universität wien

DISSERTATION

Titel der Dissertation

Addressing Biological Questions with Massive Sequence Data

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr. rer.nat.)

Verfasser:	Sascha Strauß
Matrikel-Nummer:	0749068
Dissertationsgebiet (lt. Studienblatt):	Molekulare Biologie
Betreuer:	Univ.-Prof. Dr. Arndt von Haeseler

Wien, im Dezemer 2009

Contents

List of Figures

List of Tables

1 Motivation

The instructions that are needed to build a living organism with all its structures and biochemical functions is stored in its DNA, a polymer of chemical compounds, called nucleotides. The specific order of the four different nucleotides, distinguishable by their individual bases (adenine, cytosine, guanine and thymine) encodes the information. Hence, the knowledge of the exact sequence of the nucleotides in DNA molecules is the most fundamental kind of data in biological science. Accordingly, a large interest in the generation of sequence data exists.
Two important landmarks in the history of sequence data generation are worth mentioning. The *Sanger sequencing method* enabled researchers for the first time to sequence DNA molecules efficiently (Sanger and Coulson (1975)). The invention of the *polymerase chain reaction* (PCR, Mullis *et al.* (1986)) provided a method to amplify trace amounts of DNA molecules to obtain sufficient quantities for sequencing. The combination of both methods formed the foundation for large-scale sequence projects.
Since then, permanent improvement of sequencing methods led to an ever-increasing amount of sequence data, yielding a true data flood. This development can be representatively demonstrated by the growth of GenBank, one of the major public sequence database hosted at the National Center for Biotechnology Information[1].

After starting with 606 sequences in 1982, GenBank grew exponentially and is currently hosting more than 110 million sequences (Figure 1.1). On average, the public available sequence data doubled every 30 month (Benson *et al.* (2009)). Similarly, the number of completely sequenced genomes grew exponentially as well (Fig. 1.2). This data now allows to re-address open standing questions in many fields of biological science. For example, the study of evolution particularly benefits from the massive amounts of sequence data. To get a robust resolution of splitting events between species lineages that took place hundreds of millions of years ago, lots of data from multiple taxa has to be incorporated (e.g. Bapteste *et al.* (2002), Rokas *et al.* (2005)). On the other hand, to investigate the evolution of biological systems comprised of various components, information gained from completely sequenced genomes can be utilized (e.g. Francke *et al.* (2005)).

However, the sheer amounts of data also require the development of new methods to

[1] `http://www.ncbi.nlm.nih.gov`
[2] Data source: `http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html`
[3] Data source: `http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj`

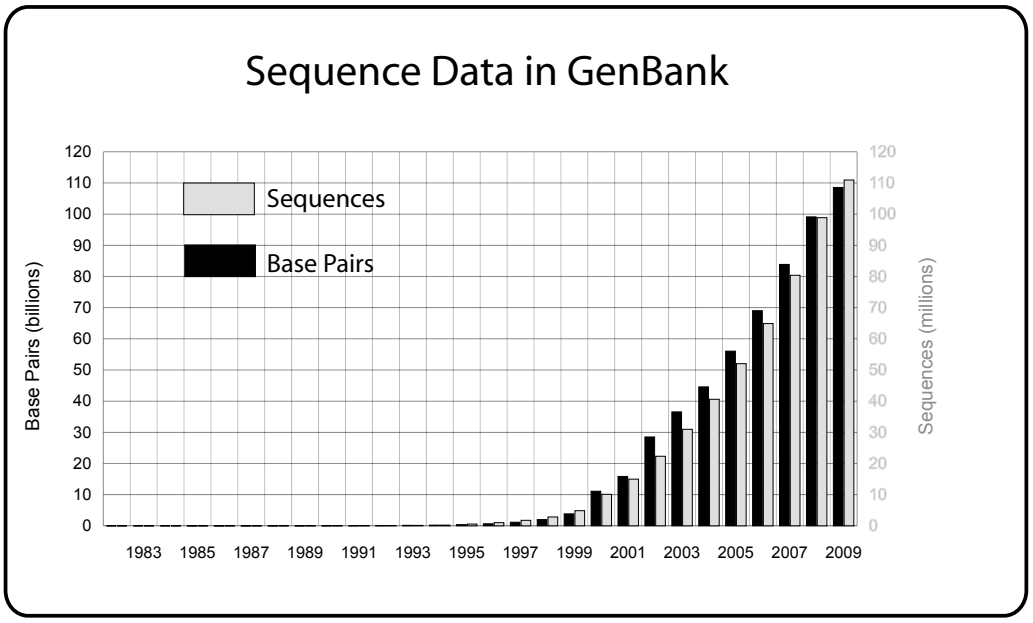


Figure 1.1: GenBank Growth The grey bars show the number of sequences (in millions) available from GenBank in each year. The black bars give the total length of the sequences in billions of base pairs[2]

handle and analyze them and to visualize the results.

In the first part of this thesis, we present our approach to incorporate massive sequence data to reconstruct the evolutionary relationships between species on a large-scale.

In the second part, we describe a systematic investigation of a unique transport system found in Gram-negative bacteria.

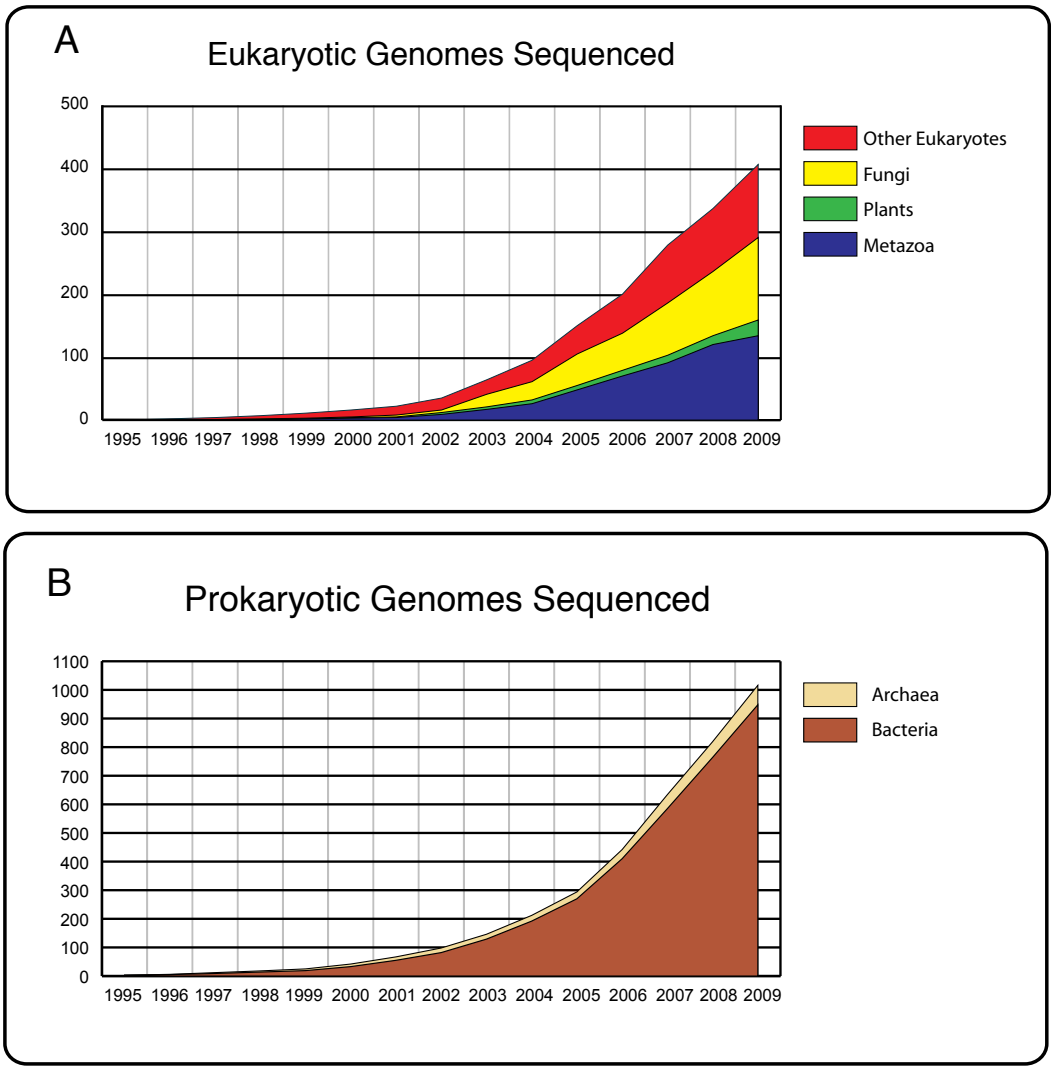


Figure 1.2: Number of Completely Sequenced Genomes These plots illustrate the increase of completely sequenced eukaryotic ((**A**)) and prokaryotic **(B)** genomes over the last fifteen years.[3]

Part I

EST-based Phylogeny Reconstruction

The time will come I believe, though I shall not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of nature.

Charles R. Darwin

2 Introduction

Between 5 and 50 million species are estimated to live on earth (May (1988)). For several hundred years, scientist are trying to grasp this tremendous biodiversity by categorizing the species. Initial efforts to establish an all-embracing systematic of all species were solely based on observable features, such as morphological characters or embryonic development. In 1735, Carl Linnaeus suggested a hierarchical nomenclature that can be used to describe groups of organisms based on shared characteristics (Linnaeus (1735)). This nomenclature, modified by Georges Cuvier and Ernst Haeckel in the 19th century, is still in use (Valentine, 2004, pp. 7-8). In his famous book "On the origin of species", Charles Darwin postulated that species are related to each other by genealogy; that they can be even traced back to a universal common ancestor (Darwin (1859)). This hypothesis is commonly accepted nowadays. The observed morphological similarities among species are therefore not random, but reflect the degree of relatedness, i.e., the time which has passed, since the lineages that lead to the extant species split up. It was not until 1950, however, that this theory was incorporated into the systematic of species. In that year, Willi Hennig proposed to use the evolutionary relationships of taxa, their *phylogeny*, as foundation for their classification (Hennig (1950)), which yielded some conflicts with the traditional Linnean taxonomy (Valentine, 2004, p. 10).
The development of efficient protein and DNA sequencing techniques during the 1970s (Niall *et al.* (1973) and Sanger and Coulson (1975)) granted access to large-scale sequence data and therewith to alternative characters, whose states can be compared in individuals to draw conclusions on their evolutionary relationships.
In the context of phylogeny reconstruction, sequence data provide some advantages over morphological data (Hillis (1987); Graur and Li (2000)). First, DNA or protein sequence data contain more independent characters that can be compared. On the genome level, 5×10^3 to 4×10^{11} nucleotides are available. Since the body plan of an organism is encoded in its DNA, the morphological data is always a smaller subset of the molecular information. Second, DNA sequences are heritable entities, which are passed to the next generation. On the contrary, character states of morphological features are often influenced by the environmental conditions the specimen is exposed to, which are not transmitted to its descendants. Third, sequence data is defined by a universal alphabet. DNA or protein sequences can be compared over large evolutionary distances, while shared morphological characters between, for example, bacteria and mammals are hard

to find. Finally, with sequence data, the extent of differences between two individuals can be quantified by counting differing positions in a sequence alignment. This allows to use sophisticated statistical methods, whereas morphological data often can only be evaluated qualitatively. These advantages were quickly realized by several authors, who developed methods to infer phylogenetic trees based on sequence data (e.g. Fitch and Margoliash (1967), Felsenstein (1981)).
Soon, these methods were employed to reconstruct phylogenies of very distantly related taxa which lineages separated millions of years ago. For that, sequence data of universally present genes, such as the 18S rRNA gene (Aguinaldo *et al.* (1997)) or the gene encoding the myosin heavy chain II (Ruiz-Trillo *et al.* (2002)) were employed. These studies provided some evidence that formerly widely accepted groupings of taxa, established by morphological data, might be wrong (Aguinaldo *et al.* (1997)). But the results were not robust enough to end all discussions. In order to reliably resolve such deep splits, sequence data from multiple genes are necessary (Rokas *et al.* (2003b), Rokas *et al.* (2003a)).
More recently, Expressed sequence tags (ESTs) were introduced into the field of phylogenetics (e.g. Bapteste *et al.* (2002)). This type of sequence data is highly abundant and accounts for the majority of the sequences stored in GenBank (Benson *et al.* (2009)). Consequently, the application of ESTs led to a further enlargement of data sets, which nowadays typically comprise more than hundred genes from an equally large number of taxa (Philippe *et al.* (2004), Dunn *et al.* (2008)).
However, despite enormous efforts, the phylogenetic relationships between the metazoan (animal) phyla are still under heavy debate and the common view is constantly modified (Halanych (2004), Philippe *et al.* (2005a), Irimia *et al.* (2007), Dunn *et al.* (2008)). This is mainly owed to the several hundreds of million years that have passed since those splits occurred. Such a huge time frame makes the reconstruction of evolutionary relationships challenging, but can hopefully be achieved by the incorporation of even more data and the development of better models of sequence evolution (Baurain *et al.* (2007)). Unfortunately, the data at hand for some of the phyla is very sparse, which might be also an explanation for the lack of resolution within some of them (Giribet (2008)).

