) can be expressed in terms of divergence matrices (Eq. 5.3). The transition probability matrix \( M^{(i,m)} \) is defined as exponential of the instantaneous rate matrix \( Q^{(i,m)} \):

\[ M^{(i,m)} = \exp(Q^{(i,m)}(t)). \]  \hspace{1cm} (5.4)

Assume \( M^{(i,m)} \) has a spectral decomposition (Bailey et al., 1990):

\[ M^{(i,m)} = U \cdot \text{diag}(e^{\lambda_{1,i,m}}, e^{\lambda_{2,i,m}}, e^{\lambda_{3,i,m}}, e^{\lambda_{4,i,m}}) \cdot U^{-1} \]  \hspace{1cm} (5.5)
5.1 Estimation of branch-specific substitution models

The matrix $M(i,m)$ is decomposed to the left eigenvectors in $U$, the eigenvalues defined in $\lambda$ and right eigenvectors. The instantaneous rate matrix $Q(i,m)$ is computed from the logarithm of the eigenvalues, where $e^{\lambda_j t_{i,m}}$ is replaced by $\lambda_j t_{i,m}$.

$$Q(i,m) = U \cdot \text{diag}(\lambda_{1t_{i,m}}, \lambda_{2t_{i,m}}, \lambda_{3t_{i,m}}, \lambda_{4t_{i,m}}) \cdot U^{-1}$$  \hspace{1cm} (5.6)

For a given alignment the substitution rate matrix $Q$ is estimated for each branch in the unrooted tree of human, chimpanzee and rhesus. The test for the homogeneity of the substitution patterns are performed by a contrast of the human and chimpanzee branch. The variations between the substitution models are quantified by a statistic that measures the squared distance $\delta$ of the corresponding substitution rate matrices:

$$\delta = \sum_{i=1}^{12} \left( q_{i\text{human}} - q_{i\text{chimp}} \right)^2. \hspace{1cm} (5.7)$$

The vectors $q_{\text{human}}$ and $q_{\text{chimp}}$ describe the off-diagonal elements of the individual rate matrices $Q$ of the human and chimpanzee branch in the tree. For the null-hypothesis that the substitution process is homogeneous over the entire tree an average substitution rate matrix is computed from the substitution rate matrices of the branches and used as null-model. The null model $q_{\text{null}}$ is defined by averaging the substitution rates in the

Figure 5.1: Return-trip transition matrices reconstructed from pairwise leaf-to-leaf transition matrices [Baake 1998]
vectors $q^i$ from all branches $i$ in the three taxon tree:

$$q_{null} = \frac{1}{n} \sum_{i=1}^{n} q^{(i)}$$  \hspace{1cm} (5.8)

The probability to reject the homogeneity assumption over the defined branches is estimated from the fraction of $\delta_{sim}$ values $\geq \delta_{obs}$. An overview of the procedure is shown in Figure 5.2.

**Figure 5.2:** Testing model homogeneity between human and chimpanzee. The substitution rate matrices are estimated for the branches in the human, chimpanzee and rhesus tree. The difference between the human and chimpanzee substitution matrices are inferred by a squared distance. The homogeneous model for the tree that is used for the simulation procedure is the average substitution rate matrix of all branches. Sequence alignments are simulated under the homogeneous model to obtain the null-distribution of $\delta$. The critical value of $\delta$ to reject the homogeneity assumption between the human and chimpanzee substitution model is determined from the simulated null-distribution.
5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees

A sliding window approach was used to screen the human genome for regions that evolve in a species-specific manner. The individual alignment window datasets that were used for the study are described in detail in chapter 2 (Table 2.1). The SBT was applied to each alignment window to identify genomic regions for which the human substitution model differs significantly from the chimpanzee substitution model. To assess the influence of the window length on the outcome of the test, different window sizes of 125, 250, 500, and 1000 Kb were used. In addition, the transcribed fraction of the 125 Kb windows as well as the transcribed fraction of individual genes was analyzed separately. The confidence limit for the SBT was set to \( p \leq 0.05 \). An alignment window for which the homogeneity assumption between human and chimpanzee substitution process was rejected is referred to as significant alignment window in the following. The results of the analysis are summarized in table 5.1.

The impact of the alignment length on the power to detect a significant difference between the models was verified by comparing the fraction of significant tests of the alignments between the different alignment window sizes. The fraction of significant alignment windows is slightly increased in the bigger window sizes with 5.69% tests in the 125 Kb to 8.1% tests of the 1000 Kb alignment windows (Table 5.1). This indicates that the power of the test increases with the sample size.

The signal to detect a difference between the human and chimpanzee substitution models is expected to be weak due to the low number of substitutional sequence differences between both species. As only a small fraction of significant alignment windows were detected (c.f. Table 5.1), the distribution of the p-values obtained from the genome-wide analysis are expected to follow almost an uniform distribution. The power of the SBT to detect a difference between the human and chimpanzee substitution model is examined by a probability-probability plot. The probability-probability plot allows to compare the cumulative p-value distribution to a cumulative uniform distribution (Figure 5.3). For the different datasets only a small deviation to an uniform distribution is observed that slightly increases for bigger window sizes.
5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total windows tested</th>
<th>Significant</th>
<th>Fraction (significant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Kb</td>
<td>2176</td>
<td>177</td>
<td>8.13%</td>
</tr>
<tr>
<td>500 Kb</td>
<td>4321</td>
<td>287</td>
<td>6.64%</td>
</tr>
<tr>
<td>250 Kb</td>
<td>8608</td>
<td>544</td>
<td>6.32%</td>
</tr>
<tr>
<td>125 Kb</td>
<td>17138</td>
<td>974</td>
<td>5.68%</td>
</tr>
<tr>
<td>TRF_{125Kb}</td>
<td>12596</td>
<td>717</td>
<td>5.69%</td>
</tr>
<tr>
<td>TRF_{Transcript}</td>
<td>7292</td>
<td>412</td>
<td>5.65%</td>
</tr>
</tbody>
</table>

Table 5.1: Results of the SBT between human and chimpanzee. Overview of the results obtained for the alignments of the 125 Kb to 1000 Kb alignment window (Dataset) and the transcribed fraction of the 125 Kb windows (TRF_{125Kb}) and the transcribed fraction of individual genes (TRF_{Transcript}). Shown is the total number of alignment windows as well as the number and the percentage of significant alignment windows.

When multiple tests are performed at a significance level of 5% one accepts to falsely reject the null hypothesis of model homogeneity in 5% of all performed tests. In my analysis, about 5% of the performed tests reject the null hypothesis. Hence, it is necessary to show that the substitution models do not differ solely by chance due to the multiple tests that have been performed. A detailed analysis of the substitution patterns is conducted to describe the differences between the human and chimpanzee model in the significant alignment windows.
5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees

Figure 5.3: Probability-Probability plot of the SBT nominal p-values. The plot shows the percentages of tests falling at or below a given p-value from the SBT in 125 KB to 1000 KB alignment windows, and the transcribed fraction datasets. The diagonal (black) illustrates a distribution of samples from a uniform distribution. The vertical line shows the 5% confidence limit. Only a small deviation of the p-value distributions to a uniform distribution is observed.
5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees

5.2.1 Normalization of $\delta$ estimated from different alignment windows

The number of analyzable ungapped alignment positions and the amount of divergence between the sequences differ among the alignment windows. With an increasing size of an alignment or increasing sequence divergence the distance measure of $\delta$ decreases due to the higher accuracy and precision of the estimates. How close the estimated $\delta$ is to the true unknown value is dependent on the accuracy and the variation of $\delta$ around the true value, i.e., on the precision of the estimate $\delta$. For the comparison of the $\delta$ statistic estimated among different alignment windows a normalization procedure is therefore required that is discussed in detail in chapter 6. The estimated $\delta$ is then scaled to a sample size of 50,000 analyzable alignment columns and a pairwise sequence divergence of 1%. For each alignment window the branch length of the unrooted three species tree is inferred using the neighbor-joining algorithm \cite{Saitou1987}.

The scaling factor for the estimates $\delta$ relative to a branch length of 0.005 expected substitutions per site for the human and chimpanzee branch as was determined from a calibration curve (chapter 6, Figure 6.6). For a variety of branch lengths, the mean of $\delta$ was computed from 1000 simulated alignments of 50 Kb in length, generated with the same substitution model for all branches in the tree. The model was defined by a random substitution rate matrix explained in detail in chapter 6. $\delta$ was measured between two branches with branch lengths between 0.001 and 0.2 expected substitutions per site. The calibration curve is defined by an interpolation function $F$ that describes $\delta$ as function of the branch length. The function allows to determine the expected $\delta$ for any given branch length $b$. The scaling factor for $\delta$ estimate is then determined by the ratio of $F_{0.005}$ and $F_b$, where $b$ denotes branch length estimated for human and chimpanzee from the given alignment window. The normalization of $\delta$ considering the alignment size and branch length is computed by:

$$\delta_{\text{norm}} = \delta \cdot F_{0.005}/F_{b} \cdot L_s/L_{50Kb} \tag{5.9}$$

$L_s$ denotes the number of analyzable alignment columns of an alignment window and $L_{50Kb}$ denotes the reference alignment size of 50 Kb. The effect of the normalization procedure is shown in Figure 5.4.
In the following, it was investigated whether the fractions of significant alignment windows or the distributions of $\delta$ differ among the chromosomes. The enrichment of significant alignment windows for each chromosome was inferred using the Fisher exact test (Table 5.3). The number of significant alignment windows are enriched on chromosome X (42/491, 8.5%), chromosome Y (6/31, 19.35%) and chromosome 6 (52/712, 7.3%). Multiple testing was considered by estimating the false discovery rate (FDR, Benjamini and Y. (1995)). The FDR denotes the expected fraction of false positive tests among the significant alignment windows. When multiple tests are considered chromosome X and Y remain significant for a FDR < 0.1. The distribution of $\delta$ estimated from the transcribed fraction of the 125 Kb alignment windows were compared between the chromosomes by an one factor ANOVA analysis. This shows that the distributions of $\delta$ are significantly different between the chromosomes (ANOVA, p≤2.2e-16). Figure 5.5

**Figure 5.4:** The distribution of $\delta$ obtained from the alignment datasets of 1000 Kb, 500 Kb, 250 Kb, 125 Kb windows, the transcribed fraction of the 125 Kb windows (TRF 125) and the transcribed fraction of individual genes (Ensembl) are shown before (red) and after normalization (blue). $\delta$ is normalized relative to a pairwise sequence divergence of 1% between human and chimpanzee and an alignment size of 50 Kb.

### 5.2.2 Overview of the significant alignment windows

In the following, it was investigated whether the fractions of significant alignment windows or the distributions of $\delta$ differ among the chromosomes. The enrichment of significant alignment windows for each chromosome was inferred using the Fisher exact test (Table 5.3). The number of significant alignment windows are enriched on chromosome X (42/491, 8.5%), chromosome Y (6/31, 19.35%) and chromosome 6 (52/712, 7.3%). Multiple testing was considered by estimating the false discovery rate (FDR, Benjamini and Y. (1995)). The FDR denotes the expected fraction of false positive tests among the significant alignment windows. When multiple tests are considered chromosome X and Y remain significant for a FDR < 0.1. The distribution of $\delta$ estimated from the transcribed fraction of the 125 Kb alignment windows were compared between the chromosomes by an one factor ANOVA analysis. This shows that the distributions of $\delta$ are significantly different between the chromosomes (ANOVA, p≤2.2e-16). Figure 5.5
shows the corresponding distributions of $\delta$ for each chromosome. Note, chromosome Y has a remarkably increased distribution of $\delta$ compared to the other chromosomes. The ANOVA analysis remains significant when chromosome Y is excluded.

So far, I have searched for alignment windows which had a significant difference between the substitution models of human and chimpanzee. Intuitively, an alignment with a large $\delta$ suggests that the differences between the human and chimpanzee substitution models also have a biological significance. Subsequently I selected those windows for further analysis that show the strongest difference among all tested alignment windows. For further analysis the alignment windows with a a significant p-value and with a $\delta$ in the 5% largest $\delta$ values were selected (Figure 5.5). The selected alignment windows comprised a total of $\sim 4.5\%$ for the different datasets (Table 5.2).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total</th>
<th>$p \leq 0.05; \delta &gt; \delta_{95%}$</th>
<th>Percentage Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Kb</td>
<td>2176</td>
<td>107</td>
<td>4.92 %</td>
</tr>
<tr>
<td>500 Kb</td>
<td>4321</td>
<td>205</td>
<td>4.74 %</td>
</tr>
<tr>
<td>250 Kb</td>
<td>8608</td>
<td>395</td>
<td>4.59 %</td>
</tr>
<tr>
<td>125 Kb</td>
<td>17138</td>
<td>771</td>
<td>4.5 %</td>
</tr>
<tr>
<td>TRF$_{125Kb}$</td>
<td>12596</td>
<td>565</td>
<td>4.49 %</td>
</tr>
<tr>
<td>TRF$_{Transcript}$</td>
<td>7292</td>
<td>328</td>
<td>4.5 %</td>
</tr>
</tbody>
</table>

**Table 5.2:** Significant alignment windows with $\delta > \delta_{95\%}$ from all tests after normalization. Alignment windows were selected with strongest difference between the estimated human and chimpanzee substitution model.

### 5.2.3 Characterization of the substitution patterns in significant alignment windows

In the next step, the aim was to identify which type of substitution contributes significantly to the rejection of the substitution model homogeneity assumption between human and chimpanzee. The $\delta$ statistic allows only to measure the overall differences between the human and chimpanzee models by the squared distance between the corresponding rate matrices. Therefore, the Single-Branch-Test was extended by defining a statistic to measure the differences between each component of the human and chimpanzee substitution model.

The individual substitution rates were tested whether they differ significantly between
Figure 5.5: Human and chimpanzee model differences for the 24 human chromosomes. The plot shows the distributions of $\delta$ estimated from the transcribed fraction of the 125 Kb alignment windows ($TREF_{125Kb}$). The horizontal line denotes the genome-wide 95% quantile of the $\delta$ distribution estimated from all alignment windows. Significant alignment windows with a $\delta$ above the horizontal line and a p-value $\leq 0.05$ are selected for further analysis.
the human and chimpanzee model for all 125 Kb alignment windows in the transcribed fraction of the human genome \((TRF_{125\text{Kb}})\). The statistic is defined as a squared difference for each of the 12 substitution rates \(i\) between \(q^{(\text{human})}\) and \(q^{(\text{chimpanzee})}\).

\[
\delta_i = (q_i^{(\text{human})} - q_i^{(\text{chimpanzee})})^2
\]  

(5.10)

The null-distribution of each \(\delta_i\) for each substitution type is inferred from the 1000 simulated alignments that were generated under a homogeneous model.

The number of how often the rate equality for each substitution type was rejected between the human and chimpanzee model was counted and compared between the set of significant alignment windows \((SBT_p \leq 0.05; \delta > \delta_{95\%})\) and the non-significant alignment windows of the \(TRF_{125\text{Kb}}\) alignment dataset (Figure 5.6). As the human and chimpanzee substitution model underlying this analysis are reversible the counts of tests that rejected equality for the individual substitution types are averaged for the six reversible substitution rate pairs.