In 2006, the Deutsche Forschungsgemeinschaft (German Research Foundation) therefore initiated a program called 'Deep Metazoan Phylogeny' (DMP)[1] in which the expertise of multiple research groups is combined. The ultimate goal of this project is to assemble a universal tree of Metazoa, based on both molecular and morphological data. In order to close the gaps in the sequence data, Expressed Sequence Tags have been coordinately generated by the members of the DMP project. In combination with the large amounts of already publicly available ESTs (Benson *et al.* (2009)), new data sets can be compiled that hopefully shed more light on those parts of the metazoan species tree that are still disputed.

[1] `http://www.deep-phylogeny.org/`

However, the manner of preparation makes a processing of ESTs necessary, before they can be used in phylogenetic analyses (Nagaraj *et al.* (2007)). Within the DMP project, we therefore developed a framework to compile EST-based sequence data sets suited to address the evolutionary relationships of animals. The framework includes the infrastructure to process and organize ESTs from hundreds of species and methods to filter these massive amounts of sequence data to gain informative subsets. Although the framework has been primarily designed to aid the DMP project, its application is not restricted to metazoan related problems and we consequently extended its use to address the phylogenies of plants and fungi as well.
In this part of the thesis, we first provide a detailed descriptions of the generation of ESTs. We then present our framework to incorporate ESTs into phylogenetic studies, followed by a demonstration of its application. Finally, we present new aspects of EST-based phylogeny reconstructions .

3 Expressed Sequence Tags

3.1 Background

In 1983, Scott Putney and his colleagues developed a rapid and relatively inexpensive method for identifying clones of particular genes in a library of clones. The library was generated by extracting the mRNA of a tissue, reverse-transcribing it into complementary DNA (cDNA) and cloning these in bacteria (Putney *et al.* (1983)). The genes the authors were interested in are characterized as being highly expressed. This means that they contribute a substantial amount to the total mRNA mass of a cell and therefore should also be represented by multiple clones in the library. Hence, they simply picked ∼180 clones randomly, sequenced parts of the contained cDNA inserts as single-pass reads and compared the translated sequences with the known protein sequences of the genes of interest. Although the authors did not know which genes are represented by the clones they chose, they still could unequivocally identify clones for the majority of genes they were looking for. Furthermore, they discovered new sequence variants of some of these proteins.
At the beginning of the 1990s, Mark D. Adams and colleagues took up this method to aid the discovery of new human genes to complement the, at that time, ongoing human genome sequencing project (Adams *et al.* (1991)). They sequenced random cDNA clones to efficiently determine transcribed regions of the genome. By mapping the cDNA sequences on the genome sequence, the locations of hitherto unknown genes were discovered. However, the information was often incomplete, because in many cases only a part of the mRNA is covered by the sequenced cDNA and lowly expressed genes might not be detected at all. Adams *et al.* (1991) did not only demonstrate the usefulness of this strategy for gene discovery on a high-throughput scale, they further coined the today commonly used term "Expressed Sequence Tags" (ESTs). ESTs are characterized as short parts of gene transcripts, usually 200-800 nucleotides in length (Nagaraj *et al.* (2007)), that have been single-pass sequenced, resulting in a relatively high error rate of about 3% (Hillier *et al.* (1996)).

Besides gene discovery, ESTs were quickly discerned as valuable tools in a broad range of applications such as gene identification (e.g. Nakamura *et al.* (1997)), SNP detection (Picoult-Newberg *et al.* (1999)) and genome annotation (e.g. McCombie *et al.* (1992)).

Their usage also led to a better understanding of gene expression mechanisms in general (e.g. Okubo *et al.* (1992)) and in cancer studies in particular (Krizman *et al.* (1999)). Consequently, many high-throughput projects were initiated, followed by a dramatic increase of the number of ESTs in the public domain. As a consequence, in 1992 the National Center of Biotechnology Information (NCBI) dedicated ESTs their own database, called *dbEST* (Boguski *et al.* (1993)). Today, dbEST is the biggest division of the NCBI sequence database, with over 63 million entries from more than 1,800 different species as of October 2009[1].

3.2 EST Generation - Overview

Several protocols for the generation of ESTs have been developed during the past few decades. In the following section we will outline the basic principles that are common to most approaches, followed by a detailed description of each step.
First, mRNA molecules are extracted from the cells and reverse transcribed into the cDNA. The mRNA template is removed, and a second strand of cDNA is synthesized. The double-stranded cDNA molecules are inserted into a vector and the vector-insert construct is transferred into bacteria for cloning. Several bacteria clones are picked randomly, the vector-insert construct is extracted and the insert is sequenced.

3.3 EST Generation - Detailed Description

3.3.1 cDNA Synthesis

Single stranded mRNA molecules are isolated from a tissue or a whole organism. Next, they are reverse transcribed into DNA, because DNA is much better suited for the cloning process than RNA. For the reverse transcription, an enzyme called *reverse transcriptase* is used which was originally found in retroviruses and which has the ability to synthesize DNA from RNA templates (Temin and Mizutani (1970)). In comparison to genomic DNA, this type of DNA is lacking the introns that have been spliced out during the maturing of the mRNA. In order to differentiate between genomic DNA and DNA based on mRNA templates, the latter is called complementary DNA (cDNA).
The reverse transcription needs to be primed, that is, the reverse transcriptase will attach itself to the mRNA molecule only in the presence of a double-stranded section of mRNA. This can be achieved by adding a short stretch of single-stranded DNA that is

[1] http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

complementary to the mRNA. Designing such a primer would in principle require that at least parts of the mRNA sequence are known, which is mostly not the case. However, this can be overcome since the majority of mRNA molecules carries a stretch of exclusively adenine nucleotides at their 3′ end, the poly-A tail. A primer that consists of thymine bases only, called oligo(dT), binds to the poly-A tail via hydrogen bonds (Spiegelman *et al.* (1971)), see Figure 3.1A. Once the primer is attached to the mRNA, the reverse transcriptase will start to tie deoxyribonucleotide triphosphates (dNTPs), the building blocks of the DNA, to the mRNA, forming an RNA/DNA hybrid, see Figure 3.1B.
Owing to the fact that the viral reverse transcriptase is missing a proofreading mechanism, its error rate is relatively high, compared to other DNA synthesizing enzymes. Depending on the chosen enzyme, between one in 1,700 and one in 30,000 bases do not correspond to the mRNA template (Roberts *et al.* (1988)).
Furthermore, the reverse transcriptase does not necessarily reach the end of the mRNA, because single-stranded RNA molecules tend to form secondary structures due to intramolecular interactions of the bases. These structural elements can prevent a passing of the enzyme. Additionally, contaminations with RNA-dismantling enzymes (RNases) can lead to truncated transcripts (Greene and Rao (1998)). In such cases, the resulting ESTs will cover only the 3′ end of the mRNA, because due to the use of the oligo(dT) primer, the reverse transcription is started from this end. To reduce this bias, a random primer can be used (Feinberg and Vogelstein (1983)). This refers to a mixture of single-stranded DNA octamers or hexamers containing all possible combinations of the four bases adenine, cytosine, guanine and thymine. They will bind to the mRNA wherever the sequence is complementary to that of the primer. Random primers have the advantage of not being restricted to the 3′ end of the mRNA, but they usually yield shorter products.
In order to clone the cDNA, a double-stranded version is needed. Therefore, an enzyme called *RNase H* is used to create nicks in the mRNA sugar-phosphate backbone in which a *DNA polymerase* will start to synthesize the second strand of the cDNA, successively replacing the mRNA completely (Alberts *et al.* (2007)), see Figure 3.2A and B.

3.3.2 Cloning of cDNA

The double-stranded cDNA can be incorporated into a cloning vector, a molecular device suitable for transporting the cDNA into a living host cell, the bacterium *Escherichia coli* for example. In the early days of cDNA cloning, vectors based on plasmids were mainly used. Plasmids are circular, non-chromosomal DNA molecules naturally occurring in bacterial cells. Since the size of the plasmids is limited, elements not essential for the cloning process have been removed from the vector, so that larger cDNA molecules can be loaded. To insert the cDNA into a vector, first a short piece of double-stranded DNA, called adapter or linker, is ligated to both blunt ends of the cDNA (see Figure 3.2(C)).

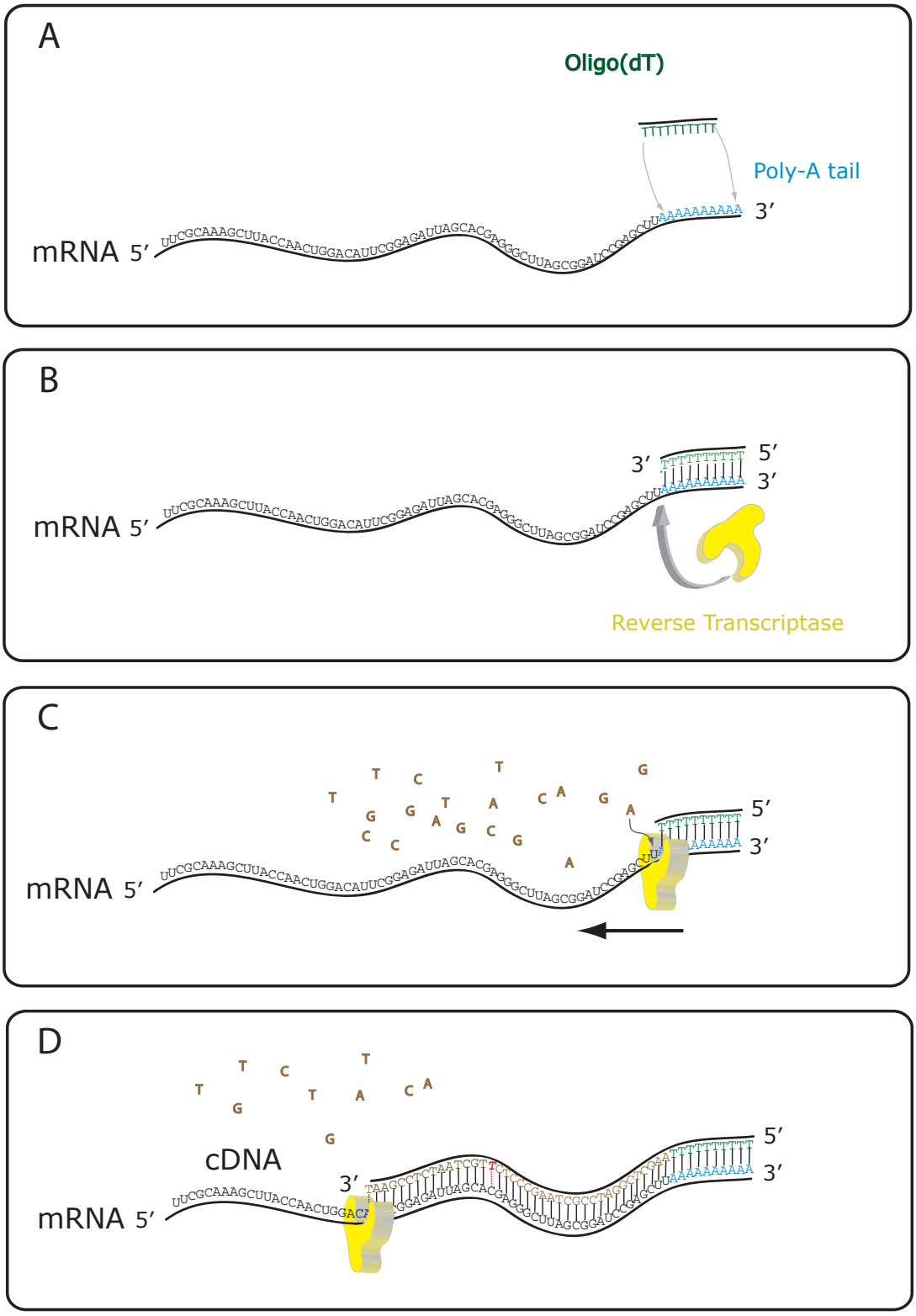


Figure 3.1: cDNA First Strand Synthesis (A) A cartooned mRNA is shown. The letters represent the nucleotides (A: adenine, C: cytosine, G: guanine, U: uracil). An oligo(dT) primer is annealed to the poly-A tail at the 3′ end of the mRNA. Once the primer is bound to the mRNA, the reverse transcriptase will attach itself to the mRNA **(B)** and start to synthesize the first strand of the cDNA using the mRNA as a template by binding the provided complementary dNTPs to it **(C)**. As the reverse transcriptase is lacking a proofreading mechanism, errors can occur during this process, highlighted in red **(D)**

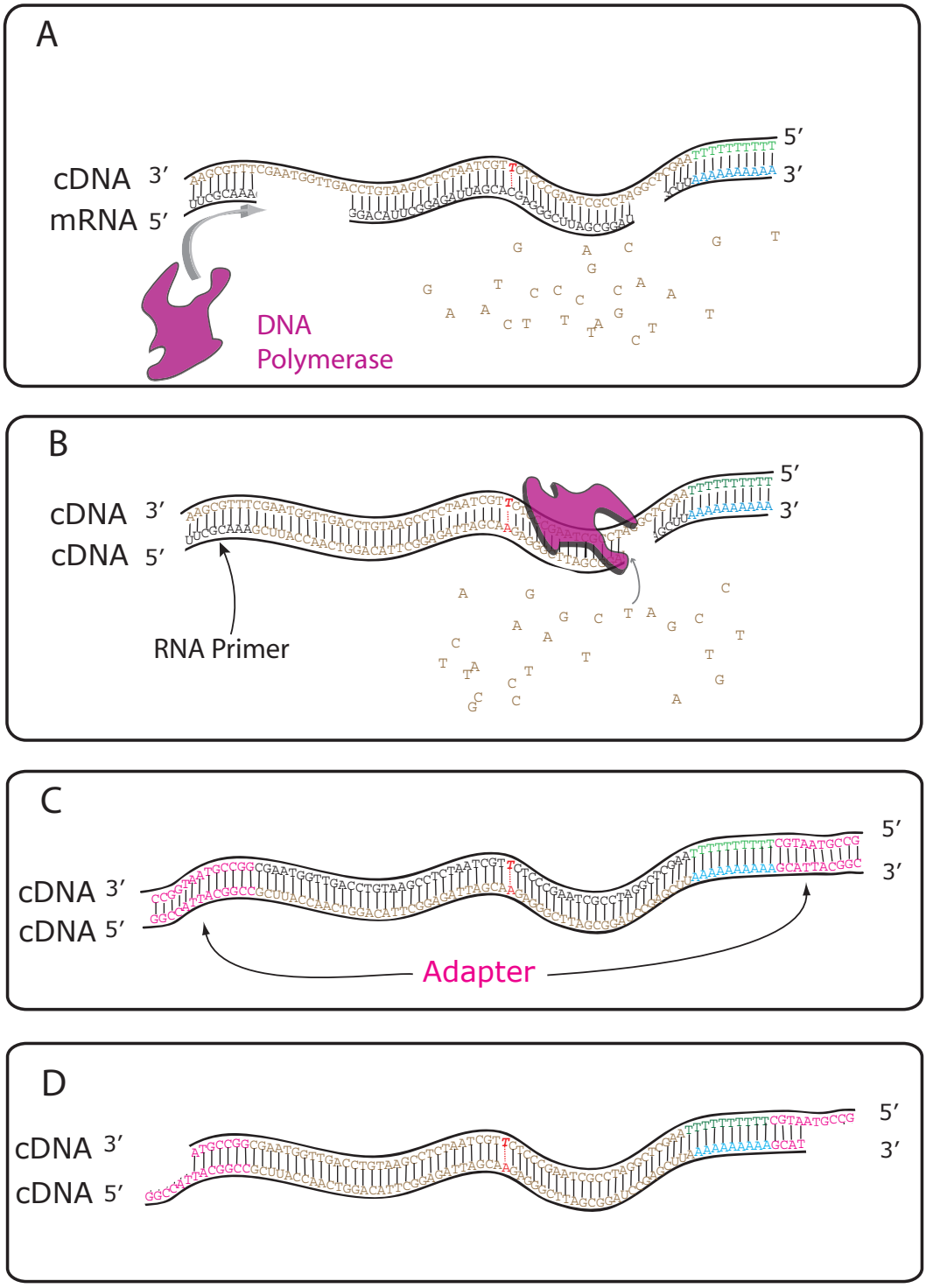


Figure 3.2: cDNA Second Strand Synthesis RNase H has created nicks within the mRNA **(A)**. The DNA polymerase can start to synthesize a second cDNA strand, filling the gaps and replacing the remaining mRNA stretches **(B)**. Errors introduced by the reverse transcriptase are adopted by the DNA polymerase and manifested in the second strand, highlighted in red. A short stretch of RNA remains on the 5′ end of the newly synthesized strand as the DNA polymerase only works in 5′-3′ direction and hence is not able to process the part on the 5′ end. It will be removed and an adapter is ligated on both ends **(C)**. By treating the molecule with a restriction enzyme, overhanging ends are formed **(D)**.

The sequence of the ligated piece is recognized by a certain restriction enzyme which cuts the adapter in such a way that at the termini, one of the two strands is longer than the other. Similarly, the vector has been cut with a different restriction enzyme at a position called multiple cloning site, generating overhanging ends that are complementary to the cut adapter (Fig. 3.3(A) and Fig. 3.3(B)). The bases of the overhanging ends form hydrogen bonds, and the cDNA molecule integrates seamlessly into the vector molecule. The vector then needs to enter a bacterial cell, but the plasmids cannot pass the cell wall. The bacterial cells are therefore treated with an electric field or calcium chloride, which will cause holes in its cell wall. The vector then passively moves into the cell, which will continue with their normal life cycle and undergo cell devision. During this process, the plasmid together with the inserted cDNA is replicated as well, so that both daughter cells will have a copy of it. By consecutive reproduction, multiple identical copies (clones) of the cDNA are generated. A collection of cloned cDNA molecules is called a *cDNA library*.

Today, plasmid-based systems have been largely superseded by a newer vector system. It is based on bacteriophages (short: phages), a group of viruses that infect bacteria (Walker and Rapley (2000)). This system can handle larger cDNA molecules and a perforation of the bacterial cell wall is not necessary because of the phages natural ability to inject genetic material into cells. A cDNA molecule is inserted into a vector molecule, similar to the plasmid based system. Afterwards, the proteins that form the coat of the phage are added. These have the ability to assemble themselves into functional units (Figure 3.3(C)), which can then inject the vector into bacterial cells.

Naturally, bacteriophages have two different life cycles. In the *lysogenic* cycle, the DNA that was injected into a bacterial cell will be integrated into its chromosome and passively copied during the bacterial cell division. In contrast, in the *lytic* cycle, the injected DNA forms a circle and is then transcribed and translated by the bacterial gene expression machinery, continuously producing new bacteriophages, including a copy of the vector sequence with a cDNA insert. Eventually, the cell literally bursts and the released bacteriophages can infect other cells.

Both cycles have been adopted for the cloning process. The lytic cycle has the advantage that the numbers of copies rapidly increases, because each infected cell produces multiple clones of the injected cDNA insert. But since the phages continuously destroy the bacterial cells, the clones cannot be maintained in the bacterial colony over multiple generations, but have to be harvested and processed immediately.

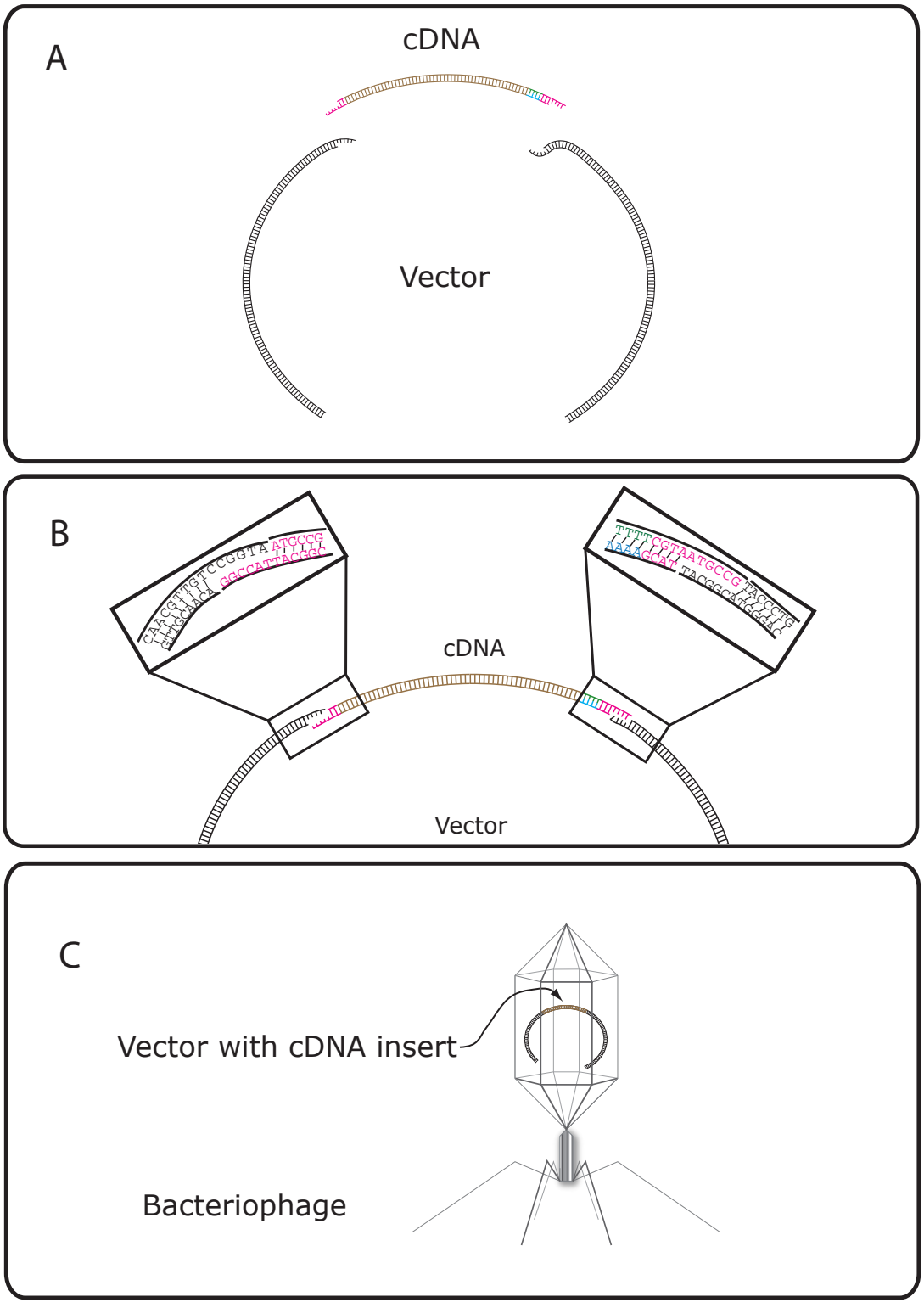


Figure 3.3: cDNA Cloning The double-stranded cDNA with the attached adapter is mixed with the vector DNA molecule that has been digested with a certain restriction enzyme **(A)**. The exposed DNA sequence of the vector perfectly matches the attached adapter of the cDNA, allowing a fusion of both molecules **(B)**. If a bacteriophage-based vector system is used, proteins forming the envelope of the phage are added, and a functional unit containing the cDNA will be assembled **(C)**.