The observed model differences between the human and chimpanzee substitution models are mainly driven by a difference between the transition rates. The significant alignment windows rejected the substitution rate equality between the human and chimpanzee model for the transitions \(A \leftrightarrow G\) in >50\%, for the transitions \(T \leftrightarrow C\) in >50\% and for each of the transversions \((A \leftrightarrow C\), \(T \leftrightarrow G\), \(G \leftrightarrow C\), \(T \leftrightarrow A\)) in <20\%. In the non-significant alignment windows, only less than \(\leq 5\%\) rejected the substitution rate equality for the different substitution types.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Percentage (%)</th>
<th>Significant</th>
<th>Expected</th>
<th>Total</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.05</td>
<td>58</td>
<td>65.35</td>
<td>1148</td>
<td>0.8533</td>
<td>0.9986</td>
</tr>
<tr>
<td>2</td>
<td>4.27</td>
<td>44</td>
<td>58.63</td>
<td>1030</td>
<td>0.9859</td>
<td>0.9986</td>
</tr>
<tr>
<td>3</td>
<td>4.57</td>
<td>40</td>
<td>49.86</td>
<td>876</td>
<td>0.9450</td>
<td>0.9986</td>
</tr>
<tr>
<td>4</td>
<td>6.96</td>
<td>46</td>
<td>37.63</td>
<td>661</td>
<td>0.0899</td>
<td>0.4315</td>
</tr>
<tr>
<td>5</td>
<td>6.04</td>
<td>40</td>
<td>37.68</td>
<td>662</td>
<td>0.3693</td>
<td>0.8057</td>
</tr>
<tr>
<td>6</td>
<td><strong>7.30</strong></td>
<td><strong>52</strong></td>
<td><strong>40.53</strong></td>
<td><strong>712</strong></td>
<td><strong>0.0374</strong></td>
<td><strong>0.2992</strong></td>
</tr>
<tr>
<td>7</td>
<td>6.21</td>
<td>54</td>
<td>49.52</td>
<td>870</td>
<td>0.2691</td>
<td>0.8057</td>
</tr>
<tr>
<td>8</td>
<td>5.98</td>
<td>33</td>
<td>31.42</td>
<td>552</td>
<td>0.4103</td>
<td>0.8206</td>
</tr>
<tr>
<td>9</td>
<td>7.32</td>
<td>38</td>
<td>29.54</td>
<td>519</td>
<td>0.0657</td>
<td>0.3942</td>
</tr>
<tr>
<td>10</td>
<td>5.23</td>
<td>33</td>
<td>35.92</td>
<td>631</td>
<td>0.7216</td>
<td>0.9986</td>
</tr>
<tr>
<td>11</td>
<td>4.20</td>
<td>25</td>
<td>33.87</td>
<td>595</td>
<td>0.9601</td>
<td>0.9986</td>
</tr>
<tr>
<td>12</td>
<td>6.11</td>
<td>37</td>
<td>34.50</td>
<td>606</td>
<td>0.3516</td>
<td>0.8057</td>
</tr>
<tr>
<td>13</td>
<td>4.26</td>
<td>14</td>
<td>18.73</td>
<td>329</td>
<td>0.9007</td>
<td>0.9986</td>
</tr>
<tr>
<td>14</td>
<td>5.39</td>
<td>20</td>
<td>21.12</td>
<td>371</td>
<td>0.6332</td>
<td>0.9986</td>
</tr>
<tr>
<td>15</td>
<td>6.33</td>
<td>25</td>
<td>22.48</td>
<td>395</td>
<td>0.3195</td>
<td>0.8057</td>
</tr>
<tr>
<td>16</td>
<td>4.23</td>
<td>16</td>
<td>21.52</td>
<td>378</td>
<td>0.9175</td>
<td>0.9986</td>
</tr>
<tr>
<td>17</td>
<td>6.44</td>
<td>29</td>
<td>25.62</td>
<td>450</td>
<td>0.2690</td>
<td>0.8057</td>
</tr>
<tr>
<td>18</td>
<td>4.98</td>
<td>14</td>
<td>16.00</td>
<td>281</td>
<td>0.7350</td>
<td>0.9986</td>
</tr>
<tr>
<td>19</td>
<td>5.64</td>
<td>19</td>
<td>19.18</td>
<td>337</td>
<td>0.5517</td>
<td>0.9458</td>
</tr>
<tr>
<td>20</td>
<td>2.33</td>
<td>7</td>
<td>17.08</td>
<td>300</td>
<td>0.9986</td>
<td>0.9986</td>
</tr>
<tr>
<td>21</td>
<td>8.05</td>
<td>12</td>
<td>8.48</td>
<td>149</td>
<td>0.1421</td>
<td>0.5684</td>
</tr>
<tr>
<td>22</td>
<td>5.86</td>
<td>13</td>
<td>12.64</td>
<td>222</td>
<td>0.4992</td>
<td>0.9216</td>
</tr>
<tr>
<td>X</td>
<td><strong>8.55</strong></td>
<td><strong>42</strong></td>
<td><strong>27.95</strong></td>
<td><strong>491</strong></td>
<td><strong>0.0053</strong></td>
<td><strong>0.0864</strong></td>
</tr>
<tr>
<td>Y</td>
<td><strong>19.35</strong></td>
<td><strong>6</strong></td>
<td><strong>1.76</strong></td>
<td><strong>31</strong></td>
<td><strong>0.0072</strong></td>
<td><strong>0.0864</strong></td>
</tr>
</tbody>
</table>

**Table 5.3:** Enrichment analysis of significant tests for the 24 human chromosomes

For each chromosome the percentage and number of significant alignment windows, the number of expected significant alignment windows and the total number of alignment windows are shown. The enrichment of significant alignment windows on a chromosome was estimated by a one-sided Fisher Exact Test (p-value). Chromosome 6, X and Y show an enrichment for significant alignment windows (confidence limit for nominal p-value $p \leq 0.05$). To account for multiple testing, a p-value correction was performed using the FDR. The FDR estimates the expected false discovery rate and is < 0.1 for the X and Y chromosome.
5.2 Testing the homogeneity assumption of the substitution model between humans and chimpanzees

Figure 5.6: Substitution rates with a significant difference between the human and chimpanzee models. The Single-Branch-Test was extended to identify which type of substitution contributed significantly to the rejection of the substitution model homogeneity assumption between human and chimpanzee detected. The analysis was performed using the TRF_{125Kb} dataset. Shown are the fractions of rejected equality between the human and chimpanzee models for the individual substitution types. The results for the set of significant alignment windows is shown in red. The results for the set of non-significant alignment windows are shown in blue. Transition rates (A \leftrightarrow G or T \leftrightarrow C ) showed in > 50% and each of the transversion rates showed in < 20% a significant difference between the human and chimpanzee model. In non-significant alignment windows \leq5% rejected rate equality for all the individual substitution types between the human and chimpanzee model.
5.2.4 Species-specific transition patterns in the human and chimpanzee genomes

In the following the distributions of the substitution rate differences between the corresponding human and chimpanzee substitution models are analyzed in more detail. The analysis is again performed on the $TRF_{125Kb}$ dataset. As the underlying substitution models are reversible, the differences between the 12 off-diagonal substitution rates in the vector of $q^{(human)}$ and $q^{(chimpanzee)}$ were averaged for the six reversible substitution rate pairs.

In the significant alignment windows, the distributions of substitution rate differences are bimodal for the transition rates and unimodal for the transversion rates (Figure 5.7). This indicates that there is a strong tendency for a significant window to have a high transition rate in humans and a low transition rate in chimpanzees and vice versa. For the transversion rates no such trend is observed. The non-significant alignment windows show no remarkable difference of the transition rates and transversion rates between the human and chimpanzee model.

The observed differences in significant alignment windows are supposed to be driven by different extents of transcription-coupled repair. In significant alignment windows a difference of the extent of the strand-specific $A \rightarrow G$ increase in either the human or the chimpanzee substitution model is therefore expected. The strand-specific $A \leftrightarrow G$ increase is quantified as the $A \leftrightarrow G / T \leftrightarrow C$ ratio and is compared between the human and chimpanzee substitution model that was estimated from the transcribed fraction 125 Kb alignment windows. Moreover, the ratios between the human and chimpanzee substitution models in significant and non-significant alignment windows were compared. Indeed, the extent of the $A \leftrightarrow G / T \leftrightarrow C$ ratios are remarkably different between the human and chimpanzee model in the significant alignment windows (Figure 5.8). The $A \leftrightarrow G / T \leftrightarrow C$ ratios between both species show a bimodal distribution with two peaks at a ratio of $\sim1.4$ and $\sim1$ in either human or chimpanzee. This indicates that there is a strong tendency for a significant window to have an increased $A \leftrightarrow G / T \leftrightarrow C$ ratio in humans and a decreased $A \leftrightarrow G T \leftrightarrow C$ ratio in chimpanzees and vice versa. In the non-significant alignment windows the $A \leftrightarrow G / T \leftrightarrow C$ ratios between both species are unimodal and show no remarkable difference (Figure 5.8).
5.3 Discussion

Human and chimpanzee DNA sequences are believed to evolve according to the same substitution model. Exceptions to this assumption have been described ([Ebersberger and Meyer, 2005, e.g.]), but a genome-wide analysis of this aspect has not been performed. It was aimed to identify regions in the transcribed fraction of the human and chimpanzee genomes that evolve species-specific. To this end, the human, chimpanzee and rhesus branch specific substitution matrices ([Baake, 1998]) were estimated for 12,596 non-overlapping windows 125 Kb in size, representing the transcribed part in the human-chimpanzee-rhesus genome alignment. For each window a squared distance $\delta$ was computed between the two substitution rate matrices, $Q_{\text{human}}$ and $Q_{\text{chimpanzee}}$. A parametric bootstrap approach was used to simulate the null-distribution of $\delta$ under the homogeneity assumption, i.e., when the substitution model is the same on all branches. If the value of $\delta$ fell outside of the 95% quantile of the null-distribution, the homogeneity assumption was rejected for the corresponding window. In total, 717 windows were found evolving in a species-specific manner. A closer look on the individual types of substitutions in these windows revealed, that mainly differences in the transition rates were responsible for rejecting the substitution model homogeneity for both species. Notably, the extent of the strand-specific $A \leftrightarrow G$ over $T \leftrightarrow C$ transition bias is markedly different between human and chimpanzee in the significant regions. The strand-specific increase of $A \rightarrow G$ over $T \rightarrow C$ substitutions on the transcribed strand is interpreted as signature of transcription coupled repair ([Green, 2003]). The results suggest that species-specific differences in the germline gene expression of the corresponding genes account for the different modes of evolution.

A major problem of the conducted analysis is the choice of an appropriate window size. On one hand a fine-scaled analysis on the gene level with small window sizes is desired. And on the other hand a sufficient sample size for accurate estimates of the substitution model is required to overcome numerical problems of zero entries in the divergence matrix. Based on a simulation study, I assessed the minimal window size to

---

**Figure 5.7 (following page):** Substitution rate differences between human and chimpanzee in significant and non-significant alignment windows The rate differences of transitions (red) between the human and chimpanzee models show a bimodal distribution only in significant alignment windows.
be at least 15 Kb for human, chimpanzee and rhesus alignments (see section 6.2). The cut-off for the minimal size for the $TRF_{125Kb}$ dataset was 23 Kb, and hence sufficient to prevent numerical problems. A higher resolution of the model differences is expected with an increasing number of observed substitutions. The fraction of significant tests should therefore increase with the size of the alignment and with the divergence of the sequences from the corresponding branches that are tested. As expected the fraction of significant tests increases slightly from the smaller (125 Kb, 5%) to the bigger (1000 Kb, 8%) sized alignment windows.

The distribution of the estimated p-values appeared almost uniformly distributed and the fraction of observed significant alignment windows was only marginally larger than 5%, which would be expected significant just by chance at the chosen significance level. This finding is not too surprising. The close evolutionary relationship between the two species suggest that the majority of genomic regions may simply evolve alike. Indeed, several studies showed that a general homogeneity of the substitution process between the human and chimpanzee was suggested [Webster et al., 2003; Ebersberger et al., 2002]. Therefore, only a low fraction of differently evolving genomic regions was expected. Their detection is then further more hindered by the limited amount of divergence between human and chimpanzee sequences making the detection of significant differences between the species-specific substitution models hard. To show that it is not likely that the identified candidate regions did not come out as significant solely due to multiple testing, the different substitution patterns between human and chimpanzee were analyzed.

For this analysis, the set of alignment windows was selected, where the homogeneity of the human and chimpanzee substitution model was rejected ($p \leq 0.05$) and that additionally resulted in a $\delta$ that exceeded the 95% quantile estimated for all alignment windows. The rationale for selecting only alignments with a large $\delta$ was to focus on

---

**Figure 5.8 (following page):** $A \leftrightarrow G / T \leftrightarrow C$ rate ratio of human and chimpanzee in significant and non-significant alignment windows. The distributions of the transition ratio $A \leftrightarrow G / T \leftrightarrow C$ estimated from the human (x-axis) and chimpanzee (y-axis) substitution model from significant alignment windows are compared in a density and a contour plot. In significant alignment windows a bimodal distribution is observed, where a mean 1.5 fold $A \leftrightarrow G / T \leftrightarrow C$ transition ratio in human corresponds to a mean 1 fold $A \leftrightarrow G / T \leftrightarrow C$ transition ratio in chimpanzee and vice versa. In non-significant alignment windows a mean of 1.25 fold $A \leftrightarrow G / T \leftrightarrow C$ ratio is observed in both species.
such windows, where the difference between human and chimpanzee substitution models is substantial, and may also be biologically significant. A total of 565 from the 717 significant alignment windows were used for further analysis ($TRF_{125Kb,p\leq 0.05,\delta>95\%}$).

The $\delta$ statistic agglomerates the difference between the human and chimpanzee substitution model as a single value distance measure. It was verified whether the $\delta$ statistic captures the sought-after differences of strand-specific substitution patterns by identifying which substitution rates contributed most to the distance measure $\delta$. A statistic for each substitution rate was defined, where the null-distribution was estimated from simulated alignments that were evolved under the homogeneity assumption for each alignment window. In significant alignment windows it was observed that either the $A \leftrightarrow G$ or the $T \leftrightarrow C$ transition were significant, whereas the difference of other substitution rates (i.e. transversions) between the human and chimpanzee models were less pronounced. Therefore, it can be presumed that the $\delta$ statistic is mostly influenced by the differences of transition rates between human and chimpanzee.

Interestingly, the difference of the individual transition rates compared between the human and chimpanzee model in significant alignment windows showed a bimodal distribution. This suggested an increased $A \leftrightarrow G$ or $T \leftrightarrow C$ in either human or chimpanzee, which results from a strand-specific increase of $A \leftrightarrow G$ over $T \leftrightarrow C$ in only one species. The comparison of the $A \leftrightarrow G$ over $T \leftrightarrow C$ ratio between the human and chimpanzee models confirmed this presumption in the significant alignment windows. A $\sim 1.4$ fold increased $A \leftrightarrow G$ over $T \leftrightarrow C$ rate in one species is opposed to a ratio of $\sim 1$ fold in the respective other species.

In summary, our analysis shows that the significant alignment windows are consistent in their pattern of difference between humans and chimpanzees. There is argues strongly against multiple testing artifacts during their selection. Rather it implies that this difference in this mode of evolution between the two species is biologically meaningful. Interestingly, the differences between the substitution models in human and chimpanzee are in agreement with the hypothesis that the corresponding regions experience differing extents of transcription-coupled repair reflecting a differential expression of the corresponding genes in the germline.
Chapter 6

Simulation and evaluation studies

The Single-Branch-Test introduced in chapter 5 was used to detect significant differences between the substitution model of human and chimpanzee. In this chapter the Single-Branch-Test is evaluated to detect a change in the substitution model between two branches in an unrooted 3 species tree. Further, the influence of the alignment size and the branch lengths in the tree on the test statistic $\delta$ is evaluated.

6.1 Evaluation of the Single-Branch-Test

In this section, the statistical power of the Single-Branch-Test (SBT) to detect a change in the substitution model is inferred from a simulation procedure. The simulation is performed by testing the homogeneity of the substitution model between two branches in a 3 species tree. Two sets of alignments are generated, one using the same model $Q$ for all branches in the tree and a second, where the substitution model $Q$ is changed to $Q^*$ in one branch (Figure 6.1).

The differences between the substitution models are more likely being detected when the number of observed substitutions is higher. This can be achieved by enlarging either the alignment size or the branch length. For the simulation study the alignment size was kept fixed to a size of 25 Kb and the length of the branches was varied. The homogeneity of the substitution model was tested between the two branches $b$ and $c$ that have the same branch length ranging from 0.001 to 0.1 substitutions per site, where the length of the outgroup branch $a$ is doubled (Figure 6.1). A set of 1,000 alignments with a size of 25 Kb were generated according to an unrooted 3 species tree. In the following a significant
test denotes a test performed for an alignment, where the homogeneity assumption of the substitution model between the two branches was rejected.

In total, 3 datasets were generated using the same model $Q$ on all branches with uniform nucleotide frequencies $\pi_0 = (0.25, 0.25, 0.25, 0.25)$. For the first dataset all rate parameters were equal with $r_0 = (1, 1, 1, 1, 1, 1)$. To reduce the effect of the random sampling error the number of alignments was increased for the second dataset to 10,000 alignments generated with $r_0 = (1, 1, 1, 1, 1, 1)$. For the third dataset the rate parameters were randomly generated ($r_r$). The random rate parameters $r_r$ were generated by the following procedure: The off diagonal elements of the matrix are sampled from a uniform distribution $\{0,1\}$, and the diagonal elements were then chosen such that the rows sum up to 1. The resulting random rate matrix $R$ was normalized by dividing all entries by the sum of the diagonal (trace). The rate parameters in $r_r$ are defined from the six rates in the upper triangle of the random rate matrix $R$.