3.3.3 EST Sequencing

A broad variety of different sequencing methods have been developed. Here we describe in detail the chain-termination method, developed by Frederick Sanger (Sanger and Coulson (1975)) also called the Sanger sequencing method. It has been widely used for the generation of ESTs.
In order to determine the nucleotide sequence of a cDNA clone and by that obtain an EST, first, both strands of the double-stranded vector molecule have to be separated so that a primer can be annealed. The vector includes a group of known nucleotides located either upstream or downstream of the multiple cloning site, called standard priming site, to which the primer will complementarily bind (Fig. 3.4A and B). The attached primer will initiate a second strand synthesis by the *DNA polymerase*. As the DNA polymerase only operates in 5′ - 3′ direction, the added primer determines whether the cDNA insert is sequenced from the 5′ or 3′ end of the original mRNA (Fig. 3.4C).
For the synthesis reaction, dNTPs are provided. Additionally, a modified version of the four dNTPs is added as well, in which the hydroxyl group at the 3′-C atom is replaced by a hydrogen atom. Therefore it is called dideoxynucleotide-tri-phosphate (ddNTP). ddNTPs are competing with the dNTPs for being incorporated into the growing DNA chain. An incorporation of the first will terminate the synthesis and the chain is not elongated anymore, forming a fragment of the original sequence (Fig. 3.4D). Such a chain synthesis is repeated multiple times for each clone, so that the growing strand is terminated at every single position of the original nucleotide sequence (Fig. 3.5). However, in case of long cDNA molecules the DNA polymerase rarely reaches the end, because at every position the chain is terminated with a certain probability. The total probability of the DNA polymerase reaching a specific position therefore decreases with increasing distance from the starting point of the synthesis.
After a fixed number of cycles, the generated fragments are separated by their size, either by a sequencing gel or by capillary electrophoresis. The resolution of these procedures is high enough to separate fragments whose lengths differ in only a single nucleotide. Since the synthesis always starts at the same point (defined by the primer), all fragments of one fraction are identical, including the terminating ddNTP. To determine the specific terminal ddNTP of a fraction, each ddNTP is labeled with a different colored dye. A laser in the sequencing machine stimulates the dye and the color is detected. The sequence of detected colors is translated into a chromatogram, in which the colored peaks correspond to the four bases (Fig. 3.6). Computer programs such as Phred (Ewing *et al.* (1998)) can analyze the order of the peaks and subsequently reconstruct the DNA sequence automatically. This process, however, is error-prone. Especially at the termini of the sequence, sequencing errors frequently occur, because here, the peaks in the chromatogram are often not clearly separated but merge into each other. This introduces a lot of noise in the chromatogram which is difficult to process. To have an easy to interpret measure of the

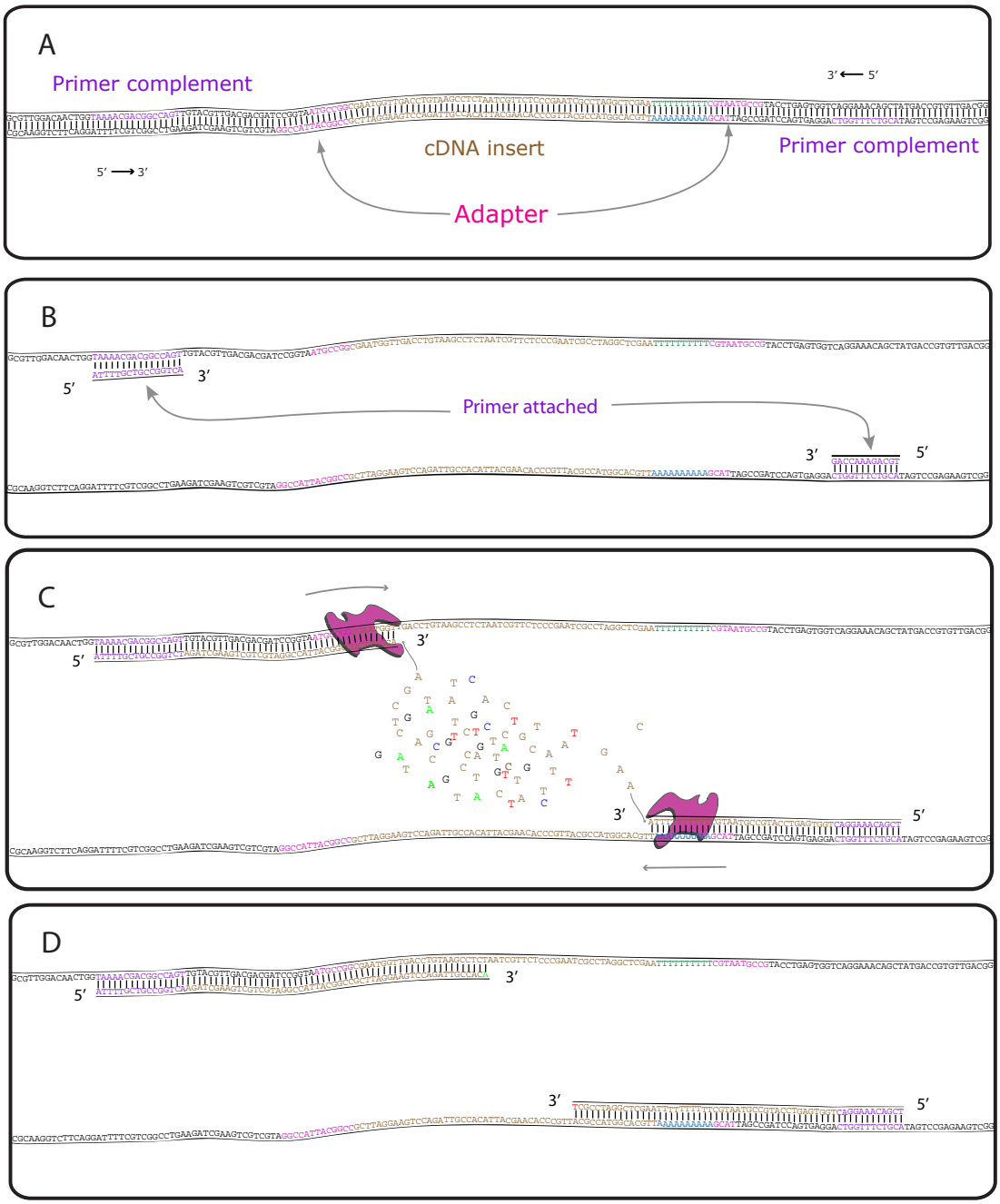


Figure 3.4: Chain-termination Reaction An excerpt of a vector molecule with integrated cDNA insert (in brown color) is shown in **(A)**. Up- and downstream of the insert, standard priming sites of the vector sequence are highlighted in purple. The remaining vector sequence is written in black. The two adapter sequences that have been previously ligated to the insert for cloning are colored in pink. The two strands are separated and one of the two primers is attached **(B)**. The DNA polymerase will synthesize a second strand in 5′-3′ direction, only **(C)**. The choice of the primer determines which strand is synthesized. The sequence (EST) then covers either 5′ end of the original mRNA or the 3′ end. For the synthesis reaction, dNTPs (in brown) and ddNTPs (ddATP in green, ddCTP in blue, ddGTP in black and ddTTP in red) are added. If one of the ddNTPs is attached to the end of the chain instead of its corresponding dNTP, the reaction is terminated **(D)**.

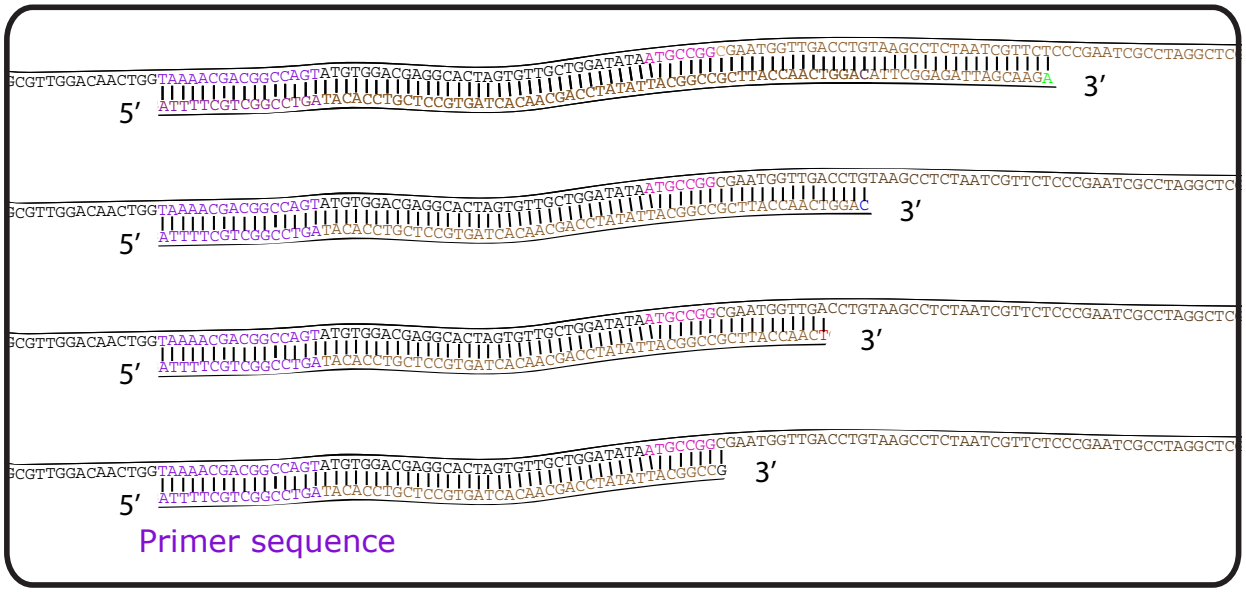


Figure 3.5: EST Sequencing Preparation The second stand synthesis reaction is executed multiple times for each clone, so that a ddNTP will be added at different positions, forming differently sized fragments of the original cDNA molecule.

reliability of each base called from the chromatogram, Ewing and Green (1998) developed a scoring scheme based on empirically determined error rates. The score is calculated by analyzing different attributes of the chromatogram, such as the homogeneity of peak spacing and signal to noise ratio. The resulting quality value q then equals the error probability for each sequence position in logarithmic scale, expressed by the formula: $q = -10 \times log_{10}(p)$, where p is the estimated error probability for that base-call given the image attributes. For example, a base quality value of 30 means that one in 1000 nucleotides will be false.

The last years witnessed a shift from the described traditional sequencing methods to the next-generation sequencing techniques such as Roche's 454[2] or Illumina's Solexa sequencing[3]. Since those techniques deliver shorter reads, they have not been yet extensively used for EST generation, although some groups did some initial experiments very recently (Roeding *et al.* (2009); Gibbons *et al.* (2009)).

[2] `http:///www.454.com`
[3] `http://www.illumina.com/technology/sequencing_technology.ilmn`

3.4 ESTs in a Phylogenetic Context

During the past years, ESTs became popular for phylogenetic studies (e.g. Dunn *et al.* (2008), Roeding *et al.* (2007)), because on one hand, plenty of data is already publicly available and on the other hand, ESTs for taxa not yet present in the sequence databases can be generated at reasonably low costs. Furthermore, since ESTs are based on mRNAs, they mainly represent protein coding regions of a genome. This is particularly useful for studies addressing evolutionary events that took place millions of years ago, because the phylogenetic signal fades slower over time on the protein level than on the DNA level (Opperdoes (2003)).

However, the advantages of using ESTs in phylogenetic analyses comes at the cost of several disadvantages. As pointed out earlier (Section 3.3), the generation of ESTs involves several stages in which the cDNA is altered from its mRNA template. Most prominently, ESTs usually do not cover the complete mRNA due to the inefficiencies of the reverse transcriptase and the sequencing process. This leads to a reduction of the phylogenetic signal (when compared to the full length mRNA sequence), because it is assumed that longer sequences contain more phylogenetic information (Philippe *et al.* (2004)). Also, EST sequences as obtained from the sequencing machine usually contain contaminations, such as parts of the vector, the adapter sequence and genetic material from the bacterial host cell that was integrated via transposable elements. When not taken care of, these contaminations can cause severe problems during phylogenetic tree reconstruction, because these sequence parts do not share an evolutionary history with the cDNA they are attached to. Finally, nucleotides of the cDNA do not necessarily correspond to their mRNA counterpart as the reverse transcriptase is not operating faultlessly. Consequently, sequences that are compared will show more differences on the nucleotide level, which makes them appear more distantly related than they really are. Finally, the quality of EST sequences is usually poor at the ends, caused by the sequencing process. Discounting this fact and not removing faulty nucleotides will lead to an overestimate of the number of substitutions that have been introduced by mutations over time.

Fortunately, methods exist to deal with such sources of error. Vector contaminations can be identified by comparing the EST against the known vector sequence and subsequently remove them. As explained, modern base-calling software do not only deliver the sequences themselves, but additionally an estimate of the correctness of each single sequence position; the base quality values. Consequently, a researcher working with the sequences has knowledge about which nucleotides can be trusted and which should be regarded with suspicion. The latter category can be simply removed or masked before performing analyses.

To counteract the other error sources, one can take advantage of the redundancy of ESTs: A gene can be transcribed in parallel, which results in the presence of multiple

mRNA molecules derived from the same gene. The pace in which a gene is processed to the final protein, the gene expression level, reflects the need of the cell for this gene product. Expression levels differ therefore not only between genes but also for one gene between different tissues, developmental stages or environmental conditions (e.g. Su *et al.* (2004)). Usually, the individual expression levels of genes are unknown when the mRNA is extracted to construct a cDNA library, but typically there are some genes (10 to 15) that account for up to 20% of the total mRNA mass of a cell. Approximately 1000-2000 genes are represented with intermediate levels of mRNA and the remaining genes are only found to be present with a few mRNA molecules or to be completely absent (Bonaldo *et al.* (1996)). It is very likely, that genes with a higher expression level are represented by several ESTs, because cDNA clones are usually randomly picked and sequenced.
A common strategy is to remove these redundancies after the sequencing, by grouping ESTs that stem from the same gene (clustering) and then assembling all overlapping ESTs in a cluster to form a longer continuous sequence, called contig. Although all ESTs in a cluster should represent the same gene, different mRNA molecules were used as templates. Hence, random errors introduced by the reverse transcriptase should not be present at the same position. Thus, errors will trigger conflicts during the consensus sequence determination of a contig and can be either corrected or at least marked as suspicious. As a further advantage of the clustering, different ESTs of the same gene can cover different parts of the mRNA. By clustering and assembling them, the final cDNA sequences can be extended, yielding a higher coverage of the gene and increase the phylogenetic signal compared to single ESTs.
A processing of ESTs is therefore straightforward and nowadays routinely done, independent of the application. Correspondingly, a broad range of tools for each step have been developed in the last 15 years (Nagaraj *et al.* (2007)). But with an increase in numbers and sizes of EST projects to be processed, there is a need for completely automated solutions. Consequently, to process the enormous amounts of EST data for the Deep Metazoan Phylogeny project, we developed a program pipeline that wraps up each individual processing step without user-interaction and that also takes care of the data management.

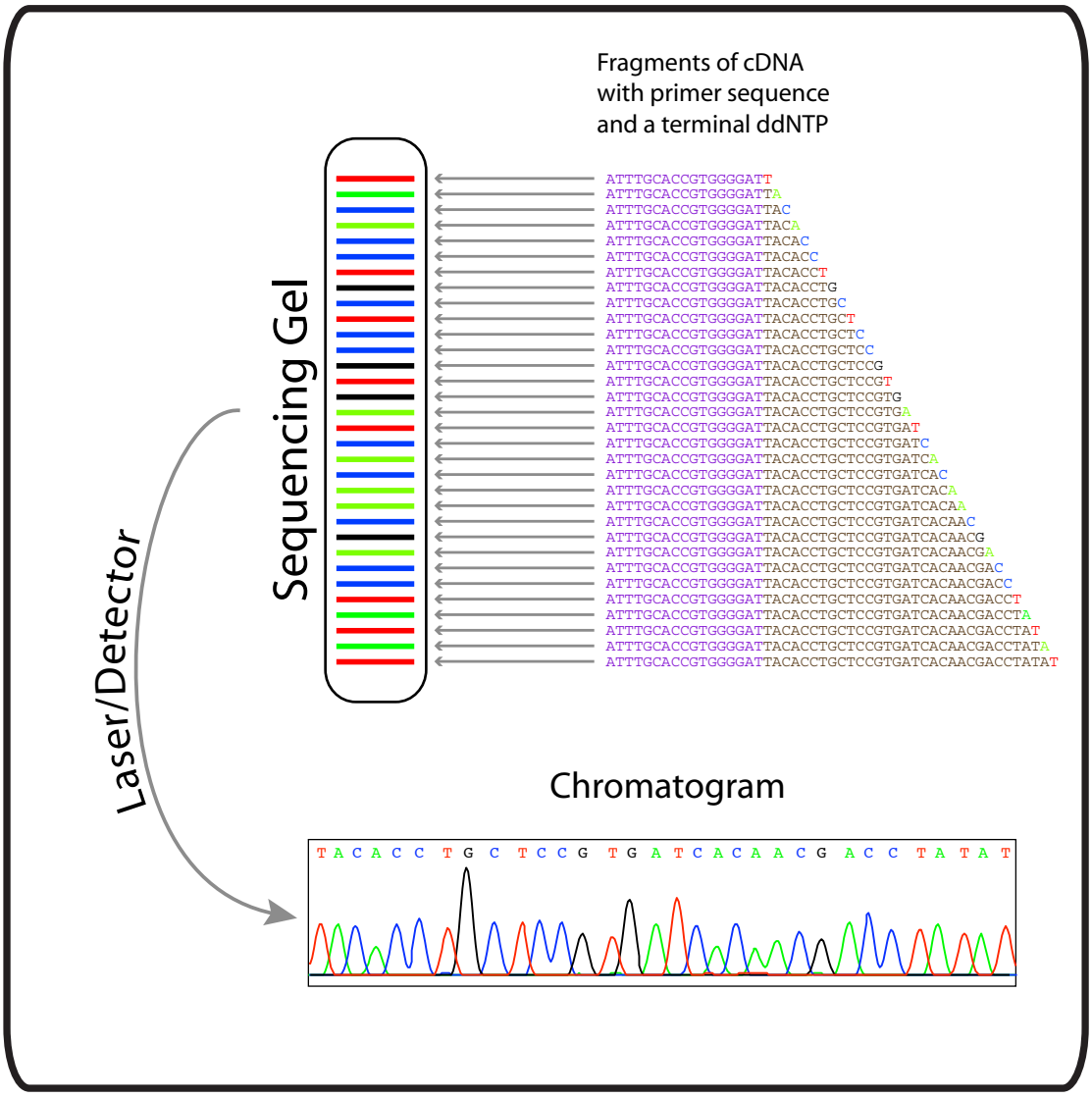


Figure 3.6: EST sequencing The synthesizing of fragments is continued for a fixed number of cycles, resulting in many fragments of different length. These fragments are then loaded on a sequencing gel for example. By that, the fragments will be separated according to their size, forming a band pattern. A laser aimed on the gel will stimulate the dye attached to the ddNTPs in bands passing the laser to emit light in its specific color. The time at which a band passes the laser corresponds to the length of the fragments in the band. The series of the four colors is detected and processed by the sequencing machine which yields a chromatogram. The chromatogram can then be translated into the actual DNA sequence.

4 EST Processing

4.1 Introduction

As already explained in the previous chapter, the state in which ESTs are obtained from the sequencer is not suitable for direct use in phylogenetic analyses. A wide choice of tools for every necessary step exists, to obtain high-quality data. Depending on the source, the available ESTs are provided on different levels of quality. Some sources only offer unprocessed ESTs as obtained directly from the sequencing machine, where they potentially contain vector contaminations and low quality regions. Other providers remove contaminations and low quality regions, but do not apply any clustering procedures. We call this state *preprocessed* hereafter. Finally, sequence data, based on cleaned and assembled ESTs, is available as well. Here, we describe a pipeline in which the complete processing of ESTs is performed automatically with a minimum of user interaction. We also take into account the different stages ESTs can be delivered in, to prevent unnecessary steps and save computational resources. An overview of the complete workflow is shown in Fig. 4.1. In the following we explain each processing step.

4.2 Detailed Description of each Processing Step

4.2.1 Base Calling

Sequence data that are directly received from the sequencer usually come as trace files. These contain the chromatograms. To infer the DNA sequence from the chromatograms, we use the program PHRED (Ewing *et al.* (1998)), which not only determines the nucleotide sequence but also provides base quality values for each nucleotide called.
Preprocessed ESTs are usually provided without base quality values, because low quality regions should have been already removed. But since some of the employed programs explicitly demand a quality value for each base, we generate artificial base quality sequences with a values of 20 for each nucleotide, which corresponds to an error rate of one false base in 100.

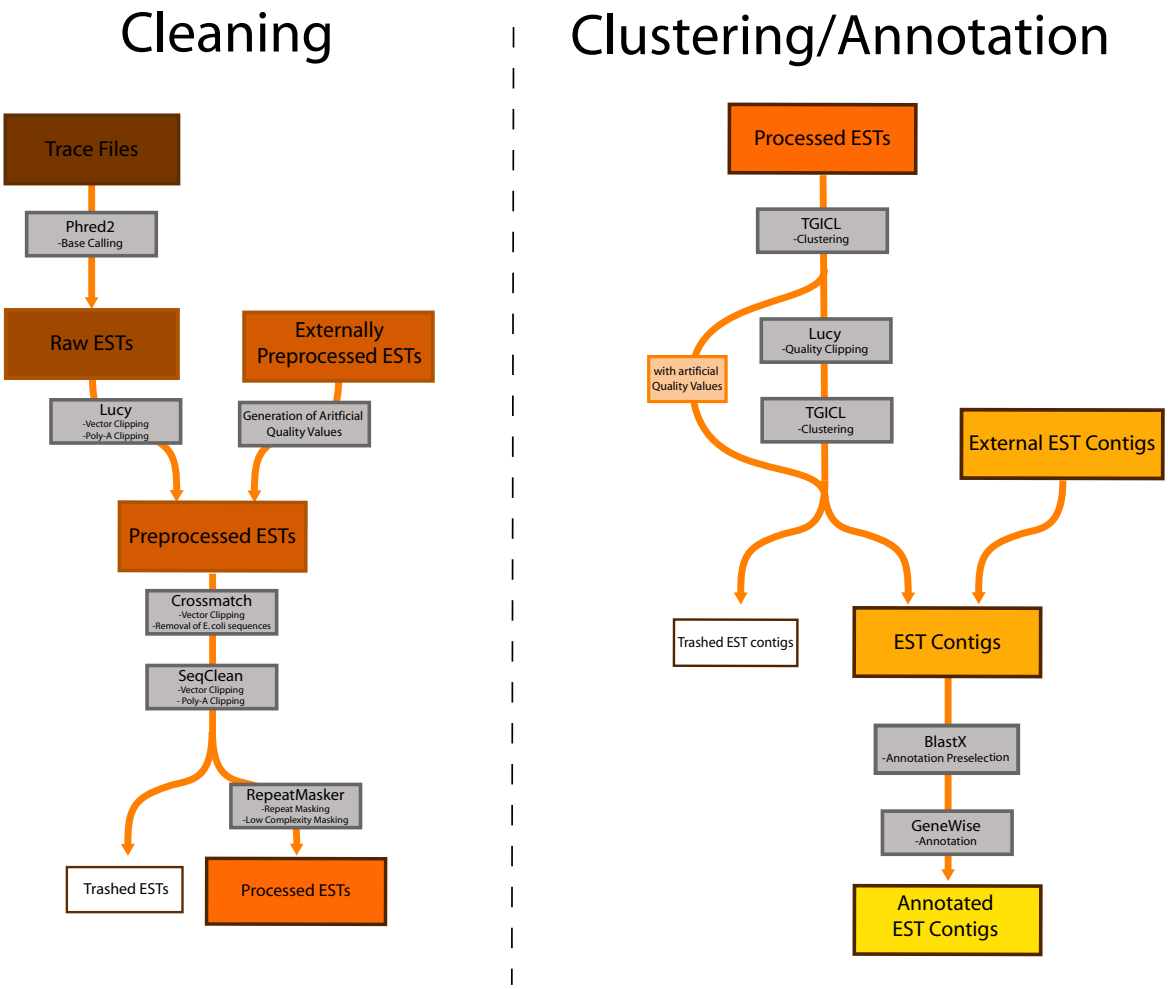


Figure 4.1: Pipeline Program Flow. Left: Overview on the cleaning steps of the pipeline First, if ESTs are delivered as trace files (chromatograms), the bases are called with the program PHRED2. Raw ESTs are then freed from vector contaminations and Poly-A tails with LUCY. For already preprocessed ESTs without base quality information, a quality value of 20 is assigned to each base. Preprocessed ESTs are scanned with CROSSMATCH for contaminations of vector sequences and bacterial genomic DNA. Subsequently, the program SEQCLEAN aims to detect remaining contaminations and Poly-A tails. ESTs shorter than 100 nucleotides are discarded. In the remaining sequences, repetitive elements and low complexity regions are masked with REPEATMASKER.

Right: Overview on the clustering and annotation steps of the pipeline Processed ESTs are clustered and assembled with TGICL. If base quality values are available, low-quality regions at the terminal parts of assembled sequences are removed and a second clustering step with TGICL is performed. These two steps are skipped if only artificial quality values are present. Subsequently, all sequences with a length of <100 bp are trashed. To achieve a tentative annotation, each EST contig is compared with the NCBI non-redundant protein database using BLASTX. The protein sequence of the best 25 BLASTX hits per contig are aligned to the contig separately using GENEWISE. The description of the protein sequence resulting in the highest GENEWISE alignment score is adopted as a tentative annotation.

The resulting EST reads together with their base quality values are then consigned to the cleaning procedure.