To arrive at a modified substitution model, $Q^*$, the nucleotide frequencies were altered to $\pi_1 = (0.15, 0.25, 0.25, 0.35)$, and $\pi_2 = (0.15, 0.2, 0.3, 0.35)$. More severely altered models $Q^*$ were generated by changing the rate parameters to $r_1 = (1, 1, 1, 1, 2, 1)$ and $r_1 = (1, 1, 1, 1, 4, 1)$. For each of the four $Q^*$ models one dataset was created. Additionally, an alignment set was generated, where individual randomly generated rate
parameters \( r_r \) were assigned in \( Q_r \) and \( Q_r^* \).

For the simulation study the sequence simulation program TRIGEN that is based on the Single-Branch-Test code was developed. The program allows to specify individual substitution models and branch lengths for each branch in an unrooted 3 taxon tree.

The power of the Single-Branch-Test to detect a difference in the substitution model between two branches was assessed by the fraction of significant tests for a set of alignments. For the alignments that were generated using the same substitution model for all branches in the tree the homogeneity of the substitution model was falsely rejected in a percentage of 3.2 % to 5.7 %. The slight variations around the percentage of 5% expected by chance due to multiple testing are explained by the random sampling error. The random sampling error is reduced when the dataset is extended from 1,000 to 10,000 alignments. The percentages for the different branch lengths showed less variations and converge closer to 5% (Table 6.1).

Relative to the observed percentage of 5% tests that are expected significant by chance, the percentage of significant tests increases notably with increasing branch lengths when the substitution model is changed in one branch. When moderate differences are introduced between \( Q \) and \( Q^* \) by changing the nucleotide frequency parameters the fraction of significant tests increases for branch lengths >0.005 expected substitutions per site. For more rigorous differences between \( Q \) and \( Q^* \), where one rate parameter was two- and fourfold increased or random rate parameters were assigned in \( Q \) and \( Q^* \), an increase was also observed for the smallest branch length of 0.002 and 0.001 expected substitutions per site (Table 6.1). However, with increasing branch length it is more likely to detect a difference between the substitution models due to more accurate estimates.

### 6.2 Minimal window size

For the accurate estimation of the substitution model an appropriate choice of the alignment size is required. When the sequence divergence is low or the alignment size is too small not all substitution types are observed in a pairwise sequence comparison. The missing observations lead to zero entries in the divergence matrix \( H \) (c.f., equation 5.1) that can cause wrong estimates of the rates in \( Q \). In the following an example for a divergence matrix with zero entries inferred from an alignment between sequence B and
Table 6.1: Evaluation study for the Single-Branch-Test. The homogeneity of the substitution model between two branches was tested for 8 alignment sets that were generated using a homogeneous model $Q^{(1-3)}$ for all branches and a non-homogeneous model, where a different model $Q^*$ was assigned to one branch (4-8). Shown are the fractions of test results obtained for each dataset consisting of 1000 alignments with a size of 25 Kb. For the simulation using a homogeneous model a dataset with 10,000 alignments with a size of 25 Kb was considered as well (2). The homogeneous substitution model (1-3) was generated with uniform nucleotide frequencies, where the rate parameters were defined in $r_0$, from a randomly generated rate parameters $r$. Moderate changes in the substitution model are implied by altering the nucleotide frequency parameters (4,5) and more rigorous changes by altering the rate parameters (6,7,8) between $Q$ and $Q^*$. The parameters that were used to generate the datasets are the nucleotide frequencies: $\pi_0 = (0.25, 0.25, 0.25, 0.25)$, $\pi_1 = (0.15, 0.25, 0.25, 0.35)$, $\pi_2 = (0.15, 0.2, 0.3, 0.35)$, the rate parameters: $r_0 = (1, 1, 1, 1, 1, 1)$, $r_1 = (1, 1, 1, 1, 2, 1)$, $r_2 = (1, 1, 1, 1, 4, 1)$ and randomly generated rate parameters $r_r$.

\[
H^{(B,C)} = \begin{pmatrix}
A & 0.260500 & 0.000000 & 0.000500 & 0.000250 \\
C & 0.000000 & 0.246000 & 0.000750 & 0.000000 \\
G & 0.000500 & 0.000750 & 0.245000 & 0.000000 \\
T & 0.000250 & 0.000000 & 0.000000 & 0.245500
\end{pmatrix}
\]

To overcome the problem of zero entries in the divergence matrix a large alignment size or a higher sequence divergence between the sequences is required. In the following a simulation study is described to infer the minimal alignment size that is required to obtain accurate estimates of $Q$. For different branch lengths of 0.001, 0.002, 0.005, 0.01, 0.02, 0.05 and 0.1 expected substitutions per site, a set of 1000 alignments are generated.
ranging from 100 bp to 30000 bp. For the simulation, a GTR model with uniform nucleotide frequencies was used, where the rate parameters were randomly generated as described in section 6.1.

The zero entries that occur in the divergence matrix \( H \) between a chosen pair of sequences are counted for each alignment and then averaged and rounded to integers for each alignment set. As expected, the number of zero entries decreases with increasing branch lengths or alignment size. For example, for a branch length of 0.005 expected substitutions per site an alignment size of at least 15 Kb is sufficient to avoid zero entries in the divergence matrix (Figure 6.2). For smaller branch lengths such as 0.002 expected substitutions per site larger alignments with a size of at least 30 Kb should be considered.

### 6.3 Evaluation of the test statistic \( \delta \)

The statistic for the Single-Branch-Test is defined by the squared distance \( \delta \) between the substitution rate vectors \( q \) of two branches \( i \) and \( j \) (equation 5.7). In this chapter the accuracy and precision (Taylor, 1996) of \( \delta \) to measure the distance between two substitution models is evaluated with respect to the influence of the alignment size and the branch lengths. The accuracy defines how close \( \delta \) is to the unknown true value and the precision describes the scatter of the estimates of \( \delta \) (Figure 6.3). When the homogeneity of the substitution model is assumed, \( \delta \) should reflect the random variations of the substitution process between the branches of a tree that converge near zero with an increasing accuracy and precision of the estimates. In this section the distributions of \( \delta \) were simulated for different alignment sizes, substitution models and branch lengths.

For the evaluation, \( \delta \) was estimated from simulated alignments for an unrooted 3 taxon tree. Two sets of alignments were generated, one using a homogeneous model for all branches in the tree, and a second using a non-homogeneous model, where the substitution model was changed for one branch. The rate parameters of the substitution models were randomly generated as described in section 6.1. The alignments were generated using the TRIGEN program (c.f. section 6.1).

The impact of the alignment size and the branch length on the measurement of \( \delta \) was inferred by the following. For the evaluation study, \( \delta \) is computed for two branches
6.3 Evaluation of the test statistic $\delta$

![Graph](image)

**Figure 6.2:** Number of expected zero entries in a divergence matrix. 1000 alignments each were generated for varying alignment sizes (100 to 30000 alignment columns) and branch lengths (0.001 to 0.1 expected substitutions per site). The average number of zero entries in the divergence matrices is shown for the different alignment sizes and branch lengths.

$b$ and $c$ in a 3 taxon tree from simulated sequence alignments (Figure 6.4 B). In the first part, the branch lengths of the tree were kept constant and the alignment size was varied from 5 Kb to 1 Mb (Figure 6.4 D.1). In the second part, the alignment size was kept constant and the length of the branches in the tree were varied between 0.001 and 0.2 substitutions per site (Figure 6.4 D.2). The third part, the alignment size and the lengths of the branches $b$ and $c$ were kept constant and the length of branch $a$ (outgroup) was varied between 0.005 to 0.25 substitutions per site (Figure 6.4 D.3).
6.3 Evaluation of the test statistic $\delta$

Figure 6.3: The accuracy describes how close the estimate $\delta$ is to the unknown true value and the precision describes the scatter of $\delta$.

6.3.1 Effects of the alignment length on the estimate of $\delta$

The accuracy of $\delta$ is expected to improve with increasing alignment size as more substitutions can be considered to describe the substitution process. In a simulation procedure $\delta$ was estimated from simulated alignments of different lengths for a tree with constant branch lengths. The lengths of the branches $b$ and $c$ were set to 0.005 expected substitutions per site and the length of branch $a$ was set to 0.01 expected substitutions per site. The alignment size was varied in 5 Kb steps from 5 Kb to 100 Kb and further to 125 Kb, 250 Kb, 500 Kb and 1 Mb (24 intervals in total).

The statistic $\delta$ is used to measure the variations between the substitution models of

Figure 6.4 (following page): Evaluation of $\delta$. For the simulation procedure sequence alignments were generated using a homogeneous (Simulation 1) and non-homogeneous (Simulation 2) substitution model. For Simulation 2, the substitution model for the branch $b$ differs from the other branches $a$ and $c$ in the tree. The values of the rate parameters were randomly generated. The statistic $\delta$ was estimated for the branches $b$ and $c$ from the simulated alignments. To measure the impact of the alignment size and branch length on the estimate of $\delta$ the following settings were used: 1) The lengths of all branches were kept constant and the sizes of the alignments were varied from 5 Kb to 1 Mb. 2) The alignment size was kept constant and the lengths of the branches in the tree were varied from 0.001 to 0.2 expected substitutions per site. 3) The alignment size and the length of branches $b$ and $c$ were kept constant and the length of branch $a$ was varied from 0.001 to 0.2 expected substitutions per site.
6.3 Evaluation of the test statistic

\[ \delta = \sum_{i=0}^{n} (q_{i}^{(b)} - q_{i}^{(c)})^2 \]

**A)** Simulation 1 (homogeneous model)  
Simulation 2 (non-homogeneous model)

\[ q^{(a)} = q^{(b)} = q^{(c)} \]

**B)** Statistic:

**C)** Random Rate Matrix

\[ Q = \begin{pmatrix}
    A & G & C & T \\
    a & \beta & \gamma & \delta \\
    \delta & \gamma & \epsilon & \eta \\
    \eta & \epsilon & \delta & \gamma
\end{pmatrix} \]

Base frequencies \( \pi_A = \pi_C = \pi_T = \pi_G = 0.25 \)

**D)**

1) Alignment Size  
Simulation 1 and 2  
Alignment size 50 Kb

ACGTGGGCTATATATAG  
ACGTGGGCTATATATAG  
ACGTGGGCTATATATAG

2) Branch-Length  
Simulation 1 and 2  
Alignment size 50 Kb

B  
0.001 to 0.2  
0.002 to 0.4

C  
A

3) Outgroup Branch-Length  
Simulation 1 and 2  
Alignment size 50 Kb

B  
0.001 to 0.2  
branch length  
a and c constant

C  
A

5 Kb to 1000 Kb

Alignment
branch b and c. In simulation 1 a homogeneous model was used for all branches in the tree. In simulation 2 a non-homogeneous model was used. For the non-homogeneous model, the substitution model was modified for branch b (Figure 6.4). The effect on the estimate of $\delta$ caused by the varying alignment lengths is remarkable, especially for length $<20$Kb (Figure 6.5). As expected, $\delta$ is substantially increased when the substitution model is changed on one branch. For alignment length $\geq 20$ Kb, the distributions of the $\delta$ estimate are by and large similar. Only for the short alignments (5-15Kb) the $\delta$ distribution is slightly shifted towards larger values. In summary, this indicates that for alignments $\geq 20$ Kb length, the alignment length has only very little influence on the $\delta$ estimate.

6.3.2 Branch length effects

The accuracy of $\delta$ is expected to improve with increasing branch lengths as more substitutions can be considered to describe the process. In total, 21 unrooted three taxon trees were generated, where the individual trees varied in their branch lengths. The branch lengths were chosen from the interval between 0.001 to 0.2 expected substitutions per site for the branches b and c. The branch length of branch a was set two twice the branch length of branch b and c. The alignment size was fixed to 50 Kb. In simulation 1 again 1000 alignments were generated for each individual tree using a homogeneous substitution model for the three branches. In simulation 2 the different substitution model was chosen for branch b compared to branch a and c.

An influence on the estimate of $\delta$ for branches b and c caused by the varying branch length is observed for short branches of $< 0.01$ expected substitutions per site. Again, as expected $\delta$ between b and c is highly increased when the substitution model is changed in branch b. For the short branches $< 0.01$ a slight shift of $\delta$ towards larger values is observed in comparison to longer branches. The branch length has therefore not a strong influence on the estimate of $\delta$ when the substitution model differs between the branches (Figure 6.6).

6.3.3 Effects of the outgroup divergence on the estimate of $\delta$

The accuracy of $\delta$ measured between the branches b and c is evaluated for a putative influence of the length of the outgroup branch a. For the simulation the length of the
branches b and c were fixed to 0.005 expected substitutions per site and the outgroup branch a was increased from 0.001 to 0.2 expected substitutions per site. In total, 21 trees were considered. For each tree, 1000 alignments were generated with uniform nucleotide frequencies, using a homogeneous model and a non-homogeneous substitution model (Figure 6.7). The outgroup branch length does not affect the δ measurement for the branches b and c. A slight increase of δ was observed only for very large outgroup branch lengths with more than 0.1 expected substitutions per site.

6.4 Normalization of δ

For the analysis of real sequence data stemming from different regions of a genome, both the divergence between sequences and the number of analyzable alignment columns can vary. In the previous section I have shown that both alignment length and extent of sequence divergence have an influence on the δ estimate. In the following a normalization procedure for δ is developed that considers the size of the alignment and the length of the branches for the estimates.

The bias on δ caused by the alignment size can be modeled by a linear factor. The factor is described as ratio of the number of alignment columns \( L_{\text{size}} \) to an arbitrarily chosen reference size of 50 Kb (\( L_{50\text{Kb}} \)). The normalized δ is obtained from the normalization function \( F_{N1} \) by:

\[
F_{N1}(\delta) = \delta \times \frac{L_{\text{size}}}{L_{50\text{Kb}}}
\] (6.1)

Figure 6.4 shows the result of the normalization on δ estimates for different alignment sizes.

The influence on δ caused by different branch length is modeled by a calibration curve that was estimated from simulated alignment sets. For a fixed alignment size of 50 Kb, 1000 alignments were generated for varying branch length, using a homogeneous model for all branches in the tree (Figure 6.4 a). For each branch length the average of δ was estimated from 1000 simulated alignments. The calibration Function \( F_{\text{cal}} \) was defined from the linear interpolation between the averages of δ estimated for each branch length as function to the branch length. The distributions of δ were normalized to a fixed
6.4 Normalization of $\delta$

reference branch length of $bl_{0.01}$ by the following:

$$F_{N2}(\delta) = \delta \times \frac{F_{cal}(bl_{0.01})}{F_{cal}(bl)}$$ (6.2)

The branch length $bl$ of the branches $b$ and $c$ was estimated from the alignments using the Neighbor-Joining method (Saitou and Nei 1987). Figure 6.4 b shows the result of the normalization procedure. The normalized $\delta$ is now independent from the branch length.
Figure 6.5: Influence of the alignment size on $\delta$. The left plot shows the distribution of $\delta$ estimates for different alignment sizes when a homogeneous model (blue, simulation 1), and a non-homogeneous model (red, simulation 2) was used. In total, 1000 alignments 50 Kb in size were generated for each alignment size. The linear interpolation (blue) between the averages of $\delta$ estimated for the different alignment sizes describe the calibration function of $\delta$ as function to the alignment sizes.
Figure 6.6: Influence of the branch length on $\delta$. The left plot shows the distribution of $\delta$ estimates for different branch length when a homogeneous model (blue, simulation 1), and a non-homogeneous model (red, simulation 2) was used. In total, 1000 alignments 50 Kb in size were generated for each branch length. The average of $\delta$ for different alignment sizes are shown in the right plot.
Figure 6.7: Influence of the branch length of branch a on $\delta$ that was estimated between branch b and c. The left plot shows the distribution of $\delta$ estimates for different branch length of branch a when a homogeneous model (blue, simulation 1), and a non-homogeneous model (red, simulation 2) was used. The branch length of b and c was 0.005 and kept constant. In total, 1000 alignments 50 Kb in size were generated for each branch length. The average of $\delta$ for different alignment sizes are shown in the right plot.
Figure 6.8: Effect of the $\delta$ normalization. The left plot shows the effect of the normalization of $\delta$ with respect to varying alignment sizes. The right plot shows the effect of the normalization of $\delta$ with respect to varying branch lengths. $\delta$ is adjusted to a branch length of 0.005 (left) and an alignment size of 50 Kb (right).
6.5 How to decrease the computing time of the Single-Branch-Test

The Single-Branch-Test uses a parametric bootstrap approach to estimate the null distribution of $\delta$ from 1,000 simulated alignments that were generated according to a homogeneous model for all branches in the tree. The speed-up of the procedure can be realized by two general approaches. The first is a simple stopping rule, where the simulation procedure runs until a fixed number of simulated test statistics had exceeded the observed one from the data, i.e., exceeded the chosen confidence limit. The second strategy is to estimate the p-values directly from a standard null distribution of $\delta$ corresponding to a fixed alignment size and branch length. For $\delta$ no standard null distribution was assumed as the estimates of the branches in the tree are expected to be highly correlated. Therefore the distribution of $\delta$ given the null-hypothesis is true are determined by a parametric bootstrap approach. However, the simulation study conducted in the previous section showed that the correlations between two branches are mainly dependent on the branch length and the alignment size. For a normalized $\delta$ value, the null distribution corresponding to fixed alignment size and branch length can be considered (c.f. section 6.4). When the parameters of a chosen standard distribution are fitted to a null distribution of $\delta$ corresponding to a fixed alignment size and branch length, the probability to reject the homogeneity of the substitution model between two branches can then be directly estimated from the cumulative distribution function of the standard distribution.

6.5.1 Estimating the p-values of the Single-Branch-Test from a Gamma distribution

The Gamma distribution model has a shape and a scaling parameter for a distribution of positive continuous values and is therefore appropriate to model the null distribution of $\delta$. The parameters of the Gamma distribution were fitted to the null distribution of $\delta$ corresponding to a fixed branch length and alignment size. The Gamma distribution has a scale parameter $k$ and a shape parameter $\lambda$. The density function is described by:
6.5 How to decrease the computing time of the Single-Branch-Test

\[ f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, \quad x > 0 \]  

(6.3)

As the null distribution of \( \delta \) is estimated only once, the number of samples was increased from 1,000 to 10,000 samples to further reduce the effects of random sampling errors. The null distribution of \( \delta \) was estimated from 10,000 alignments that were generated according to a homogeneous model for all branches in the tree with branch lengths of 0.005 (expected substitutions per site) and an alignment size of 50 Kb. The parameters of the Gamma distribution were estimated using the \texttt{fitdistr()} function in R from the MASS package \cite{Venables2002}. The null distribution of \( \delta \) was then compared to 10,000 randomly sampled values of the Gamma distribution according to the estimated scale parameter \( k=2.44 \) and shape parameter \( \lambda=43.41 \). The Gamma distribution shows a good fit to the null distribution of \( \delta \) (Figure 6.9). With the estimated parameters of the Gamma null distribution the probability that the given null-hypothesis is true can be directly estimated from the Gamma cumulative distribution function.

To show, whether the parametric approach can be applied on real data the p-values of the Single-Branch-Test estimated for the \( TRF_{125KB} \) dataset obtained from the parametric bootstrap approach were compared to p-values that were estimated from the Gamma distribution with the given parameters. For each alignment window in the \( TRF_{125KB} \) dataset the p-value corresponding to the normalized \( \delta \) value was estimated from the Gamma cumulative distribution function. In Figure 6.10 the p-values estimated from the parametric bootstrap null distribution and from the gamma null distribution of \( \delta \) show as expected a high correlation with \( R^2 = 0.94 \). The estimation of the p-values from a standard distribution extensively accelerate the computation time of the Single-Branch-Test as the simulation procedure to obtain the null distribution of \( \delta \) between two branches is not required.
Figure 6.9: Fitting the parameters of a Gamma distribution to the null distribution of $\delta$. $\delta$ was estimated from 10,000 simulated alignments generated using a homogeneous model for all branches in a 3 taxon tree with an alignment size of 50 Kb. The Gamma distribution is randomly sampled for 10,000 values from a Gamma distribution with the scale parameter $k=2.44$ and shape parameter $\lambda=43.41$. 
Figure 6.10: Comparing p-values of the $TRF_{125kb}$ dataset estimated from the parametric bootstrap approach and from the gamma distribution function.

$R^2 = 0.94$
6.6 Discussion

In this chapter the sensitivity and reliability of the Single-Branch-Test to detect a change in the substitution model was evaluated. Moreover, the minimal size of the alignment that is required for accurate estimates was inferred, and the influence of the alignment size and branch length on the test statistic $\delta$ was assessed.

In the first part, the homogeneity of the substitution model between two branches was tested for simulated alignments using the Single-Branch-Test. A set of alignments were generated using different branch lengths under a homogeneous and a non-homogeneous substitution model for the branches. In the non-homogeneous substitution model, the substitution model differed only between the two branches that were tested in the Single-Branch-test. The probability to reject falsely the homogeneity assumption between two branches was in agreement with the expected 5% tests that are expected significant by chance due to multiple testing (Table 6.1).

The differences of the substitution model between the branches could be better resolved by larger alignment sizes and when branch lengths are increased. This is since more substitutions were available, a better inference of the underlying substitution model is facilitated. The more precise the parameters are estimated the smaller can the model distance be to be still significant. The fraction of significant tests therefore increases with the size of the alignments or the branch length. The fraction of significant tests increased substantially when branch-specific substitution models were used. More pronounced model differences were more often detected than only subtle differences as e.g. differences in the equilibrium frequency. In summary, the results show that the Single-Branch-Test is reliable and sensitive enough to detect a difference in the substitution model between two branches. In chapter 5 a study was conducted to test the homogeneity of the substitution model between the human and chimpanzee branch. The branch length between human and chimpanzee corresponds in average to a branch length of 0.005 expected substitutions per site. For this branch length the Single-Branch-Test is expected to detect only datasets with strong differences between the substitution models as only the simulated datasets with the strongest substitution model differences a percentage of 99.1% correctly rejected the nullhypothesis (dataset 7, Table 6.1).

In the second part, a simulation was conducted to infer the minimal size of the alignment that is required to obtain accurate estimates of the substitution model. Zero entries
in a divergence matrix are a problem for the estimation of the substitution model when the alignment size is limited and the sequence divergence is low. Such zero entries can cause numerical problems for the estimation of the substitution rates in $Q$, and hence need to be avoided.

For the estimation of $Q$ between human and chimpanzee, the simulation suggests that for a branch length of 0.005 an alignment size of 15 Kb is sufficient to avoid the occurrence of zero entries in the divergence matrix. Note, that for the simulated datasets randomly generated rate matrices were used that have putatively larger variations between the rates within a model than in a model that was estimated from real data. The approximate cut-off of a minimal alignment size 15 Kb for a branch length $> 0.005$ is therefore sufficient for real or simulated data.

The statistic $\delta$ is used in the Single-Branch-Test to measure the distance between the substitution models of two branches. In the third part of this chapter, a simulation study was presented that assessed the impact of varying alignment sizes and branch lengths and outgroup branch length to the measure of $\delta$. The increase of both the alignment size and the branch length decreases the measured distance of $\delta$ between two branches due to a higher accuracy and precision of the estimates. On the other hand, the outgroup branch length proved to have little impact on $\delta$ that is measured between the other two branches. Therefore it can be presumed that the branch-specific substitution models can be accurately estimated independent of the choice of the outgroup.

A normalization procedure was developed to account for influences on the accuracy of $\delta$ caused by different alignment sizes and branch lengths. The branch length effect on $\delta$ was modeled by fitting a curve to the mean values of $\delta$ obtained from simulated datasets for differing alignment sizes and branch lengths. I could show that after the normalization the measure of $\delta$ was independent of the alignment size and branch length. For the study conducted in chapter 5 the alignment windows that rejected the homogeneity of the model between the human and chimpanzee branch with the largest distance $\delta$ were selected among all alignment windows. The variations of the alignment length and branch lengths among the alignment windows can lead to a bias for selecting shorter alignment windows or alignment windows with a low sequence divergence when the normalization would not be considered.

The parametric bootstrap approach of the Single-Branch-Test is time intensive for a genome-wide analysis that uses large alignment sizes. For each alignment the null
distribution of $\delta$ is determined from a simulated dataset of 1,000 alignments that are simulated with a homogeneous model. One way, to reduce the computation time is to reduce the number of simulations. To estimate whether a test rejects the null hypothesis it suffices to run the simulation until the observed $\delta$ exceeds $\delta_{null}$ in 5% of the simulations. However, for a test that rejects the null-hypothesis significantly the computation time would not be reduced as only less than 5% of the simulated $\delta$ values are larger than the observed $\delta$.

A more effective alternative is to model the null-distribution of $\delta$ by a known standard distribution. The shape and rate parameters of a Gamma distribution were estimated for a null distribution of $\delta$ from an alignment dataset with fixed alignment size and branch length (e.g. 50 Kb and a branch length of 0.005). The p-values corresponding to the normalized $\delta$ values of the TRF$_{125}$Kb dataset were recomputed from a Gamma cumulative distribution function according to the estimated shape and rate parameters. The comparison of the p-values estimated from the bootstrap approach to the p-values estimated from the Gamma distribution are as expected highly correlated. This approach would allow to substantially reduce the computing time especially for large-scale datasets with large alignment sizes as the simulation procedure would not be required in the Single-Branch-Test. Such a speed-up would be desirable, compared to the bootstrap method, which is currently used.
Chapter 7

Functional analysis of candidate genes

This chapter describes a Gene Ontology enrichment analysis to characterize the biological role of the set of candidate genes that were identified from the significant alignment windows described in chapter 5.

In chapter 5 a set of alignment windows were identified, for which human and chimpanzee showed a significantly different extent of the strand-specific $A \leftrightarrow G$ rate. The difference of the strand-specific $A \leftrightarrow G$ rate between the two species over the other can be interpreted as a signal that the expression of the corresponding genes has changed either in humans or chimpanzees. These genes may therefore play a role in processes leading to phenotypic differences between humans and chimpanzees.

A biological meaningful interpretation of the results obtained from a genome-wide screen is essential to draw conclusions about the underlying affected processes in the cell. Among the most popular and most widely used are Gene Ontology (Ashburner et al. 2000), Panther (Thomas et al. 2003) and the molecular signature database (Subramanian et al. 2005). The molecular signature database (Subramanian et al. 2005) provides curated collections of genes that are derived from experiments in which chemical and genetic approaches are applied to perturbate gene expression, canonical pathways and several pathway databases such as KEGG (Kanehisa and Goto 2000), BioCarta (http://www.biocarta.com). The pathway databases provide manually curated pathway maps that are based on well studied protein-protein interactions.

The Gene Ontology (Ashburner et al. 2000) provides a structured and defined vocab-
ularity describing the role of genes and gene products in eukaryotic cells. Terms from this vocabulary are assigned to genes to reflect functional information about this gene provided by a variety of sources, e.g. from experimental analysis (e.g., direct assay) from an experimental analysis (e.g., a direct assay), from author statements in the literature or computational analysis (e.g., inferred from sequence similarity). The assignments in the Gene Ontology are manually revised by a curator and complemented by automatic assignments that are inferred from computational analyses that are not manually revised. Gene products are described in the Gene Ontology (GO) by a terminology that is divided in three domains: 1) Biological Process, 2) Molecular Function and 3) Cellular Component. Gene Ontology terms are organized as nodes in a directed acyclic graph. A directed acyclic graph is a graph with no directed cycles, i.e., a path that starts in any node $n_i$ does not lead back to the same node $n_i$.

A catalogue of statistical approaches can be applied to obtain the relevant biological terms that describe the characteristics of a list of genes. Although, there is no unified strategy most methods compare a list of candidate genes derived from a genome-wide screen or experiment to a set of reference genes and test for an enrichment of associated ontology terms. The reference gene list is usually defined by all genes available in the organism’s genome or by the list of genes that were considered in the analysis. The most common approach is to infer an over- or under-representation of ontology terms in the candidate gene set. The statistical methods to test for an enrichment commonly applied are the hyper-geometric test (Robinson et al., 2002, one sided Fisher exact test), Fisher exact test (Zeeberg et al., 2003; Beissbarth and Speed, 2004), $\chi^2$ test (Khatri et al., 2002) and a binomial test (Maere et al., 2005).

### 7.1 Gene Ontology enrichment analysis

The Gene Ontology annotation for all human genes in Ensembl was retrieved from hsapiens gene Ensembl (build 53) using the ”biomaRt” (Durinck et al., 2005) package from Bioconductor (Gentleman et al., 2004). The pipeline for the gene ontology enrichment analysis was implemented using the ”topGO” R package (Alexa et al., 2006) and the GOstat R package (Beissbarth and Speed, 2004) from Bioconductor in R version R-2.8.1. The significance level of the enrichment for a gene ontology term was determined by a hyper-geometric test (one sided Fisher Exact Test).


7.1.1 Selected candidate and reference genes

An overview of the number of genes overlapping within the genomic coordinates of the 125 Kb alignment windows in the transcribed fraction of the human genome is shown in Table 7.1. An alignment window overlaps in average with 2.79 individual genes. For significant alignment windows the candidate gene set comprised a total of 1,420 genes. The reference gene set comprises a total of 25,236 genes.

<table>
<thead>
<tr>
<th></th>
<th>Alignment windows</th>
<th>Genes (avg %)</th>
<th>Genes (total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRF125Kb</td>
<td>12,596</td>
<td>2.79</td>
<td>26,656</td>
</tr>
<tr>
<td>p ≤ 0.05</td>
<td>717</td>
<td>2.83</td>
<td>1,956</td>
</tr>
<tr>
<td>Candidate p≤0.05,δ&gt;95%</td>
<td>565</td>
<td></td>
<td>1,420</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td>25,236</td>
</tr>
</tbody>
</table>

Table 7.1: Gene content within the TRF125Kb dataset. The numbers of significant and non-significant alignment windows, the average and total number of genes overlapping with the alignment windows are shown. The candidate genes were defined by the genes that overlap with the 565 significant alignment windows having the strongest substitution model differences between human and chimpanzee measured by δ. The remaining genes were used as reference.

7.1.2 Significantly enriched Gene Ontology terms

From all 26,656 Ensembl genes covered by the alignment windows only 50% were annotated with a gene ontology term. For the enrichment analysis only terms assigned to ≥ 3 candidate genes were considered for the analysis. The enrichment analysis was performed separately for the three classes Biological Process (BP); 917 terms, Molecular Function (MF); 386 terms and Cellular Component (CC); 198 terms (Table 7.2). The p-value distributions resulting from the gene ontology enrichment analysis are shown in a PP-Plot (Figure 7.1). The deviation from the uniform distribution is more prominent for the Biological Process and Molecular Function and less pronounced for Cellular Component terms. Gene Ontology terms with a nominal p-value ≤ 0.05 in the enrichment analysis are selected as significantly enriched in the candidate gene set. For the GO class Biological Process a total of 153 significant terms, for the Molecular Function 60 significant terms, and for Cellular Component 16 significant terms were inferred (Table 7.2).
Table 7.2: Overview of the results from the Gene Ontology enrichment analysis. Shown are the number of significant terms (Terms (p≤0.05)), the total number of tested terms, number of annotated candidates, the number of annotated reference genes and the total number of candidate genes (Candidates (total)) for the three GO domains separately. In total, 908/1420 (63.9%) candidate genes were annotated with at least one term of the three GO domains. Overall, 15,780/26,656 (59.2%) genes were annotated with at least one term of the three GO domains.

<table>
<thead>
<tr>
<th></th>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Cellular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms (p≤0.05)</td>
<td>153</td>
<td>60</td>
<td>16</td>
</tr>
<tr>
<td>Terms (tested)</td>
<td>917</td>
<td>386</td>
<td>198</td>
</tr>
<tr>
<td>Candidates</td>
<td>750 (52.8%)</td>
<td>797 (58.1%)</td>
<td>825 (58.1%)</td>
</tr>
<tr>
<td>Annotated Reference</td>
<td>12,978 (48.7%)</td>
<td>13,983 (52.5%)</td>
<td>14,234 (53.4%)</td>
</tr>
</tbody>
</table>

Figure 7.1: Probability-probability plot of the nominal p-values inferred from the enrichment analysis in the three gene ontology domains (Biological Process, Molecular Function, Cellular Component). The black diagonal line represents a cumulative uniform distribution. Terms were selected with a p-value ≤ 0.05 (vertical line) for further analysis.