4.2.2 Cleaning

Unprocessed ESTs contain the remains of the vector, the adapter sequence, as well as poly-A tails. We identify and remove them with the program Lucy (Chou and Holmes (2001)). For the vector removal, a file has to be provided which contains the specific sequence of the vector 100 nucleotides up- and downstream of the multiple cloning site (see Section 3.3.2). Owing to this precise information, Lucy performs best in benchmarking studies (Chen *et al.* (2007)). Lucy is usually also used to remove low quality sequence parts. We disabled the low quality clipping here, but postponed it to after ESTs are clustered and assembled, see Section 4.2.3.
Preprocessed ESTs should be free of contaminations. While this is true for most of the ESTs, Chen *et al.* (2007) still found a vector contamination rate of 1.63% in dbEST, a public archive of preprocessed ESTs. At first glance, such a small percentage seems to be negligible. However, the authors found that the contaminated ESTs are not uniformly distributed across the projects. Most of the ESTs containing foreign DNA are concentrated in a relatively small number of projects. Consequently, if one of these projects is used as data source, the amount of contaminations can be substantial. We therefore scan preprocessed ESTs for vector contaminations as well.
However, for preprocessed ESTs vector information is often not available, which hampers the usage of Lucy. Thus, in addition to Lucy, we also use Crossmatch[1], which scans for vector and adapter contaminations in a more general way. It compares each EST against NCBI's vector sequence database, UniVec[2], which contains a sequence collection of commonly used vectors and adapters. Additionally, Crossmatch searches for similarities to the *Escherichia coli* genome, as parts of it can be integrated into the cDNA inserts via transposable elements during the cloning process. Hence, ESTs which we already processed with Lucy are passed through Crossmatch as well.
Finally, SeqClean[3] once again checks for the presence of left over contaminations with a comparison against UniVec. Furthermore, it looks for Poly-A tails and evaluates every sequence regarding its length, the number of undetermined bases and the proportion of low-complexity regions. Sequences with a length of less than 100 nucleotides or with too many uninformative bases are of poor quality. Thus they are discarded in this step.
The sequential application of different cleaning programs reliably removes any vector contamination from ESTs. However, all programs generate false positive vector predictions,

[1] http://www.phrap.org/phredphrapconsed.html
[2] http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
[3] `http://compbio.dfci.harvard.edu/tgi/software/`

i.e., genuine cDNA parts are erroneously identified as part of the vector and subsequently removed. On the one hand, the loss of useful data increases with the application of each program. On the other hand, since the purpose of our pipeline is to process EST data for phylogeny reconstructions, any unrecognized contamination can have a severe influence on the results. Therefore, we decided that to minimize the amount of contaminations, the loss of some useful data was acceptable.
For the last step of the cleaning section, we use RepeatMasker (Smit *et al.* (1996)) to mask repetitive elements and low complexity regions. Repetitive elements and low complexity regions are problematic for the clustering process, because they can cause sequence similarities between unrelated ESTs, which feigns a common origin. If not taken into account, the clustering could lead to artificial constructs of unrelated sequences, in other words, chimerics. While low complexity regions are easy to recognize, repetitive elements can consist of complex sequence patterns, which makes them harder to predict ab initio. Fortunately, they can be detected by comparing each EST to a database of known repetitive elements, called Repbase (Jurka *et al.* (2005)). RepeatMasker identifies low complexity regions and repetitive elements and masks them by writing the corresponding sequence stretches in lower case characters in the sequence files. This soft-masking is supported by many programs which ignore such regions during the clustering accordingly.

4.2.3 Clustering

Initial Clustering Step Clustering is performed with the TGICL package (Pertea *et al.* (2003)). Initially, cleaned EST sequences are grouped by searching for sequences that share identical or almost identical subsequences. This task is accomplished by the program MGBlast, included in the TGICL package. Its core is a modified version of the MegaBlast algorithm (Zhang *et al.* (2000)), a greedy, and thus very fast search strategy, that quickly finds highly similar sequence pairs. The modification of MGBlast compared to MegaBlast concerns the filtering of the output for minimum overlap length and sequence identity. In detail, sequences have to overlap for at least 40 base pairs to be considered as connected. Within the overlapping part, a minimal pairwise sequence identity of 95% is required. However, at the terminal parts, up to 30 mismatches are allowed to account for low sequence quality, which is typically found at the terminal parts of ESTs. MGBlast further recognizes lower case written nucleotides as masked. Masked parts will be ignored during the initial search for hits, but hits starting in an unmasked region next to a masked one can be extended into the latter.

Assembly of ESTs Once the initial groups of pairwise similar ESTs have been formed, the ESTs are assembled into contigs. In this step, sequences of each group are checked for incompatibilities, i.e., stretches of differing nucleotides. In contrast to the initial

grouping, base quality values are now considered. Conflicts between sequences are more likely accepted if one or both sequences have low quality values at the mismatching positions. Consequently, conflicts due to sequencing errors have less impact on the clustering procedure. Finally, all sequences within a cluster that have no unresolved incompatibilities are assembled into a contig with the program CAP3 (Huang and Madan (1999)), and a consensus sequence is generated. For each position in the consensus sequence, the base quality value is calculated from the quality values in the individual ESTs. A base that is confirmed by several independent ESTs is likely to be correct and it thus gets a higher assigned base quality value in the assembled consensus sequence than in the individual ESTs. Conflicts in the consensus sequence are resolved by setting the base with the higher quality value in the underlying ESTs at the corresponding position in the consensus sequence. Since conflicting bases should be treated with caution, they are marked by lower quality values. If all concurrent bases for a conflicting position have an equal quality value, the conflict cannot be resolved and the position gets an assigned quality value of zero.

Quality Clipping of EST Contigs If we are in possession of base quality values, contigs are trimmed with Lucy by removing low quality regions, so that the average error probability of the remaining sequence is 0.025 in each contig. The strategy of performing the quality clipping after the clustering is based on the observation, that also low quality regions, especially at the terminal parts, can contain useful information for connecting ESTs, that share only a short stretch of sequence. However, low quality regions can also prevent overlapping ESTs from being clustered, since the number of mismatching positions can exceed the chosen thresholds. Thus, we repeat the clustering step with TGICL after quality clipping. Afterwards, sequences with a length of less than 100 nucleotides are discarded.
In preprocessed ESTs, low quality regions are already removed, which makes a quality clipping and a second clustering step unnecessary. Furthermore, since in this case all bases in all ESTs have an equally high quality value of 20 (see 4.2.1), all conflicts will be marked as unresolved and will be assigned a quality value of zero in the consensus sequence. A few successive positions with an assigned quality value of 0 in a contig are sufficient to trigger the low quality detection by Lucy. Subsequently, the contig will be divided at the conflicting position and the shorter end of the contig will be discarded, although it contains high quality sequence. Since such conflicts also frequently occur near to the middle of contigs, the decline of sequence data is substantial. Therefore, we do not apply quality clipping to contigs based on ESTs with artificial base quality values.

For the sake of simplicity, we call sequences that passed the assembly step, EST contigs, regardless if they are consensus sequences of several ESTs or are singletons.

4.2.4 Annotation

Usually, the only information available about an EST contig is the sequence itself, and the organism/tissue the mRNA was extracted from. For many applications however, it is desirable to know which gene an EST was derived from. Such information can be used, for example, to extend an existing gene set for phylogenetic tree reconstruction. To assign ESTs to certain genes, it is common to perform a BLAST search (Altschul *et al.* (1997)) with the ESTs against a non-redundant protein database (*nr-pdb*), provided by the National Center of Biotechnology Information (NCBI)[4], for example.
Since ESTs usually do not cover the full-length of the mRNA, the reading frame of the coding sequence contained in an EST is not known. To this end, we perform a BLASTX search against the nr-pdb from the NCBI. BLASTX translates the query nucleotide sequence in all six reading frames into the corresponding amino acid sequences and searches for significant hits for all six virtual protein sequences. However, insertions and deletions, possibly introduced by sequencing errors or during the generation of the cDNA, will cause frame shifts during the virtual translation. This is a fact not acknowledged by BLASTX. A frame shift can split long continuous hits in several short ones, which has consequences for the ordering of the BLAST hits. To circumvent errors related to this matter, we introduced an additional step. We extract protein sequences from the BLAST database appearing in the BLASTX report. Each of the protein sequences is aligned to the EST contig that was used as a query with GeneWise (Birney *et al.* (2004)). This program performs a codon-alignment of a DNA sequence against protein sequences, allowing insertions/deletions in a codon, by introducing gaps to compensate for them. Thus, even in the presence of sequence errors, a correct full-length codon alignment between an EST contig and a protein sequence can be obtained. Consequently, the GeneWise alignment score is a more accurate selection criterion than the BLAST score to choose the most similar protein sequence. Moreover, we obtain a high quality prediction of the coding sequence, its reading frame and the corresponding amino acid sequence. The latter can be used in amino acid based tree reconstructions.
To speed up the annotation process we empirically determined the optimal number of protein sequences an EST contig has to be compared with in order to get the highest possible GeneWise score. It turned out that with a 95% probability, the protein that yields the highest GeneWise score is among the best 25 BLASTX hits (see Section 5.4). Therefore, we only consider the best 25 BLASTX hits for the annotation instead of every hit an EST contig triggered. Furthermore, an E-Value cutoff of 0.0001 is set for the BLASTX search.

[4] `ftp://ftp.ncbi.nlm.nih.gov/blast/db/`

4.3 Notes on the Implementation

4.3.1 Programming

With the exception of the programs included in the TGICL package, all tools mentioned so far are stand-alone. Each has to be started manually requiring a specifically formatted input and generating an output in a defined format. Obviously, processing hundreds of EST collections by sequentially starting ∼10 individual programs by hand, is not efficient; not to mention the necessary reformatting of files between the individual steps. The only way of processing the amounts of data needed to reach the goals of the Deep Metazoan Phylogeny project is by automating the processing. The scripting language PERL[5] is perfectly suited for this task, for several reasons. First, it has superb text manipulation abilities, which makes reformatting of text files easy. Second, it allows powerful process management to start and control the individual programs. Last but not least, it provides interfaces to several advanced database management systems, which are needed to efficiently store and manage such huge amounts of data (see also Section 4.3.2).
Based on these considerations, we wrote a PERL program, EST**CC.PL**, that takes care of the data transfer from and to the database, process management and reformatting of files. The program is built in a modular fashion. For each step, an enclosed module is written which provides flexibility for future modifications. For example as stated in Section 3.3.3, new sequencing technologies are currently entering the field of EST generation. Those might necessitate the incorporation of alternative clustering programs, which can be easily integrated into the pipeline by programming a new module. A detailed overview of all modules is included in the appendix, see Section B.1 and Figures B.1 – B.4 .

Besides the described implementation of the pipeline, we also developed a simplified version which omits the database interface and annotation capabilities. It has been integrated into the HaMStR online tool, as described in Ebersberger *et al.* (2009b).

4.3.2 Data Organization

We organize ESTs in projects on two different levels. An *EST project* is defined as a collection of ESTs from one species that can be processed by our pipeline with the same parameter settings. Besides the cleaned EST sequences, it contains the corresponding quality values, the raw data and related information such as project descriptions and parameter settings used for the cleaning of the ESTs. ESTs within an EST project are typically, but not necessarily, from one data source. EST projects are updateable, i.e. new ESTs from a certain species and data source can be added to an already existing

[5] `www.perl.org`

EST project. However, if the newly generated EST require different parameter settings for the processing, because, for example, a different vector system was used, we set up a new EST project.
Contigs that emerged from an individual clustering form a *clustering project*. Clustering projects contain the consensus sequences of the contigs and all information that is linked to these contigs, such as the corresponding quality values, annotation data, the components of the individual contigs and the underlying EST projects. A clustering project can be based on one or more EST projects, but all ESTs must stem from the same species. This allows to unite ESTs from different data sources. In contrast to EST projects, clustering projects are not updateable. The addition of new ESTs will likely alter existing contigs. By initiating a new clustering project for each clustering, we obtain a precisely defined set of sequence data which can be referenced by a unique identifier in analyses.