7.1.3 Graphic representation of enriched Gene Ontology terms

The tabular output of a gene ontology enrichment analysis includes redundant information for related terms (e.g., parent to child relations) and is often difficult to interpret. This is the more the case when the number of significantly enriched terms is high. The
most intuitive approach to facilitate the interpretation of the results is a visual graph representation of all significantly enriched terms.

The significant terms for the three GO domains (Table 7.2) were arranged in a sub-ontology that was generated by the following procedure. A gene ontology graph object was built containing the set of significant terms using the GOgraph function of the GOstat R package (Beissbarth and Speed, 2004). To reduce the size of the resulting graph, a procedure was implemented that iteratively deletes non significant parental term nodes from the graph. In this iterative procedure the corresponding significant child terms of a deleted node are connected to the parental nodes of the deleted node. The resulting GO graphs for Molecular Function and Cellular Component sub-ontology were plotted using the Rgraphviz (Carey et al., 2005) R package and then edited manually. The graph for the Biological Process sub-ontology was imported to the Cytoscape graph visualization tool (Shannon et al., 2003) to perform the graph layout and manual editing. The GO terms were subsequently organized into a map of functional groups based on their shared roles in biological processes.

The Biological Process graph has in total 153 nodes corresponding to the significantly enriched GO terms (Figure 7.2). The significant enriched terms are related to embryonic developmental processes, anatomical structure morphogenesis, and organ development, such as gut, brain, spinal cord, limb, skeleton and neurogenesis, gliogenesis and axogenesis. Further, terms for cell morphogenesis, cell migration, learning and pain are overrepresented. Significant terms for signal transduction are enriched for a variety of receptor signaling pathways, e.g., Wnt receptor, epidermal growth factor receptor and glutamate signaling pathway. Further, terms are enriched for the regulation of replication, transcription, protein translation and chromatin and histone modification.

For the Molecular Function sub-ontology 60 terms were significantly enriched that are shown in Figure 7.3. Terms for a variety of receptor activities are seen more often than expected by chance, such as the β-adrenergic receptor activity, androgen receptor activity, opioid receptor activity, vascular endothelial growth factor receptor activity, C-C chemokine receptor activity and interleukin-1 receptor activity and prostaglandin E receptor activity. An enrichment was also observed for binding of cytokine, GTP, cAMP and GTPase and for transmembrane transporter activity that involves a variety of ion transporter, symporter and antiporter activity for sulfate, sodium, hydrogen, amino-acids and sugar. Further, terms related to the regulation of transcription involve
transcription factor activity, transcription factor repressor activity, ribonucleases and histone deacetylase activity are enriched.

The analysis of the Cellular Component sub-ontology revealed an enrichment for 16 terms that are shown in Figure 7.4. Here, terms representing the keratin filament, secretory granule, histone deacetylase complex, the cAMP-dependent protein kinase complex, and the post- and pre-synaptic membrane and synapse are enriched.

7.2 Enrichment analysis for curated gene sets

This section gives an overview of the candidate genes by constructing a map of the genes and their association to pathways and experimental gene set collections from the Molecular Signature database (Subramanian et al., 2005). The Molecular Signature database is divided in 5 collections for positional gene sets (C1), curated gene sets (C2), motif gene sets (C3), computational gene sets (C4) and gene ontology gene sets (C5). To identify overrepresented gene sets an enrichment analysis was conducted for the curated gene sets that are termed in the following as C2 gene sets. The candidate and reference genes used for the enrichment analysis are defined in Table 7.1. The gene

Figure 7.2 (following page): Map of the GO Biological Process (BP) enrichment analysis. Shown is a graph of 153 significantly enriched terms. The boxes show the GO identifier, the expected number of genes (EXP) and, in brackets, the number of significant and annotated genes assigned to the term. The box-color reflects the p-value (P) that range from red (p ≤ 0.001) to yellow (p ≤ 0.05). The edges denote parent to child relations among the GO terms. Terms were grouped into 5 main categories i) developmental processes in embryogenesis, ii) signal transduction, iii) regulation of biological processes, iv) chromatin/histone modification and v) biosynthetic processes. The GO terms related to developmental processes cover organ development such as brain, gut, muscle, spinal cord, limb, skelet, epidermis, anatomical structure development and embryonic morphogenesis. Further, terms for neurogenesis, glial cell differentiation, axonogenesis, learning, cell migration and cell morphogenesis are enriched. The second category shows an enrichment for a variety of signal transduction processes, such as Wnt receptor, glutamate, epidermal growth factor receptor signaling pathway, protein Kinase C, arf protein, second-messenger mediated signaling and Intra-S-DNA damage checkpoint. The third category is comprised by terms involved in regulatory biological processes. Among them are regulations of replication, transcription, protein translation. The fourth group describes terms for translational protein modification (e.g., methylation, alkylation) and histone deacetylation. The fifth group describes biosynthetic processes for GTP, CTP and UTP. For a tablular format of the results see Table A.1
7.2 Enrichment analysis for curated gene sets

Limb
Spinal Cord
CNS
Brain
Gut
Skeleton
Embryonic
Pain
Learning
Cell Morphogenesis
Neurogenesis
Regulation
Chromatin/Histone Modification
Biosynthetic Process
UTP
CTP
GTP
Cytoskeleton
Development
Signal Transduction
Wnt
EgfR
ARF
Glutamate
Protein Kinase
Checkpoint
Second-messenger
Translation
Transcription
DNA Replication
DNA Repair
Metabolic Process
Organ
Embryonic
Ephoderm
Epithelial
Axonogenesis
Cell Projection
Cell Migration
Gliogenesis
Calcium
Axonogenesis
Cell Projection
Embryonic
Epidermis
Cytoskeleton
Ephoderm
Epithelial
Translation
Transcription
DNA Replication
DNA Repair
Metabolic Process
Organ
set assignments to human gene symbols were obtained from the Molecular Signature Database repository. In total, 645 candidate genes (of 1,420) and 11,054 reference genes (of 25,236) were assigned to at least one C2 term (of 1,891 C2 terms). The enrichment for each C2 term was inferred by a hypergeometric test (one-sided Fisher’s Exact Test). For the resulting set of enriched terms a graph is constructed, exported to Cytoscape (http://www.cytoscape.org/) and then manually edited.

In Figure candidate genes are shown that correspond to significantly enriched curated gene sets. An enrichment of gene sets derived from studies that investigated the modulation of transcription to UV-C light and 4-nitroquinoline-1-oxide (4NQO) induced DNA damage for different time intervals. Further an enrichment was found for Hox genes, signaling pathways, cell cycle regulators, autophagy and cell growth. The full table and term descriptions of all enriched C2 dataset terms is shown in the appendix section B.

![Figure 7.3](following page): Map of the GO Molecular Function (MF) Enrichment Analysis. The figure shows in total, 60 significant terms (p \leq 0.05) that are arranged in the subfigures (A-F). The terms were processed to subgraphs that are organized in a hierarchical structure, where the more general terms are located at the top and the more specific terms are at the bottom of the graph. The boxes show the GO identifier, the expected number of genes (EXP), and in brackets, the number of significant and annotated genes assigned to the term. The significance level of the p-value (P) is reflected by the box-color ranging from yellow (p \leq 0.05) to red (p \leq 0.001). A) An enrichment is observed for cytokine binding including interleukin-1 receptor activity and C-C chemokine binding receptor activity. Further, the functions for opioid and vascular endothelial growth factor receptor activity are enriched. B) Molecular functions are enriched that are related to cAMP and GTP binding, where the enrichment is strongest for terms leading to GTP binding. The unconnected terms describe an enrichment for the \(\beta\) adrenergic receptor activity, (Rho) GTPase binding, GTPase activity and GTPase activator activity. C) The subgraph consists of 17 related terms that describe transmembrane transporter activity for ions, amino acids and carbohydrates. In particular terms are enriched in ion transporters for sulfate and cation antiporters, cation symporters for sodium with neurotransmitters, sodium with amino-acids and hydrogen with sugar. D) Shown are enriched terms for transcription regulator activity for transcription repressors and transcription factors. Further, genes for the deacetylation of histones and androgen and chaperone binding are enriched. E,F) A collection of the remaining unconnected enriched GO terms. Terms with only 3 genes for pancreatic ribonuclease activity (endonucleolytic cleavage of RNA), phosphatidate phosphatase activity, carbamoyl-phosphate synthase activity and cytoskeletal adaptor activity. Further, an enrichment for glycosaminoglycan and single-stranded DNA binding is shown. For a tabular format of the results see Table A.2.
7.3 Discussion

The genome-wide analysis of the transcribed fraction in the human-chimpanzee-rhesus genome alignment identified 565 significant 125 Kb alignment windows, where the ho-
mogeneity of the substitution patterns between human and chimpanzee was rejected and that showed the $\sim 5\%$ most extreme model difference between both species. The differences of the substitution patterns in the human and chimpanzee models were suggested to result from different extents of transcription-coupled repair in the two species. This provides an initial indicative that the corresponding genes are differentially expressed in the germline of human and chimpanzee. A gene ontology enrichment analysis was conducted to elucidate, whether the identified genes might play a role in embryonic developmental processes that determine the phenotypes of human and chimpanzee.

From the 565 significant alignment windows, a total of 1,420 genes were defined as candidate gene set and 25,236 reference genes overlap with the remaining alignment windows. An enrichment for the three domains of the gene ontology showed that the candidate gene set is enriched in terms related to embryonic developmental processes, such as organ development and neurogenesis, signal transduction, and a variety of regulatory biological processes such as the regulation of replication, transcription, protein translation and post-translational protein modification that modulate gene expression. In summary, the functional annotation of the candidate gene set imply a regulatory role for developmental and morphological processes that could indeed account for the phenotypic differences between human and chimpanzee.

There are several alternative strategies available that could be applied to test for the enrichment of a term vocabulary such as provided by Gene Ontology. Some methods require a weighting measure or statistic that is assigned to each gene, e.g., that was experimentally inferred. The simplest approach is then to compare the distributions of the measures between the candidate and reference genes for each gene set. To assess the significance of the difference between the two distributions a variety of standard statistical methods can be employed, such as the Kolmogorov-Smirnov test (Ben-Shaul et al., Figure 7.5 (following page)).

Figure 7.5 (following page): Map of the Molecular Signature Database C2 dataset Enrichment Analysis. Shown are the significantly enriched C2 terms with a nominal p-value $\leq 0.05$ (yellow) and the corresponding assigned candidate genes (red). The enrichment for each C2 term was performed using a hypergeometric test. Terms were grouped (in dashed lines) for UVC light modulated transcription studies in fibroblasts (da Costa et al., 2005; Gentile et al., 2003; Kyng et al., 2005). Hox genes and Hox related gene sets (da Costa et al., 2005; Gentile et al., 2003; Kyng et al., 2005). Other enriched gene sets are related to Cytomegalovirus (CMV) and Reovirus infection response, cell communication, autophagy, signalling and cell growth.
the student t-test (Boorsma et al., 2005) and the Wilcoxon rank test (Barry et al., 2005). Recently introduced methods, the so called gene set enrichment analysis (GSEA) method (Mootha et al., 2003) has the advantage not to require a predefined candidate and reference gene list. The method computes a summary statistic for each gene set from the weighting measures that are assigned to each gene. The null distribution of the summary statistic is computed for each term repeatedly from random permutations of the score assignments of all individual genes. The statistics commonly used are rank statistics, e.g., Kolmogorov-Smirnov score (Subramanian et al., 2005) or summary statistics (Tian et al., 2005; Jiang and Gentleman, 2007). The latter described methods are appropriate when only subtle changes are expected in an experiment, e.g., by a coordinated change of a geneset that are not detected with the conventional methods, such as by a hyper-geometric test (Mootha et al., 2003).
Chapter 8

Outlook

In this thesis a set of candidate genes has been identified showing species-specific substitution patterns. It was shown that the underlying changes in the substitution patterns agree with what would be expected when the genes are subject to species-specific extents of transcription-coupled-repair. This implies that these genes are presumably differentially expressed in the germline of humans and chimpanzees. Interestingly, a Gene Ontology enrichment analysis revealed that the candidate genes play a role in major developmental processes, which are capable to account for different anatomical and behavioral phenotypes of humans and chimpanzees. The proposed gene expression changes are in-silico predictions and are difficult to confirm by experimental analyses. An experimental evaluation of differential gene expression during early embryonic developmental stages between humans and chimpanzees is not feasible due to ethical reasons as embryonic tissue samples would be required. For human and chimpanzee only gene expression data from adult tissues are available. It will therefore remain difficult to find strong correlations of strand-specific substitution patterns detected in this study to actual gene expression levels. Choosing species for an analysis as described here that are more accessible to experimental research, e.g., the species pair Mus Musculus and Mus spretus (Dejager et al., 2009) would facilitate a wet-lab analysis to confirm in-silico predicted gene expression changes. In mammalian species the transcription signature in the substitution patterns of transcribed genomic regions is expected to be same (Green, 2003). Mus musculus and Mus spretus for example have a similar sequence divergence compared to the sequence divergence between humans and chimpanzees (Dejager et al., 2009). The analysis to test for the homogeneity of the substitution patterns could be conducted between Mus musculus and Mus spretus using Rattus norvegicus as outgroup.
However, this analysis is not possible since the genome sequence of Mus spretus is not yet available.

For the use of other eukaryotic model organisms it would be first necessary to elucidate the explicit substitution patterns in transcribed and non-transcribed genomic regions. The observed strand asymmetries in transcribed genomic regions are likely to differ between vertebrates, invertebrates and plant species (Touchon et al. 2004). This could be performed by a genome-wide analysis as described in chapter 4 for different eukaryotic species using the Likelihood-Ratio-Test for strand symmetry.

The factors driven by transcription-related processes that affect genomic DNA sequences are also used for the identification of novel transcribed genomic regions in the human genome (Glusman et al. 2006). In Glusman et al. (2006), "transcription footprints" are detected by scoring for strand asymmetry measured by G+T excess, substitution patterns between interspersed repeats, repeat orientations and polyadenylation sites along the human genome. The modular approach from Glusman et al. (2006) could also consider branch-specific substitution patterns estimated from a multiple genome alignment to complement and improve predictions of genomic regions that are transcribed in the germline.

The branch-specific substitution model described by (Baake, 1998) that is used in the Single-Branch-Test assumes reversibility of the substitution process. This assumption does not allow to distinguish between the $A \rightarrow G$ and $G \rightarrow A$ rate and between the $T \rightarrow C$ and $C \rightarrow T$ rate. Only few studies consider the use of non-reversible branch-specific (non-homogeneous) substitution models (e.g., Lake 1997, Barry and Hartigan 1987). In Chang (1996) an example is shown, where different sets of substitution matrices for the branches of a tree can correspond to the same observed sequences in an alignment. This might be a problem for methods that require a closed-form solution (Lake 1997) or for iterative optimization methods that converge to local optima (Barry and Hartigan 1987). In Oscamou et al. (2008) the error in the estimation of non-reversible non-homogeneous substitution models was evaluated and shown to allow accurate estimates. The error was measured by the mean distance between the true substitution matrix used for a simulation of 3 taxon alignments and the inferred substitution matrices. However, the accuracy of such methods could also be measured by estimating the variations between branches from simulated data generated using a homogeneous compared to non-homogeneous substitution model. A variety of statistics can be applied to
measure the differences of the substitution patterns when more than two branches are considered (Weiss and von Haeseler 2003; Hamady et al. 2006). Further evaluations might provide a better understanding when such methods to estimate non-reversible non-homogeneous models give accurate estimates and when they fail.

It will be a challenge to identify the explicit factors that lead to differential gene expression patterns that explain phenotypic variations between different species. Gene expression differences might result from changes in the regulatory stages of gene expression starting from the regulation of DNA transcription to the regulation of post-translational processes. Sequence changes in regulatory genomic regions such as gene promoters, silencers, insulators and enhancers might alter binding specificity of transcription-factors and modify the control of transcriptional initiation, enhancement or repression (Donaldson and Gottgens 2006). Differences in the RNA processing can alter the stability of the transcribed mRNA (e.g., increased by polyadenylation) and processing of alternative splice forms (Calarco et al. 2007). Another factor is changes of DNA methylation patterns, where transcription is generally repressed in genomic regions that are methylated (Irvine et al. 2002). Changes on the chromatin structure e.g., via histone modifications or chromatin-remodeling enzymes alter the accessibility of the DNA for transcriptional activity (Li et al. 2007). The positioning of chromosomes in the nuclear matrix can also influence transcriptional activity of genomic regions (Lanctot et al. 2007). Another factor is genomic rearrangements that can cause a change of the genomic neighboring environment of a gene. This might have a direct or indirect influence on the expression patterns of a gene (Marques-Bonet et al. 2004). Gene duplications and gene losses observed between or within mammalian species lead to gene copy number variations (CNV) (Perry et al. 2006). CNVs are suggested to affect gene expression by gene dosage effects, by affecting the chromatin structure or changes of the neighboring genomic regulatory sequences (Kleinjan and van Heyningen 2005). Another suggested regulatory mechanism for transcription is processes involved in stalling or pausing the RNA polymerase at the stage of early elongation (Tamkun 2007). However, the explicit factors that would explain differential gene expression patterns between humans and chimpanzees are speculative and unclear to date.

The substitutional changes derived from transcription-coupled-repair denoted as transcription signature are thought to be neutral and assumed to reflect the strength of gene expression. The identified set of candidate genes that showed a different extend of the
transcription signature between human and chimpanzee might affect changes in the reg-
ulatory network of the metabolome, transcriptome and proteome. It would therefore be
useful to study the role of the candidate genes in the protein-protein interaction network
to identify new key regulators of developmental processes. This might help to find new
signaling pathways during development and might help to understand related human
diseases leading to e.g., developmental anomalies, intellectual impairment or growth
retardation.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A.,
Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis,
A., Lewis, S., Matsue, J., Richardson, J., Ringwald, M., Rubin, G. and Sherlock,

Baake, E. (1998) What can and what cannot be inferred from pairwise sequence com-

Applications to the Natural Sciences*. John Wiley and Sons.


in functional gene categories using continuous measures of gene expression. *Bioinfor-


da Costa, R., Riou, L., Paquola, A., Menck, C. and Sarasin, A. (2005) Transcriptional profiles of unirradiated or UV-irradiated human cells expressing either the cancer-
prone XPB/CS allele or the noncancer-prone XPB/TTD allele. *Oncogene*, 24, 1359–74.


Acknowledgement

This thesis would not have been possible without the encouragement and constant support from Arndt von Haeseler and Ingo Ebersberger. I owe my deepest gratitude to both of you for supervising this project from the very beginning to present time. Thank you for all the fruitful discussions and patience during my Phd. Thank you (especially Ingo Ebersberger) for carefully reading and commenting the chapters and figures and for all your support as supervisor(s) and friend(s). I also want to especially thank Tamara Zoranovic and Josef Penninger that supported the development of the gene ontology enrichment analysis pipeline presented in chapter 7. Thank you Tamara for all the motivation, your patience and all the precious discussions. Also i would like to thank all my other collaborators, like Federica Catalanotti, Manuela Baccarini, Shane Cronin and Yumiko Imai. You sharpened my view on the molecular level. So many people I need to thank at this place. I am sorry that I may not be able to name all of you here. I am indebted to all my colleagues in the CIBIV and the Düsseldorf group, especially to my closest colleagues Tanja Gesell and Minh Bui Quang, which also have read and commented chapter 3. Thank you Tanja for all our conversations and your support. Thank you Martin Grabner for all the discussion we had and all advices about teaching. It was an invaluable experience to be a teaching assistant in your courses that you gave. Thank you Matthias Dehmer for your support and all fruitful discussions. Thank you Heiko Schmidt, Martin Grabner and Wolfgang Fischl, which are responsible for the outstanding computer and backup system at the CIBIV. I also want to thank all visitors of the CIBIV and numerous people I met that discussed my work with me. Thank you Christian Schlötterer for the retreats in the mountains with your group, for very helpful comments on a talk I gave and for providing the opportunity to talk with some of your invited speakers. Also I want to thank the MFPL, IMP and IMBA, who make this international environment possible around the Campus by inviting numerous cutting-edge speakers in their seminars. Thank you also to the WWTF and DFG for financial support. Thank you to all my dear friends. Thank you Sonia, Pedro, Joaquim and Susete for your patience, personal support and your presence.
### Appendix A

#### Gene Ontology Enrichment Analysis - Supplement

**A.1 Gene Ontology Biological Process**

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003002</td>
<td>regionalization</td>
<td>128</td>
<td>23</td>
<td>7.40</td>
<td>0.000001</td>
</tr>
<tr>
<td>GO:0009932</td>
<td>anterior/posterior pattern formation</td>
<td>93</td>
<td>19</td>
<td>5.37</td>
<td>0.000001</td>
</tr>
<tr>
<td>GO:0007389</td>
<td>pattern specification process</td>
<td>177</td>
<td>25</td>
<td>10.23</td>
<td>0.000030</td>
</tr>
<tr>
<td>GO:0048706</td>
<td>embryonic skeletal development</td>
<td>52</td>
<td>12</td>
<td>3.01</td>
<td>0.000031</td>
</tr>
<tr>
<td>GO:0001501</td>
<td>skeletal development</td>
<td>171</td>
<td>22</td>
<td>9.88</td>
<td>0.000350</td>
</tr>
<tr>
<td>GO:0011110</td>
<td>positive regulation of cellular component organization and biogenesis</td>
<td>126</td>
<td>14</td>
<td>4.97</td>
<td>0.000380</td>
</tr>
<tr>
<td>GO:0001505</td>
<td>anatomical structure morphogenesis</td>
<td>908</td>
<td>82</td>
<td>57.10</td>
<td>0.000490</td>
</tr>
<tr>
<td>GO:0001757</td>
<td>intra-S DNA damage checkpoint</td>
<td>4</td>
<td>3</td>
<td>0.23</td>
<td>0.000740</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>anatomical structure development</td>
<td>1931</td>
<td>142</td>
<td>111.59</td>
<td>0.001060</td>
</tr>
<tr>
<td>GO:0001702</td>
<td>gastrulation with mouth forming second</td>
<td>10</td>
<td>4</td>
<td>0.58</td>
<td>0.001750</td>
</tr>
<tr>
<td>GO:0025902</td>
<td>developmental process</td>
<td>2980</td>
<td>204</td>
<td>172.21</td>
<td>0.002950</td>
</tr>
<tr>
<td>GO:0007547</td>
<td>parturition</td>
<td>6</td>
<td>3</td>
<td>0.35</td>
<td>0.003370</td>
</tr>
<tr>
<td>GO:0008088</td>
<td>positive regulation of actin filament polymerization</td>
<td>6</td>
<td>3</td>
<td>0.35</td>
<td>0.003370</td>
</tr>
<tr>
<td>GO:0045109</td>
<td>intermediate filament organization</td>
<td>6</td>
<td>3</td>
<td>0.35</td>
<td>0.003370</td>
</tr>
<tr>
<td>GO:0045104</td>
<td>intermediate filament cytoskeleton organization and biogenesis</td>
<td>12</td>
<td>4</td>
<td>0.69</td>
<td>0.003770</td>
</tr>
<tr>
<td>GO:0048566</td>
<td>embryonic gut development</td>
<td>12</td>
<td>4</td>
<td>0.69</td>
<td>0.003770</td>
</tr>
<tr>
<td>GO:0048705</td>
<td>skeletal morphogenesis</td>
<td>45</td>
<td>8</td>
<td>2.60</td>
<td>0.003840</td>
</tr>
<tr>
<td>GO:0050794</td>
<td>regulation of cellular process</td>
<td>5993</td>
<td>382</td>
<td>346.34</td>
<td>0.004030</td>
</tr>
<tr>
<td>GO:0048704</td>
<td>embryonic skeletal morphogenesis</td>
<td>37</td>
<td>7</td>
<td>2.14</td>
<td>0.004710</td>
</tr>
<tr>
<td>GO:0001739</td>
<td>induction of an organ</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005640</td>
</tr>
<tr>
<td>GO:0007419</td>
<td>ectodermal gut development</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005640</td>
</tr>
<tr>
<td>GO:0000992</td>
<td>specification of organ identity</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005640</td>
</tr>
<tr>
<td>GO:0031299</td>
<td>retinal ganglion cell axon guidance</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005640</td>
</tr>
<tr>
<td>GO:0048567</td>
<td>ectodermal gut morphogenesis</td>
<td>21</td>
<td>5</td>
<td>1.21</td>
<td>0.005960</td>
</tr>
<tr>
<td>Gene Ontology ID</td>
<td>Description</td>
<td>Count</td>
<td>Unique Count</td>
<td>P-Value</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------</td>
<td>-------</td>
<td>--------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>GO:0043687</td>
<td>post-translational protein modification</td>
<td>1173</td>
<td>88</td>
<td>0.006900</td>
<td></td>
</tr>
<tr>
<td>GO:0048513</td>
<td>organ development</td>
<td>1286</td>
<td>95</td>
<td>0.007680</td>
<td></td>
</tr>
<tr>
<td>GO:0007275</td>
<td>multicellular organismal development</td>
<td>2176</td>
<td>151</td>
<td>0.007240</td>
<td></td>
</tr>
<tr>
<td>GO:0007420</td>
<td>brain development</td>
<td>166</td>
<td>18</td>
<td>0.007450</td>
<td></td>
</tr>
<tr>
<td>GO:0048519</td>
<td>regulation of biological process</td>
<td>6815</td>
<td>390</td>
<td>0.007890</td>
<td></td>
</tr>
<tr>
<td>GO:0007420</td>
<td>induction</td>
<td>8</td>
<td>3</td>
<td>0.006560</td>
<td></td>
</tr>
<tr>
<td>GO:0006479</td>
<td>cell-cell signaling during cell fate commit</td>
<td>8</td>
<td>3</td>
<td>0.006560</td>
<td></td>
</tr>
<tr>
<td>GO:0008213</td>
<td>protein amino acid methyltransferase</td>
<td>23</td>
<td>5</td>
<td>0.009690</td>
<td></td>
</tr>
<tr>
<td>GO:0006183</td>
<td>protein amino acid alkylation</td>
<td>23</td>
<td>5</td>
<td>0.009690</td>
<td></td>
</tr>
<tr>
<td>GO:0032273</td>
<td>positive regulation of protein polymerization</td>
<td>15</td>
<td>4</td>
<td>0.009500</td>
<td></td>
</tr>
<tr>
<td>GO:0007154</td>
<td>cell communication</td>
<td>3701</td>
<td>243</td>
<td>0.009100</td>
<td></td>
</tr>
<tr>
<td>GO:0007205</td>
<td>activation of protein kinase C activity</td>
<td>26</td>
<td>5</td>
<td>0.013900</td>
<td></td>
</tr>
<tr>
<td>GO:0007204</td>
<td>regulation of cellular component organization and biogenesis</td>
<td>114</td>
<td>13</td>
<td>0.016390</td>
<td></td>
</tr>
<tr>
<td>GO:0009218</td>
<td>pyrimidine ribonucleotide biosynthetic process</td>
<td>16</td>
<td>4</td>
<td>0.015200</td>
<td></td>
</tr>
<tr>
<td>Gene Ontology Biological Process</td>
<td>Count</td>
<td>Pearson's r</td>
<td>FDR Correction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------</td>
<td>-------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016570 histone modification</td>
<td>49</td>
<td>2.83</td>
<td>0.021630</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006241 CTP biosynthetic process</td>
<td>11</td>
<td>0.64</td>
<td>0.022170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008105 asymmetric protein localization</td>
<td>14</td>
<td>0.64</td>
<td>0.022370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009208 pyrimidine ribonucleoside triphosphate metabolic process</td>
<td>11</td>
<td>0.64</td>
<td>0.022370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009209 pyrimidine ribonucleoside triphosphate biosynthetic process</td>
<td>11</td>
<td>0.64</td>
<td>0.022370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0013144 regulation of cell projection organization and biogenesis</td>
<td>11</td>
<td>0.64</td>
<td>0.022170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0013146 positive regulation of cell projection organization and biogenesis</td>
<td>11</td>
<td>0.64</td>
<td>0.022170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0046036 CTP metabolic process</td>
<td>11</td>
<td>0.64</td>
<td>0.022370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0046039 GTP metabolic process</td>
<td>11</td>
<td>0.64</td>
<td>0.022370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0000902 cell morphogenesis</td>
<td>355</td>
<td>20.52</td>
<td>0.023580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0032989 cellular structure morphogenesis</td>
<td>355</td>
<td>20.52</td>
<td>0.023580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006468 protein amino acid phosphorylation</td>
<td>728</td>
<td>42.07</td>
<td>0.024220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006874 cellular calcium ion homeostasis</td>
<td>101</td>
<td>5.84</td>
<td>0.031220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009887 organ morphogenesis</td>
<td>488</td>
<td>25.49</td>
<td>0.028030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0032383 biopolymer metabolic process</td>
<td>497</td>
<td>287.74</td>
<td>0.028100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016575 histone deacetylation</td>
<td>12</td>
<td>0.69</td>
<td>0.026570</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0032269 negative regulation of cellular protein metabolic process</td>
<td>98</td>
<td>5.66</td>
<td>0.025670</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0017148 negative regulation of translation</td>
<td>22</td>
<td>1.27</td>
<td>0.035220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005074 calcium ion homeostasis</td>
<td>103</td>
<td>5.95</td>
<td>0.035360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0020208 neurogenesis</td>
<td>324</td>
<td>18.72</td>
<td>0.035550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0091919 ribonucleoside metabolic process</td>
<td>13</td>
<td>0.75</td>
<td>0.035580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045740 positive regulation of DNA replication</td>
<td>13</td>
<td>0.75</td>
<td>0.035580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048265 response to pain</td>
<td>13</td>
<td>0.75</td>
<td>0.035580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048547 gut morphogenesis</td>
<td>13</td>
<td>0.75</td>
<td>0.035580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007611 learning and/or memory</td>
<td>66</td>
<td>3.81</td>
<td>0.035640</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006275 regulation of DNA replication</td>
<td>43</td>
<td>2.48</td>
<td>0.035900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007398 ectoderm development</td>
<td>143</td>
<td>8.26</td>
<td>0.037020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048667 neuron morphogenesis during differentiation</td>
<td>143</td>
<td>8.26</td>
<td>0.037020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048812 neurite morphogenesis</td>
<td>143</td>
<td>8.26</td>
<td>0.037020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0049228 cell motility</td>
<td>443</td>
<td>25.60</td>
<td>0.037060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0051674 localization of cell</td>
<td>443</td>
<td>25.60</td>
<td>0.037060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0022604 regulation of cell morphogenesis</td>
<td>55</td>
<td>3.18</td>
<td>0.038040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045107 appendage morphogenesis</td>
<td>55</td>
<td>3.18</td>
<td>0.038040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0035108 limb morphogenesis</td>
<td>55</td>
<td>3.18</td>
<td>0.038040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009044 cellular morphogenesis during differentiation</td>
<td>157</td>
<td>9.07</td>
<td>0.038090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0032147 activation of protein kinase activity</td>
<td>67</td>
<td>3.87</td>
<td>0.038410</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045860 positive regulation of protein kinase activity</td>
<td>144</td>
<td>8.32</td>
<td>0.038920</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0001001 glial cell differentiation</td>
<td>33</td>
<td>1.91</td>
<td>0.039380</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO Term ID</td>
<td>GO Description</td>
<td>Count</td>
<td>Term Count</td>
<td>FDR</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>GO:0031327</td>
<td>negative regulation of cellular biosynthetic process</td>
<td>33</td>
<td>5</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>GO:0011570</td>
<td>DNA integrity checkpoint</td>
<td>33</td>
<td>5</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>GO:0048568</td>
<td>embryonic organ development</td>
<td>33</td>
<td>5</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>GO:0000030</td>
<td>cell projection organization and biogenesis</td>
<td>270</td>
<td>23</td>
<td>15.60</td>
<td></td>
</tr>
<tr>
<td>GO:00006221</td>
<td>pyrimidine nucleotide biosynthetic process</td>
<td>23</td>
<td>4</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>GO:00541495</td>
<td>positive regulation of cytoskeleton organization and biogenesis</td>
<td>23</td>
<td>4</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>GO:0007409</td>
<td>axonogenesis</td>
<td>132</td>
<td>13</td>
<td>7.63</td>
<td></td>
</tr>
<tr>
<td>GO:0032090</td>
<td>cell part morphogenesis</td>
<td>271</td>
<td>23</td>
<td>15.66</td>
<td></td>
</tr>
<tr>
<td>GO:00448885</td>
<td>cell projection morphogenesis</td>
<td>271</td>
<td>23</td>
<td>15.66</td>
<td></td>
</tr>
<tr>
<td>GO:0038674</td>
<td>positive regulation of kinase activity</td>
<td>146</td>
<td>14</td>
<td>8.44</td>
<td></td>
</tr>
<tr>
<td>GO:0019932</td>
<td>second-messenger-mediated signaling</td>
<td>215</td>
<td>19</td>
<td>12.42</td>
<td></td>
</tr>
<tr>
<td>GO:0001708</td>
<td>cell fate specification</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:0009147</td>
<td>pyrimidine nucleotide triphosphate metabolic process</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:0016202</td>
<td>regulation of striated muscle development</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:0030968</td>
<td>endoplasmic reticulum unfolded protein response</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:00348646</td>
<td>digestive tract morphogenesis</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:00486434</td>
<td>regulation of muscle development</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:0020920</td>
<td>regulation of chemotaxis</td>
<td>14</td>
<td>3</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>GO:00008185</td>
<td>calcium ion transport</td>
<td>160</td>
<td>15</td>
<td>9.25</td>
<td></td>
</tr>
<tr>
<td>GO:0032259</td>
<td>methylation</td>
<td>69</td>
<td>8</td>
<td>3.99</td>
<td></td>
</tr>
<tr>
<td>GO:0010659</td>
<td>negative regulation of organelle organization and biogenesis</td>
<td>57</td>
<td>7</td>
<td>3.29</td>
<td></td>
</tr>
<tr>
<td>GO:001605</td>
<td>Wnt receptor signaling pathway</td>
<td>121</td>
<td>12</td>
<td>6.99</td>
<td></td>
</tr>
<tr>
<td>GO:00434142</td>
<td>biopolymer modification</td>
<td>1539</td>
<td>104</td>
<td>88.94</td>
<td></td>
</tr>
<tr>
<td>GO:0000875</td>
<td>cellular metal ion homeostasis</td>
<td>108</td>
<td>11</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td>GO:0032011</td>
<td>ARF protein signal transduction</td>
<td>46</td>
<td>6</td>
<td>2.66</td>
<td></td>
</tr>
<tr>
<td>GO:0032012</td>
<td>regulation of ARF protein signal transduction</td>
<td>46</td>
<td>6</td>
<td>2.66</td>
<td></td>
</tr>
<tr>
<td>GO:0005673</td>
<td>epithelial cell proliferation</td>
<td>46</td>
<td>6</td>
<td>2.66</td>
<td></td>
</tr>
<tr>
<td>GO:0033123</td>
<td>regulation of cellular metabolic process</td>
<td>2739</td>
<td>177</td>
<td>158.29</td>
<td></td>
</tr>
<tr>
<td>GO:00009890</td>
<td>negative regulation of biosynthetic process</td>
<td>363</td>
<td>29</td>
<td>20.98</td>
<td></td>
</tr>
<tr>
<td>GO:0018958</td>
<td>phenol metabolic process</td>
<td>70</td>
<td>8</td>
<td>4.05</td>
<td></td>
</tr>
<tr>
<td>GO:0037175</td>
<td>epidermal growth factor receptor signaling pathway</td>
<td>35</td>
<td>5</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>GO:0010558</td>
<td>negative regulation of macromolecule biosynthetic process</td>
<td>349</td>
<td>28</td>
<td>20.17</td>
<td></td>
</tr>
</tbody>
</table>