4.3.3 Data Storage

To achieve the ambitious goals of the Deep Metazoan Phylogeny project, millions of ESTs from hundreds of taxa have to be incorporated. Organizing such huge amounts of data requires a storage strategy that is beyond handling of simple data files on a hard drive. First of all, searching and extracting specific sequences from a very large collection stored in flat files can be complicated and time-consuming. Furthermore, besides the sequence itself, a whole collection of additional information called metadata is attached to each EST contig. This includes, among other, information about the source, applied cloning strategies, composition of each contig, base quality values and annotations. Storing all information in a single file will quickly lead to file sizes that are not manageable. Distributing this information over different files requires cross-referencing, which can hardly be done efficiently.
Another important aspect is, that local files are only suited for single user access. If modified by several users in parallel, individual changes can easily be overwritten by other users which leads to a loss or inconsistency of the stored data.
Relational Database Management Systems (RDMS, Date (1999)) are designed for such fields of application. Here, the information is organized in tables. The entries in each table can be indexed which allows a quick access to specific entries. Metadata can be distributed over additional tables, which are then connected by cross-references via unique identifiers. The *Structured Query Language* (SQL) allows to formulate complex queries to get multi-layered information from several tables at once. Different relational database management systems exist. Among these are open-source solutions such as

MySQL[6], PostgreSQL[7], but also commercial products like Oracle[8] and Access developed by Microsoft[9]. They all adhere to the SQL standards in general, but differ in detail. We have chosen Oracle, because it is best suited for large scale data storage.
To store the complex data of the DMP project, we designed a database scheme, *dbDMP*, with about 25 tables. These tables contain all the data of individual EST and clustering projects. Figure B.5 in the appendix provides a detailed diagram of the scheme and Section B.2.1, also in the appendix, contains a textual description of each table.
The processing pipeline (Section 4.2) is directly connected to this database and receives its input from and stores its output into the appropriate tables. Raw EST sequences are downloaded from it and the resulting EST contigs are uploaded into the corresponding tables, including processed ESTs and all metadata.

The DMP project has its dedicated website[10]. It provides an interface for the DMP member to the DMP database. Here, the user can gain access to the processed sequence data they have submitted, but also search and download processed public data. In principle, the DMP website provides access to all the data of the DMP project. However, certain data should be kept confidential and visible only for certain members within the DMP consortium. This is to allow the submitter to first evaluate and analyze his or her data before it is made accessible for other researchers. Public data on the other hand should be accessible right from the start by all members. We thus implemented a group based access management system. The DMP users are organized in groups, corresponding to the institutions they are affiliated with. Access to EST and clustering projects can be restricted to specified groups and then later made public within the DMP project. For that we used an Oracle-specific feature, called virtual private database. This technique provides the possibility of specifying the database content, which each user is allowed to access.

4.3.4 Developed Tools

Due to the constant generation of ESTs, the public databases need to be regularly checked for data not yet included in dbDMP. To aid this process, we wrote the PERL script CHECK_EST_TAXA.PL that parses a list containing scientific species names and the number of ESTs available from a data source. Such a list is, for example, provided on the dbEST pages[11]. For every taxon in that list, the script compares the numbers of EST available from the source and present in dbDMP. Based on these comparisons, a PDF

[6] http://www.mysql.com/
[7] http://www.postgresql.org/
[8] www.oracle.com
[9] http://office.microsoft.com/en-us/access/default.aspx
[10] www.deep-phylogeny.org
[11] http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

file is generated that contains a detailed report in which each taxon of the source is listed with the number of ESTs available from that source and the number of ESTs, if any, present in dbDMP. Additionally, it compiles two lists. The first list contains the names of only those taxa that are not yet represented in dbDMP. The second list contains those species for which ESTs are already present in dbDMP but in smaller quantities. A filter mechanism checks the systematics of every taxon and only considers those ESTs for the output that belong to a defined taxonomic classification. The classification can be set by the user and is not restricted to a taxonomy level, e.g. a kingdom name or a genus name are both valid filter terms. The filtering mechanism is helpful if one wants to update EST data of a certain part of the species tree. Based on the report, the user then can select which taxa he or she wishes to incorporate.

The developed pipeline processes EST sequences automatically. However, the ESTs need to be present in dbDMP, before the processing can be started. The integration of ESTs into dbDMP requires several steps. First, the sequence data has to be downloaded from the source, including all metadata such as species name, descriptions of the EST generation and if available, base quality values. Once the data is acquired, either a new EST project has to be initiated, or, if an existing project should be updated, already present data needs to be identified and removed, in order to prevent redundancies. Afterwards, the data needs to be uploaded into the appropriate tables of our database. Doing these steps manually is tedious, especially if hundreds of projects need to be processed. We therefore wrote a further PERL script, FETCH_EST_DATA.PL, which automates the process for data available in dbEST, our main data source (see Section 6.1). As input, it expects a list of scientific species names. Thus, we can feed the output of CHECK_EST_TAXA.PL directly into FETCH_EST_DATA.PL. The latter downloads all available ESTs from dbEST for each specified taxon and generates artificial quality values, because they are not provided by dbEST. If the species is not yet represented by ESTs in dbDMP, a new EST project has to be initiated. For this purpose, the script automatically downloads taxonomy-related information from the NCBI taxonomy database[12] and creates a new entry in the dbDMP taxon table (see Section B.2.1 in the appendix). A new entry is also created in the corresponding EST project table.
If an existing EST project is updated, the script filters the downloaded sequences for identical records in dbDMP and removes affected ESTs from the sequence file accordingly. Finally, the data is uploaded into the appropriate tables of dbDMP and a text file is generated, which contains a shell script to execute the processing pipeline with the right parameter settings.
In combination with those tools, a single user is able to manage and process hundreds of EST projects including millions of EST sequences (see Section 6.2), given sufficient computational resources, of course.

[12] `http://www.ncbi.nlm.nih.gov/taxonomy`

5 Evaluation of the Processing Pipeline

5.1 Introduction

Under optimal conditions, we would expect that the EST processing and clustering pipeline identifies all overlapping ESTs in one project that stem from the same mRNA, and assembles them into one contig. Furthermore, we would expect to see no contig that consists of ESTs from different mRNAs, which are called chimeric sequences. However, certain factors can interfere with the clustering procedure. For example, the overlap between ESTs can be too short to be recognized by the clustering routines, because ESTs cover different parts of the mRNA. Other problems are more on a technical side: The sequence quality is usually poor at the ends of ESTs (see Section 3.3.3), indicating that the sequences are usually unreliable in these regions. Too many sequencing errors can make ESTs appear to be incompatible and consequently, they are not clustered into a contig. The same holds true for unrecognized vector sequences, also found at the terminal parts of the sequences. On the other hand, gene duplications or highly conserved domains present in different genes can lead to a high sequence similarity between two unrelated ESTs, which are then spuriously assembled together.
Although the components and parameter settings of the pipeline were chosen to account for such error sources, we need to consider that errors in the EST assemblies still occur. In order to estimate the error rate, we evaluated our EST processing pipeline using a controlled EST set for which we know the origin of each EST and its position in the gene it was derived from. To gain the position information, ESTs are mapped onto the genome sequence, which will reveal their place of origin. ESTs with overlapping positions of origin should stem from the same gene and thus represent the same mRNA. Consequently, they should be assembled into one contig. After processing such a set of controlled ESTs with our pipeline, we can count the number of ESTs that should have been clustered but are not, and thereby assess the false negative rate. Similarly, we can assess the false positive rate, i.e., the number of ESTs which are clustered into one contig, although they do not share a common origin.

Table 5.1: Evaluation of Clustering Results This table gives the status of each of the 9766 analyzed ESTs, regarding their membership to a contig (rows) and if they overlap with another EST by their genome localization (columns). 'By Position' indicates the number of ESTs that do and do not overlap with other ESTs by their genomic position. 'By Pipeline' gives the number of ESTs assembled into contigs or remaining as singletons.

		By Position		
		Separated	Overlapping	
By Pipeline	Singleton	5988	917	6905
	Contig	127	2734	2861
		6115	3651	9766

5.2 Analysis

From the over 8 million human ESTs present in dbEST[1], we randomly selected 10,000. To identify their exact location in the human genome, we used the program BLAT (Kent (2002)) to align the ESTs with the human genome sequence version 18 (hg18), available at the USCS genome browser[2]. We have chosen human as a reference organism, because its finished genome sequence, together with the availability of large amounts of ESTs, form an optimal basis for our evaluation.
We processed all 10,000 ESTs with our pipeline, using standard parameter settings. Notably, 113 ESTs were discarded because of a too high vector contamination, although the ESTs are marked as being preprocessed. Another 194 ESTs were removed, because after vector and poly-A tail clipping they did not exceed the required length of 100 nucleotides. Finally, 40 ESTs were removed due to a high proportion of low-complexity regions. In total, 9766 ESTs remained for analysis. 2861 of these were assembled into 838 contigs, while the remaining 6905 ESTs remained as singletons. In the following evaluation step, we checked if the genomic origins of all ESTs in a contig have overlapping coordinates. Moreover, we also analyzed whether all ESTs with overlapping genomic localization have been assembled into one contig. Table 5.1 summarizes the results.

127 ESTs were assembled by the pipeline although they do not overlap with any other EST, according to their genomic position. The false positive rate is therefore $(127/6115) \times 100 = 2.1\%$. 917 ESTs were not clustered by the pipeline but do overlap with other ESTs from the set. This corresponds to a false negative rate of $(917/3651) \times 100 = 25.1\%$.

[1] http://www.ncbi.nlm.nih.gov/dbEST/
[2] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/

5.3 Discussion

By clustering a set of 9766 known ESTs, we could assess the error rates of the pipeline. While a false positive rate of 2.1% is acceptable, the false negative rate seems to be very high at first glance. One-fourth of ESTs, that overlap with other ESTs, are not united in a contig. This raises the question of whether the parameter settings of the pipeline could be improved to lower the false negative rate.
In our pipeline, the status of ESTs regarding their membership to a contig is solely determined by the sequences itself. ESTs that show similar sequence parts and that fulfill the criteria described in Section 4.2.3 are considered as overlapping. Specifically, the length of the similar fragment and the number of mismatches allowed are the relevant parameters. Allowing more mismatches and shorter minimal putative overlapping sections would obviously lower the false negative rate. But at the same time it would increase the probability that two unrelated ESTs meet those criteria as well, which would yield more false positives. Hence, the false negative and the false positive rate are connected via these parameter settings. Thus, lowering the false negative rate while maintaining a constantly low false positive rate is not possible. We consider false positives to be much worse than false negatives. In the worst case, ESTs that falsely remain singletons will lead to shorter sequences which contain less informative positions, compared to a contig of several ESTs. On the contrary, chimeric sequences, generated by false positives, consist of sequences that do not share a common evolutionary history. They distort the phylogenetic signal and can lead to false conclusions about phylogenetic relationships. We therefore give preference to a lower false positive rate rather than a lower false negative rate.
Additionally, when determining the status of an EST concerning its membership to a contig by its genomic location, the number of false negatives is expected to be overestimated. So far, we have not taken into consideration that one gene can give rise to several distinct mRNAs due to alternative splicing (e.g. Graveley (2001)). Thus, although ESTs can overlap in the genomic position they map to, they do not necessarily cover the same exons. In such cases, a clustering of ESTs will –and has to– fail, since they represent different transcripts of a gene. Figure 5.1 illustrates this with an example taken from the analysis.
Furthermore, we did not take the length of the overlap into consideration. Theoretically, in our test scenario two ESTs can share only a single nucleotide to be considered as overlapping. But this is of course too short to proof their relationship and to justify their assembly into one contig . For this reason, we consider the error rates as adequate and conclude that our pipeline is well suited to process ESTs for phylogenetic reconstructions.

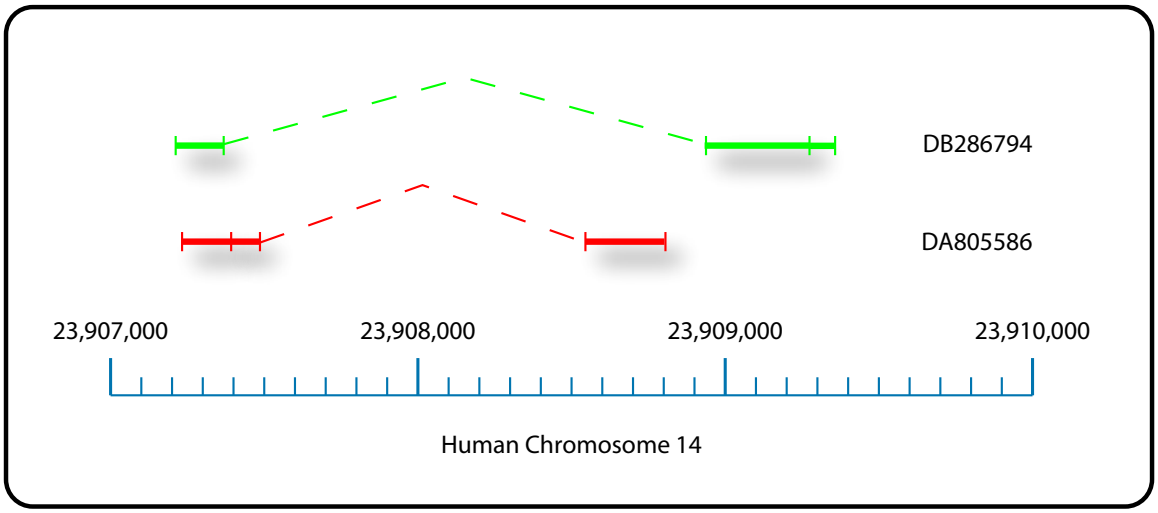


Figure 5.1: Unclustered ESTs with overlapping genome coordinates This example demonstrates that ESTs with overlapping genome coordinates should not necessarily be clustered into one contig. Shown is a graphical representation of BLAT hits for two human ESTs with the Genbank accession numbers DB286794 and DA805586. Both triggered three hits with human chromosome 14, represented as green (DB286794) and red (DA805586) bars. The location of the hit sequences is given in nucleotide positions at the bottom and reveals the intron-exon structure of the gene. Both ESTs share one exon starting at position 23,907,232, but all other exons covered by these ESTs are exclusively found in one of them. Obviously, they represent different transcripts of the same gene.