A.1 Gene Ontology Biological Process
## A.2 Gene Ontology Molecular Function

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0019955</td>
<td>cytokine binding</td>
<td>110</td>
<td>15</td>
<td>6.27</td>
<td>0.001400</td>
</tr>
<tr>
<td>GO:0004039</td>
<td>beta-adrenergic receptor activity</td>
<td>5</td>
<td>3</td>
<td>0.28</td>
<td>0.001700</td>
</tr>
<tr>
<td>GO:000406</td>
<td>carbamoyl-phosphate synthase activity</td>
<td>6</td>
<td>3</td>
<td>0.34</td>
<td>0.003200</td>
</tr>
<tr>
<td>GO:0004528</td>
<td>transcription regulator activity</td>
<td>1326</td>
<td>98</td>
<td>75.58</td>
<td>0.004100</td>
</tr>
<tr>
<td>GO:0008195</td>
<td>phosphatidate phosphatase activity</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005400</td>
</tr>
<tr>
<td>GO:0014061</td>
<td>guanyl nucleotide binding</td>
<td>368</td>
<td>33</td>
<td>20.98</td>
<td>0.006700</td>
</tr>
<tr>
<td>GO:0032661</td>
<td>guanyl ribonucleotide binding</td>
<td>368</td>
<td>33</td>
<td>20.98</td>
<td>0.006700</td>
</tr>
<tr>
<td>GO:0000525</td>
<td>GTP binding</td>
<td>360</td>
<td>32</td>
<td>20.52</td>
<td>0.008500</td>
</tr>
<tr>
<td>GO:0004564</td>
<td>transcription repressor activity</td>
<td>216</td>
<td>21</td>
<td>12.31</td>
<td>0.018000</td>
</tr>
<tr>
<td>GO:0004522</td>
<td>pancreatic ribonuclease activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0004550</td>
<td>nucleoside diphosphate kinase activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0004957</td>
<td>prostaglandin E receptor activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0004551</td>
<td>cyclic nucleotide binding</td>
<td>27</td>
<td>5</td>
<td>1.54</td>
<td>0.016800</td>
</tr>
<tr>
<td>GO:0004465</td>
<td>histone deacetylase activity</td>
<td>18</td>
<td>4</td>
<td>1.03</td>
<td>0.016900</td>
</tr>
<tr>
<td>GO:0003558</td>
<td>protein deacetylase activity</td>
<td>18</td>
<td>4</td>
<td>1.03</td>
<td>0.016900</td>
</tr>
<tr>
<td>GO:0015075</td>
<td>ion transmembrane transporter activity</td>
<td>848</td>
<td>63</td>
<td>48.33</td>
<td>0.017800</td>
</tr>
<tr>
<td>GO:0015291</td>
<td>secondary active transmembrane transporter activity</td>
<td>268</td>
<td>24</td>
<td>15.28</td>
<td>0.019000</td>
</tr>
<tr>
<td>GO:0004565</td>
<td>sequence-specific DNA binding</td>
<td>457</td>
<td>37</td>
<td>26.05</td>
<td>0.019800</td>
</tr>
<tr>
<td>GO:0004529</td>
<td>solute:antipporter activity</td>
<td>49</td>
<td>7</td>
<td>2.79</td>
<td>0.020200</td>
</tr>
<tr>
<td>GO:0004687</td>
<td>chaperone binding</td>
<td>19</td>
<td>4</td>
<td>1.08</td>
<td>0.020400</td>
</tr>
<tr>
<td>GO:0004524</td>
<td>solute:antipporter activity</td>
<td>147</td>
<td>15</td>
<td>8.38</td>
<td>0.020500</td>
</tr>
<tr>
<td>GO:0004523</td>
<td>deacetylase activity</td>
<td>20</td>
<td>4</td>
<td>1.14</td>
<td>0.024400</td>
</tr>
<tr>
<td>GO:0004552</td>
<td>cAMP binding</td>
<td>20</td>
<td>4</td>
<td>1.14</td>
<td>0.024400</td>
</tr>
<tr>
<td>GO:0015291</td>
<td>Rho GTPase binding</td>
<td>30</td>
<td>5</td>
<td>1.71</td>
<td>0.025600</td>
</tr>
<tr>
<td>GO:0015156</td>
<td>sul fate transmembrane transporter activity</td>
<td>12</td>
<td>3</td>
<td>0.68</td>
<td>0.027600</td>
</tr>
<tr>
<td>GO:0003255</td>
<td>ribonucleotide binding</td>
<td>1802</td>
<td>121</td>
<td>102.71</td>
<td>0.028300</td>
</tr>
<tr>
<td>GO:0003255</td>
<td>purine ribonucleotide binding</td>
<td>1802</td>
<td>121</td>
<td>102.71</td>
<td>0.028300</td>
</tr>
<tr>
<td>GO:0001559</td>
<td>cation:antipporter activity</td>
<td>21</td>
<td>4</td>
<td>1.20</td>
<td>0.029900</td>
</tr>
<tr>
<td>GO:0002659</td>
<td>substrate-specific transmembrane transporter activity</td>
<td>94</td>
<td>66</td>
<td>52.32</td>
<td>0.029900</td>
</tr>
<tr>
<td>GO:0001559</td>
<td>solute:antipporter activity</td>
<td>53</td>
<td>7</td>
<td>3.02</td>
<td>0.029900</td>
</tr>
<tr>
<td>GO:0005351</td>
<td>sugar:hydrogen symporter activity</td>
<td>42</td>
<td>6</td>
<td>2.39</td>
<td>0.030600</td>
</tr>
<tr>
<td>GO:0005402</td>
<td>cation:sugar symporter activity</td>
<td>42</td>
<td>6</td>
<td>2.39</td>
<td>0.030600</td>
</tr>
<tr>
<td>GO:0004576</td>
<td>purine nucleotide binding</td>
<td>1893</td>
<td>126</td>
<td>107.90</td>
<td>0.032100</td>
</tr>
<tr>
<td>GO:0016493</td>
<td>C-C chemokine receptor activity</td>
<td>32</td>
<td>5</td>
<td>1.82</td>
<td>0.033300</td>
</tr>
<tr>
<td>GO:0016493</td>
<td>C-C chemokine binding</td>
<td>32</td>
<td>5</td>
<td>1.82</td>
<td>0.033300</td>
</tr>
<tr>
<td>GO:0016493</td>
<td>carbohydrate transmembrane transporter activity</td>
<td>43</td>
<td>6</td>
<td>2.45</td>
<td>0.033900</td>
</tr>
<tr>
<td>GO:0016493</td>
<td>sugar transmembrane transporter activity</td>
<td>43</td>
<td>6</td>
<td>2.45</td>
<td>0.033900</td>
</tr>
<tr>
<td>GO:0004576</td>
<td>icosanoid receptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO:0004576</td>
<td>prostaglandin receptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO:0008093</td>
<td>cytoskeletal adaptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO ID</td>
<td>GO Term</td>
<td>Annotated</td>
<td>Significant</td>
<td>Expected</td>
<td>P-value</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------------------------------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>GO:0019955</td>
<td>cytokine binding</td>
<td>110</td>
<td>15</td>
<td>6.27</td>
<td>0.001400</td>
</tr>
<tr>
<td>GO:0004939</td>
<td>beta-adrenergic receptor activity</td>
<td>5</td>
<td>3</td>
<td>0.28</td>
<td>0.001700</td>
</tr>
<tr>
<td>GO:0004086</td>
<td>carbamoyl-phosphate synthase activity</td>
<td>6</td>
<td>3</td>
<td>0.34</td>
<td>0.002000</td>
</tr>
<tr>
<td>GO:0030528</td>
<td>transcription regulator activity</td>
<td>1326</td>
<td>98</td>
<td>75.58</td>
<td>0.004100</td>
</tr>
<tr>
<td>GO:0008195</td>
<td>phosphatidate phosphatase activity</td>
<td>7</td>
<td>3</td>
<td>0.40</td>
<td>0.005400</td>
</tr>
<tr>
<td>GO:0019001</td>
<td>guanyl nucleotide binding</td>
<td>368</td>
<td>33</td>
<td>0.57</td>
<td>0.006600</td>
</tr>
<tr>
<td>GO:0032561</td>
<td>guanyl ribonucleotide binding</td>
<td>368</td>
<td>33</td>
<td>0.57</td>
<td>0.006600</td>
</tr>
<tr>
<td>GO:0005525</td>
<td>GTP binding</td>
<td>360</td>
<td>32</td>
<td>20.52</td>
<td>0.008500</td>
</tr>
<tr>
<td>GO:0016564</td>
<td>transcription repressor activity</td>
<td>216</td>
<td>21</td>
<td>12.31</td>
<td>0.011800</td>
</tr>
<tr>
<td>GO:0004522</td>
<td>pancreatic ribonuclease activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0004550</td>
<td>nucleoside diphosphate kinase activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0004957</td>
<td>prostaglandin E receptor activity</td>
<td>10</td>
<td>3</td>
<td>0.57</td>
<td>0.016400</td>
</tr>
<tr>
<td>GO:0030551</td>
<td>cyclic nucleotide binding</td>
<td>27</td>
<td>5</td>
<td>1.54</td>
<td>0.016800</td>
</tr>
<tr>
<td>GO:0004407</td>
<td>histone deacetylase activity</td>
<td>18</td>
<td>4</td>
<td>1.03</td>
<td>0.020100</td>
</tr>
<tr>
<td>GO:0033558</td>
<td>protein deacetylase activity</td>
<td>18</td>
<td>4</td>
<td>1.03</td>
<td>0.020100</td>
</tr>
<tr>
<td>GO:0015075</td>
<td>ion transmembrane transporter activity</td>
<td>848</td>
<td>63</td>
<td>48.33</td>
<td>0.017800</td>
</tr>
<tr>
<td>GO:0015291</td>
<td>secondary active transmembrane transporter activity</td>
<td>268</td>
<td>28</td>
<td>15.26</td>
<td>0.019000</td>
</tr>
<tr>
<td>GO:0043565</td>
<td>sequence-specific DNA binding</td>
<td>457</td>
<td>37</td>
<td>2.65</td>
<td>0.022100</td>
</tr>
<tr>
<td>GO:0015294</td>
<td>solute:cation symporter activity</td>
<td>147</td>
<td>15</td>
<td>8.38</td>
<td>0.024100</td>
</tr>
<tr>
<td>GO:0019213</td>
<td>deacetylase activity</td>
<td>20</td>
<td>4</td>
<td>1.14</td>
<td>0.024400</td>
</tr>
<tr>
<td>GO:0030552</td>
<td>cAMP binding</td>
<td>20</td>
<td>4</td>
<td>1.14</td>
<td>0.024400</td>
</tr>
<tr>
<td>GO:0017048</td>
<td>Rho GTPase binding</td>
<td>30</td>
<td>5</td>
<td>1.14</td>
<td>0.024400</td>
</tr>
<tr>
<td>GO:0015116</td>
<td>sulfate transmembrane transporter activity</td>
<td>12</td>
<td>3</td>
<td>0.68</td>
<td>0.026900</td>
</tr>
<tr>
<td>GO:0032553</td>
<td>ribonucleotide binding</td>
<td>1802</td>
<td>121</td>
<td>102.71</td>
<td>0.028300</td>
</tr>
<tr>
<td>GO:0032555</td>
<td>purine ribonucleotide binding</td>
<td>1802</td>
<td>121</td>
<td>102.71</td>
<td>0.028300</td>
</tr>
<tr>
<td>GO:0015491</td>
<td>cation:cation antiporter activity</td>
<td>21</td>
<td>4</td>
<td>1.20</td>
<td>0.028900</td>
</tr>
<tr>
<td>GO:0022891</td>
<td>substrate-specific transmembrane transporter activity</td>
<td>918</td>
<td>66</td>
<td>52.32</td>
<td>0.029100</td>
</tr>
<tr>
<td>GO:0015300</td>
<td>solute:solute antiporter activity</td>
<td>53</td>
<td>7</td>
<td>3.02</td>
<td>0.029900</td>
</tr>
<tr>
<td>GO:0005351</td>
<td>sugar:hydrogen symporter activity</td>
<td>42</td>
<td>6</td>
<td>2.39</td>
<td>0.030600</td>
</tr>
<tr>
<td>GO:0005402</td>
<td>cation:sugar symporter activity</td>
<td>42</td>
<td>6</td>
<td>2.39</td>
<td>0.030600</td>
</tr>
<tr>
<td>GO:0017076</td>
<td>purine nucleotide binding</td>
<td>1893</td>
<td>126</td>
<td>107.90</td>
<td>0.032100</td>
</tr>
<tr>
<td>GO:0016493</td>
<td>C-C chemokine receptor activity</td>
<td>32</td>
<td>5</td>
<td>1.82</td>
<td>0.033300</td>
</tr>
<tr>
<td>GO:0019957</td>
<td>C-C chemokine binding</td>
<td>32</td>
<td>5</td>
<td>1.82</td>
<td>0.033300</td>
</tr>
<tr>
<td>GO:0015144</td>
<td>carbohydrate transmembrane transporter activity</td>
<td>43</td>
<td>7</td>
<td>3.02</td>
<td>0.033900</td>
</tr>
<tr>
<td>GO:0051119</td>
<td>sugar transporter activity</td>
<td>43</td>
<td>7</td>
<td>3.02</td>
<td>0.033900</td>
</tr>
<tr>
<td>GO:0004953</td>
<td>icosanoid receptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO:0004954</td>
<td>prostanoid receptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO:0004955</td>
<td>prostaglandin receptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
<tr>
<td>GO:0008093</td>
<td>cytoskeletal adaptor activity</td>
<td>13</td>
<td>3</td>
<td>0.74</td>
<td>0.034400</td>
</tr>
</tbody>
</table>

A.3 Gene Ontology Cellular Component
## Appendix B

### C2 Dataset Enrichment Analysis - Supplement

<table>
<thead>
<tr>
<th>C2 Term (Curated Gene Set)</th>
<th>Fraction (Significant)</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>P-value (Hypergeometric)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 UVC_TTD_4HR_DN</td>
<td>0.15</td>
<td>149</td>
<td>23</td>
<td>8.69</td>
<td>0.00001647</td>
<td>Down-regulated at 8 hours following treatment of XPB/TTD fibroblasts with 3J/m² UVC</td>
</tr>
<tr>
<td>2 UVC_XPCS_4HR_DN</td>
<td>0.13</td>
<td>211</td>
<td>28</td>
<td>12.31</td>
<td>0.00003746</td>
<td>Down-regulated at 4 hours following treatment of XPB/CS fibroblasts with 3J/m² UVC</td>
</tr>
<tr>
<td>3 HOX_GENES</td>
<td>0.24</td>
<td>46</td>
<td>11</td>
<td>2.68</td>
<td>0.00005008</td>
<td>HOX genes related to hematopoiesis</td>
</tr>
<tr>
<td>4 UVC_TTD_ALL_DN</td>
<td>0.11</td>
<td>325</td>
<td>35</td>
<td>18.96</td>
<td>0.00032400</td>
<td>Down-regulated at any timepoint following treatment of XPB/TTD fibroblasts with 3J/m² UVC</td>
</tr>
<tr>
<td>5 UVC_XPCS_ALL_DN</td>
<td>0.10</td>
<td>428</td>
<td>42</td>
<td>24.97</td>
<td>0.00062190</td>
<td>Down-regulated at any timepoint following treatment of XPB/CS fibroblasts with 3J/m² UVC</td>
</tr>
<tr>
<td>6 WANG_HOXA9_VS_MEIS1_UP</td>
<td>0.27</td>
<td>22</td>
<td>6</td>
<td>1.28</td>
<td>0.00128600</td>
<td>Genomic signature of progenitors immortalized by Hoxa9 versus Hoxa9 plus Meis1 Increased expression in Hox9-immortalized progenitors</td>
</tr>
<tr>
<td>7 HEMATOPSTEM_ALL_UP</td>
<td>0.23</td>
<td>31</td>
<td>7</td>
<td>1.81</td>
<td>0.00171200</td>
<td>Up-regulated in populations of human hematopoietic stem cells (CD34+/CD38−/Lin−) from bone marrow, umbilical cord blood, and peripheral blood stem-progenitor cells, compared to the stem cell-depleted population (CD34+/CD38/Lin++)</td>
</tr>
<tr>
<td>8 AGUIRRE_PANCREAS,CHR8</td>
<td>0.18</td>
<td>49</td>
<td>9</td>
<td>2.86</td>
<td>0.00186000</td>
<td>Genes on chromosome 8 with copy-number-driven expression in pancreatic adenocarcinoma</td>
</tr>
<tr>
<td>9 UVC_TTD_4HR_DN</td>
<td>0.10</td>
<td>269</td>
<td>28</td>
<td>15.70</td>
<td>0.00207500</td>
<td>Down-regulated at 4 hours following treatment of XPB/TTD fibroblasts with 3J/m² UVC</td>
</tr>
<tr>
<td>10 ET743PT650_COLONCA_DN</td>
<td>0.19</td>
<td>43</td>
<td>8</td>
<td>2.51</td>
<td>0.00302900</td>
<td>Downregulated by both Et-743 and Pt-650 in HCT116 cells (Fig. 6 A)</td>
</tr>
<tr>
<td>11 HSA04140_REGULATION_OF_AUTOPHAGY</td>
<td>0.24</td>
<td>21</td>
<td>5</td>
<td>1.23</td>
<td>0.00620000</td>
<td>Genes involved in regulation of autophagy</td>
</tr>
<tr>
<td>12 CMV_HCML_TIMECOURSE_4HR_DN</td>
<td>0.20</td>
<td>30</td>
<td>6</td>
<td>1.75</td>
<td>0.00687300</td>
<td>Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consecutive timepoints), with maximum change at 4 hours</td>
</tr>
<tr>
<td>13 S1GPIP3_SIGNALING_IN_B_LYMPHOCYTES</td>
<td>0.19</td>
<td>32</td>
<td>6</td>
<td>1.87</td>
<td>0.00950300</td>
<td>Genes related to PIP3 signaling in B lymphocytes</td>
</tr>
<tr>
<td>Gene(s)</td>
<td>Enrichment</td>
<td>p-value</td>
<td>Odds Ratio</td>
<td>Genes expression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------</td>
<td>---------</td>
<td>------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapamycin Upregulated</td>
<td>0.11</td>
<td>133</td>
<td>15</td>
<td>7.76</td>
<td>Genes upregulated in response to rapamycin starvation</td>
<td></td>
</tr>
<tr>
<td>HELP Type Cells Downregulated</td>
<td>0.25</td>
<td>16</td>
<td>4</td>
<td>0.93</td>
<td>Genes downregulated in human germinal center helper-T (Th) cells versus other CD4+ T-cell types such as naive T cells (from which the others originate), CXCR5+ CCR7+ central memory cells, and CCR7- effector memory cells.</td>
<td></td>
</tr>
<tr>
<td>HSC, HSC Fetal</td>
<td>0.10</td>
<td>199</td>
<td>20</td>
<td>11.61</td>
<td>Up-regulated in mouse hematopoietic stem cells from fetal liver (HSC Shared + Fetal)</td>
<td></td>
</tr>
<tr>
<td>HSC, LT HSC Fetal</td>
<td>0.10</td>
<td>226</td>
<td>22</td>
<td>13.19</td>
<td>Up-regulated in mouse long-term functional hematopoietic stem cells from fetal liver (LT-HSC Shared)</td>
<td></td>
</tr>
<tr>
<td>HSC, LT HSC Shared</td>
<td>0.10</td>
<td>226</td>
<td>22</td>
<td>13.19</td>
<td>Up-regulated in mouse long-term functional hematopoietic stem cells from both adult bone marrow and fetal liver (Cluster i, LT-HSC Shared)</td>
<td></td>
</tr>
<tr>
<td>Takeda, NUP8, HOXA9 16D Downregulated</td>
<td>0.10</td>
<td>174</td>
<td>18</td>
<td>10.15</td>
<td>Effect of NUP98-HOXA9 on gene transcription at 16 d after transduction Down</td>
<td></td>
</tr>
<tr>
<td>HSA01490, Cell Communication</td>
<td>0.12</td>
<td>133</td>
<td>13</td>
<td>6.59</td>
<td>Genes involved in cell communication</td>
<td></td>
</tr>
<tr>
<td>HSC, HSC Shared</td>
<td>0.10</td>
<td>189</td>
<td>19</td>
<td>11.03</td>
<td>Up-regulated in mouse hematopoietic stem cells from both adult bone marrow and fetal liver (Cluster ii, HSC Shared)</td>
<td></td>
</tr>
<tr>
<td>UVC, XPcs, 8HR Downregulated</td>
<td>0.09</td>
<td>367</td>
<td>32</td>
<td>21.41</td>
<td>Down-regulated at 8 hours following treatment of XPB/CS fibroblasts with 3J/m^2 UVC</td>
<td></td>
</tr>
<tr>
<td>CCR5 Pathway</td>
<td>0.30</td>
<td>10</td>
<td>3</td>
<td>0.58</td>
<td>CCR5 is a G-protein coupled receptor expressed in macrophages that recognizes chemokine ligands and is targeted by the HIV envelope protein GP120.</td>
<td></td>
</tr>
<tr>
<td>Takeda, NUP8, HOXA9 8D Downregulated</td>
<td>0.10</td>
<td>167</td>
<td>17</td>
<td>9.74</td>
<td>Effect of NUP98-HOXA9 on gene transcription at 8 d after transduction Down</td>
<td></td>
</tr>
<tr>
<td>Cell, Cycle Regulator</td>
<td>0.22</td>
<td>18</td>
<td>4</td>
<td>1.05</td>
<td>Obsolete by GO - was not defined before being made obsolete</td>
<td></td>
</tr>
<tr>
<td>UVC, Low, C2 Downregulated</td>
<td>0.22</td>
<td>18</td>
<td>4</td>
<td>1.05</td>
<td>Down-regulated at 6-12 hours following treatment of WS1 human skin fibroblasts with UVC at a low dose 10 J/m^2 (cluster c2)</td>
<td></td>
</tr>
<tr>
<td>UVC, TTD-XPCS, Common Downregulated</td>
<td>0.11</td>
<td>131</td>
<td>14</td>
<td>7.64</td>
<td>Down-regulated at any timepoint following treatment of both XPB/CS and XPB/TTD fibroblasts with 3 J/m^2 UVC</td>
<td></td>
</tr>
<tr>
<td>CMV, HCMV Timecourse 16HRS Downregulated</td>
<td>0.21</td>
<td>19</td>
<td>4</td>
<td>1.11</td>
<td>Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consecutive timepoints), with maximum change at 16 hours</td>
<td></td>
</tr>
<tr>
<td>HSA00534, Heparan Sulfate Biosynthesis Downregulated</td>
<td>0.21</td>
<td>19</td>
<td>4</td>
<td>1.11</td>
<td>Genes involved in heparan sulfate biosynthesis</td>
<td></td>
</tr>
<tr>
<td>CMV, HCMV Timecourse All Downregulated</td>
<td>0.09</td>
<td>340</td>
<td>29</td>
<td>19.84</td>
<td>Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consecutive timepoints)</td>
<td></td>
</tr>
<tr>
<td>Hess, HOXA, MEIS1, Downregulated</td>
<td>0.14</td>
<td>51</td>
<td>7</td>
<td>2.98</td>
<td>Genes downregulated in Hoxa9/Meis1 transduced cells vs control</td>
<td></td>
</tr>
<tr>
<td>Hess, HOXA, MEIS1, Upregulated</td>
<td>0.14</td>
<td>51</td>
<td>7</td>
<td>2.98</td>
<td>Genes upregulated in Hoxa9/Meis1 transduced cells vs control</td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>Gene Name</td>
<td>Log2 Ratio</td>
<td>FDR</td>
<td>Affy Change Call</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------------</td>
<td>------------</td>
<td>------</td>
<td>-----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>CALCIUM,REGULATION_IN,CARDIAC,CELLS</td>
<td>0.10</td>
<td>124</td>
<td>13</td>
<td>7.24</td>
<td>0.02865000</td>
</tr>
<tr>
<td>34</td>
<td>DIAB,NEPH,DN</td>
<td>0.08</td>
<td>330</td>
<td>28</td>
<td>19.26</td>
<td>0.02952000</td>
</tr>
<tr>
<td>35</td>
<td>UVC,LOW,ALL,DN</td>
<td>0.13</td>
<td>52</td>
<td>7</td>
<td>3.03</td>
<td>0.03036000</td>
</tr>
<tr>
<td>36</td>
<td>ST,ADRENERGIC</td>
<td>0.16</td>
<td>31</td>
<td>5</td>
<td>1.81</td>
<td>0.03207000</td>
</tr>
<tr>
<td>37</td>
<td>TAKEDA,NUP8,HOXA/Q10,UP</td>
<td>0.10</td>
<td>139</td>
<td>14</td>
<td>8.11</td>
<td>0.03211000</td>
</tr>
<tr>
<td>38</td>
<td>CELL,GROWTH,AND,OR,MAINTENANCE</td>
<td>0.13</td>
<td>53</td>
<td>7</td>
<td>3.09</td>
<td>0.03329000</td>
</tr>
<tr>
<td>39</td>
<td>ST,PHOSPHOINOSITIDE3,KINASE,PATHWAY</td>
<td>0.16</td>
<td>32</td>
<td>5</td>
<td>1.87</td>
<td>0.03625000</td>
</tr>
<tr>
<td>40</td>
<td>CMV,HCMV,TIMECOURSE,8HR,DN</td>
<td>0.23</td>
<td>13</td>
<td>3</td>
<td>0.76</td>
<td>0.03646000</td>
</tr>
<tr>
<td>41</td>
<td>REOVIRUS,HEK293,UP</td>
<td>0.09</td>
<td>196</td>
<td>18</td>
<td>11.44</td>
<td>0.03759000</td>
</tr>
</tbody>
</table>
Appendix C

Zusammenfassung

Appendix D

Curriculum Vitae

Ricardo de Matos Simões

MFPL/CIBIV
Dr. Bohrgasse 9
1030 Vienna
Phone: +43 1 4277 24026
Fax: +43 1 4277 24098
Email: ricardo.de.matos.simoes@univie.ac.at
Homepage: http://www.cibiv.at/~ricardo

Personal Details

Date of Birth: 25.05.1977
Birthplace: Rheydt/Mönchengladbach, Germany
Nationality: Portuguese
Academic Education

2010  Phd Thesis: "Species-specific Evolving Regions in the Human and Chimpanzee Genomes", Supervisor: Prof. Arndt von Haeseler, Dr. Ingo Ebersberger

2006-2010 (WS05/06-SS10)  Phd student (Bioinformatics), University of Vienna, Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL)

2005-2006 (SS05)  Research assistant, Heinrich-Heine-Universitaet Dusseldorf, Institut fuer Bioinformatik (Prof. Arndt von Haeseler), Sequence analysis of chloroplast DNA from non-photosynthetic Monotropaceae plant species, collaboration with Benjamin Killian, Dr. Jens Pahnke, Prof. Dr. William Martin

2005 (WS04/05)  Diploma Thesis "Identifizierung artspezifisch evolvierender Genomregionen im Menschen und Schimpansen", Institut fuer Bioinformatik, Heinrich-Heine-Universitaet Dusseldorf

1999-2005 (WS99-WS04/05)  Study of Biology (Diploma) at the Heinrich-Heine-Universitaet Dusseldorf

Teaching Experience

2010 (WS09/10)  R programming (15 hours), Veterinaermedizinische Universitaet Wien, parallel course to an introductory statistics course for biologist held by Dr. Mareike Fischer

2008 (SS08)  Perl programming lecture (8 hours), Veterinaermedizinische Universitaet Wien, introductory programming lecture with practical exercises

2008 (WS08/09)  Applications in Bioinformatics (8 hours), FH Campus Wien, 2 lectures about alignment methods for biological sequences, methods for phylogenetic analysis, including practical exercises in a unix environment

2007 (WS07/08)  R programming tutorial (8 hours), Veterinaermedizinische Universitaet Wien, lecture and practical exercises in R, parallel course to an introductory statistics course for biologist held by Dr. Steffen Klaere
2007 (WS07/08) Applications in Bioinformatics (8 hours), FH Campus Wien, 2 lectures about alignment methods for biological sequences, methods for phylogenetic analysis, including practical exercises in unix

Other Employments/Teaching Assistant

2006-2010 (SS2006-SS2010) EDV-course for molecular biologist, teaching assistant (5 days block course) University Vienna, (Dr. Martin Grabner)

2005 (SS05) Maintainance of a publication database, Heinrich-Heine-Universitaet Dusseldorf, Institut fuer Informatik (Prof. Dr. Michael Leuschel)

2004 (WS04/05) Created a custom cDNA clone annotation database, Heinrich-Heine-Universitaet Dusseldorf, BMFZ Biologisch-Medizinisches Forschungszentrum (Dr. Andreas Bayer, Dr. Karl Koehrer)

2004 (SS04) Informatik II, Tutor for seminal lecture exercises, Heinrich-Heine-Universitaet Dusseldorf, Institut fuer Informatik (Prof. Dr. Martin Mauve)

2004 (SS04) Bioinformatik II, Tutor for a practical perl programming course, Detecting CpG Islands using HMM (Hidden Markov Models) Heinrich-Heine-Universitaet Dusseldorf, Institut fuer Bioinformatik (Prof. Dr. Arndt von Haeseler, Dr. Ingo Ebersberger)

2004 (SS04) Bioinformatik I, Teaching assistant for practical exercises, Introductory bioinformatics course about alignment and phylogenetic analysis, Heinrich-Heine-Universitaet Dusseldorf, Institut fuer Botanik III (Prof. Dr. William Martin)
Publications