5.4 Annotation

As described in Section 4.2.4, we have implemented a two-step annotation approach. First, EST contigs are used as queries for significant hits in a non-redundant protein database with BLASTX. Second, each contig/hit pair is codon-wise aligned with GeneWise. Eventually, the description of the hit sequence yielding the highest GeneWise score is adopted as tentative annotation. However, the calculation of each GeneWise alignment costs time, so that by comparing the query contig with all BLASTX hits forms a severe bottleneck for the processing pipeline. Since the BLAST hits are sorted by their E-value, the protein sequences of the most significant hits will be compared first with the query contig. Hence, it is expected that the majority of the protein sequences yielding the highest GeneWise score will be found among the top ranking BLASTX hits. To evaluate how many BLASTX hits have to be considered in order to get the highest scoring GeneWise alignments with a high probability, we annotated the 7,743 human EST contigs obtained by the clustering described in Section 5.2. For each contig, we noted

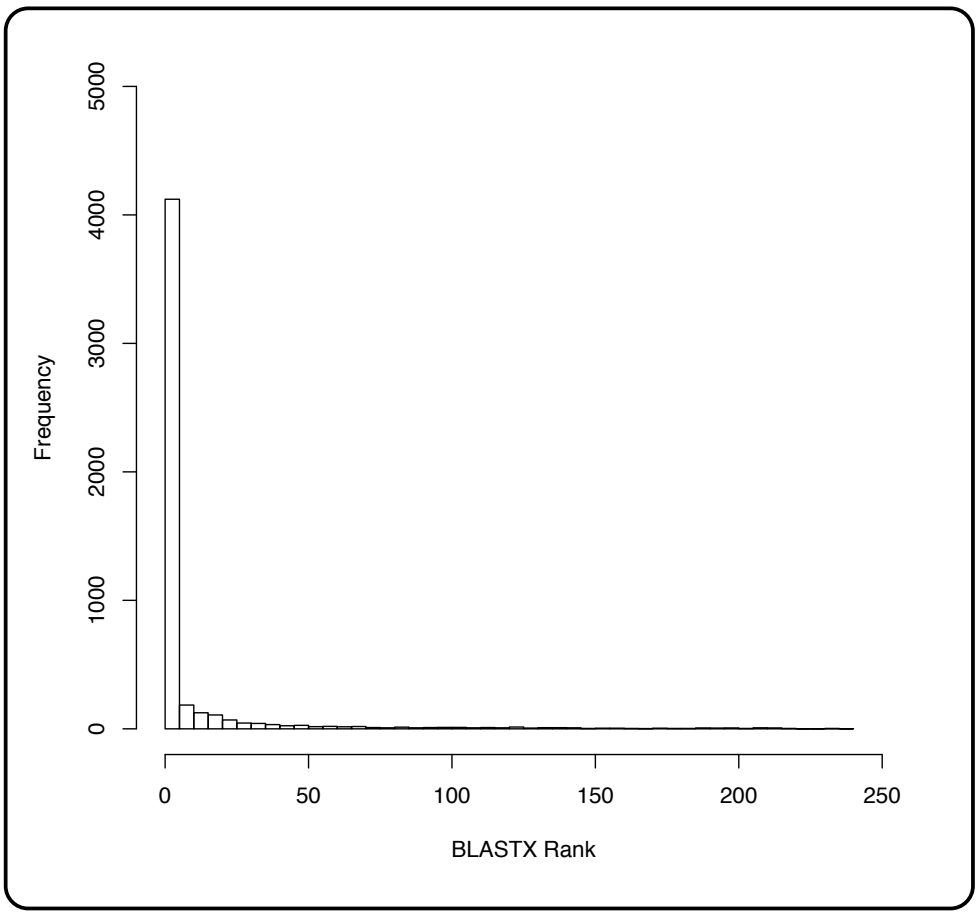


Figure 5.2: Distribution of BLASTX ranks This histogram shows the distribution of BLASTX ranks for proteins that yielded a highest-scoring GeneWise alignment with one of the human ESTs contigs.

---

the BLASTX rank of the protein sequence that yielded the highest GeneWise scores. As protein database for the BLASTX search, we used NCBI's non-redundant protein database, which contains 12% sequences derived from human (Benson *et al.* (2009)). Hence, our results could be biased by the choice of species. We therefore ignored all hits during the BLASTX search that stem from human, to get conservative results.
BLASTX found at least one significant (non-human) hit for 4,854 of the 7,743 contigs. The distribution of the BLASTX hit ranks for proteins yielding a highest-scoring GeneWise alignment is shown in Figure 5.2. The vast majority of proteins that were eventually adopted as annotation are also the top ranking BLASTX hits, although, for one contig the best scoring GeneWise alignment was achieved with the 236th best BLASTX hit. However, for 4610 of these contigs (95%), the most similar protein sequence was found among the top 25 BLASTX hits. We therefore will use only the

best 25 BLASTX hits to determine the highest-scoring GeneWise alignments.

6 Data Processing

With the pipeline at hand, we started to process all the EST data that is publicly available and store it into dbDMP. For this purpose we explored different data sources.

6.1 Data Sources

The most comprehensive sources for public ESTs are the three major sequence databases hosted at the NCBI[1], the EBI[2] and the DDBJ[3]. They contain the majority of all ESTs generated world-wide. Since the content of these three databases is regularly synchronized, it is sufficient to download the data of only one of them. We used the EST division of the NCBI database, called dbEST[4] (Boguski *et al.* (1993)) as our main source to feed dbDMP. ESTs provided by dbEST should be free of contaminations and low quality regions, but are unassembled[5]. We call this state *preprocessed* and handle ESTs in that stage as depicted in Section 4.1.
We further queried the Trace Archive[6], also hosted at the NCBI. Here, unprocessed ESTs are stored, part of which is also provided in a preprocessed form by dbEST.
In addition, we incorporated data from the Gene Index project[7] (Lee *et al.* (2005)). This database provides non-redundant sequence data based on ESTs from dbEST and full-length coding sequences. The construction of a Gene Index has been described in Quackenbush *et al.* (2000). In brief, all ESTs for a species are downloaded from dbEST and cleaned of contaminations and stretches of low sequence quality. Additionally, all known gene sequences for that species are downloaded from GenBank and each coding sequence is extracted according to the annotations found in the GenBank entry. The cleaned ESTs and the coding sequences are clustered and assembled into tentative consensus sequences (TCs). We handle TCs as externally generated contigs and integrate them on the contig

[1] `http://www.ncbi.nlm.nih.gov/`
[2] `http://www.ebi.ac.uk/Databases/`
[3] `http://www.ddbj.nig.ac.jp/`
[4] `http://www.ncbi.nlm.nih.gov/dbEST/index.html`
[5] `http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html`
[6] `http://www.ncbi.nlm.nih.gov/Traces/trace.cgi`
[7] `http://compbio.dfci.harvard.edu/tgi/tgipage.html`

level (see Fig. 4.1), omitting underlying ESTs. Currently, there are Gene Indices of 124 species available. 97 of these have already been integrated into dbDMP.

For some species, sequence data is provided by several public sources. To avoid redundancies in dbDMP, we internally ordered all potential sources hierarchically according to our judgment of the data quality. If there are multiple sources for a certain species, we used the data from the highest ranking source. Sequence data gained from complete genome assemblies form the top of our hierarchy. Genomic data is supposed to be superior for phylogenetic analyses, because ESTs often do not cover the complete coding sequence (see Section 3.4). Hence, if the genome of a species is completely sequenced, we usually refrain from storing ESTs from this species in dbDMP.
Sequence data that stem from the Gene Index project form the second highest level of our hierarchy. The processing pipeline used to edit and cluster the ESTs for a Gene Index shares a lot of components with our own (c.f. Quackenbush *et al.* (2000) and Section 4.2). However, in contrast to our own approach, the sequence data provided by the Gene Index project has been complemented with full-length coding sequences. We therefore do not expect to improve the results if we process the ESTs by ourselves and we would rather save computational resources. Finally, we prefer preprocessed ESTs over unprocessed ones. For example, if ESTs from a species are available from both dbEST and the Trace Archive, we prefer the preprocessed data from dbEST. The processing of ESTs, e.g., the removal of vector, requires detailed information such as the vector sequence, the cloning site and the used adapters. This information is sometimes incorrect or not provided at all (Chen *et al.* (2007)). Moreover, finding the optimal parameter settings to reliably identify and remove contaminations in unprocessed ESTs is tedious, since it can hardly be automated. For this reason, we consider unprocessed ESTs from the Trace Archive only if there are no alternative sources of higher quality.
This hierarchy is more a guideline than a strict rule and some deviations from it can be found in dbDMP for several reasons. In some cases, a genome assembly or a Gene Index was released when the ESTs of the corresponding species had been already integrated into dbDMP. For other species, a lower ranked source has been more heavyset than a higher ranked one. This mainly concerns dbEST and the Trace Archive. Some sequencing groups prefer to submit their data in an unprocessed form only to the Trace Archive, so that for a particular taxon the number of ESTs in the Trace Archive can exceed the number of ESTs available from dbEST by several orders of magnitude. Eventually, there were cases where we doubted the quality of protein sequences obtained from automatic predictions in newly assembled genomes. The identification of protein coding parts in a genome is not always accurate, especially if mainly computational methods are used (Dybas *et al.* (2008)). The annotation of a genome is usually continuously refined by a genome sequencing group and protein sequences gained from well advanced genome sequencing projects are of high quality. For very early releases, however, the sequence

should not be considered error-free. To have an alternative source of sequence data, we therefore decided in individual cases to integrate EST data for species for which a genome sequence is available too.
In addition to public data, we also integrated ESTs that were generated by members of the Deep Metazoan Phylogeny project. Usually, they were delivered as trace files (see Section 4.2.1) and we have processed them accordingly (see Fig. 4.1).

Once the ESTs for a species have been retrieved from a data source, we usually pool these sequences into an EST project (see Section 4.3.2). Thus, all ESTs of a certain species from dbEST are typically handled within a single project, although the ESTs might be uploaded into dbEST by several sequencing groups.

6.2 Growth of dbDMP

In the second half of 2006, we started processing ESTs downloaded from both public data repositories and submitted by members of the DMP project. The continuous growth of dbDMP is illustrated in Figure 6.1. Overall, the sequence data in dbDMP currently encompasses almost 17 million ESTs and about 11.5 million contigs from more than 1,200 species.
While we have already processed plenty of ESTs from taxa belonging to the kingdoms Metazoa, Fungi and Viridiplantae, lower eukaryotes have not been in our focus and, consequently, they are currently only marginally represented in dbDMP (Fig. 6.1).
The pattern of data accumulation over time looks similar for the four taxonomic groups. The amount of sequence data increased rapidly to a certain point, but then the growth stagnated. These locations in the plot mark time points at which we had processed all the public data that were available at that time and which we wished to incorporate into dbDMP. Afterwards, we only processed sequence data that were recently uploaded into the public databases or submitted by DMP members.

Figure 6.2 shows the distribution of sizes for projects currently available in dbDMP. The majority of projects is relatively small. 75% of all projects contain less than ∼11,000 ESTs or, after clustering, ∼8,000 contigs. However, for over 1,200 species sequence data is present (Fig. 6.1C). The application of new sequencing technologies will almost certainly lead to an increase of EST project sizes in the future as initial studies demonstrate (Roeding *et al.* (2009); Gibbons *et al.* (2009)).

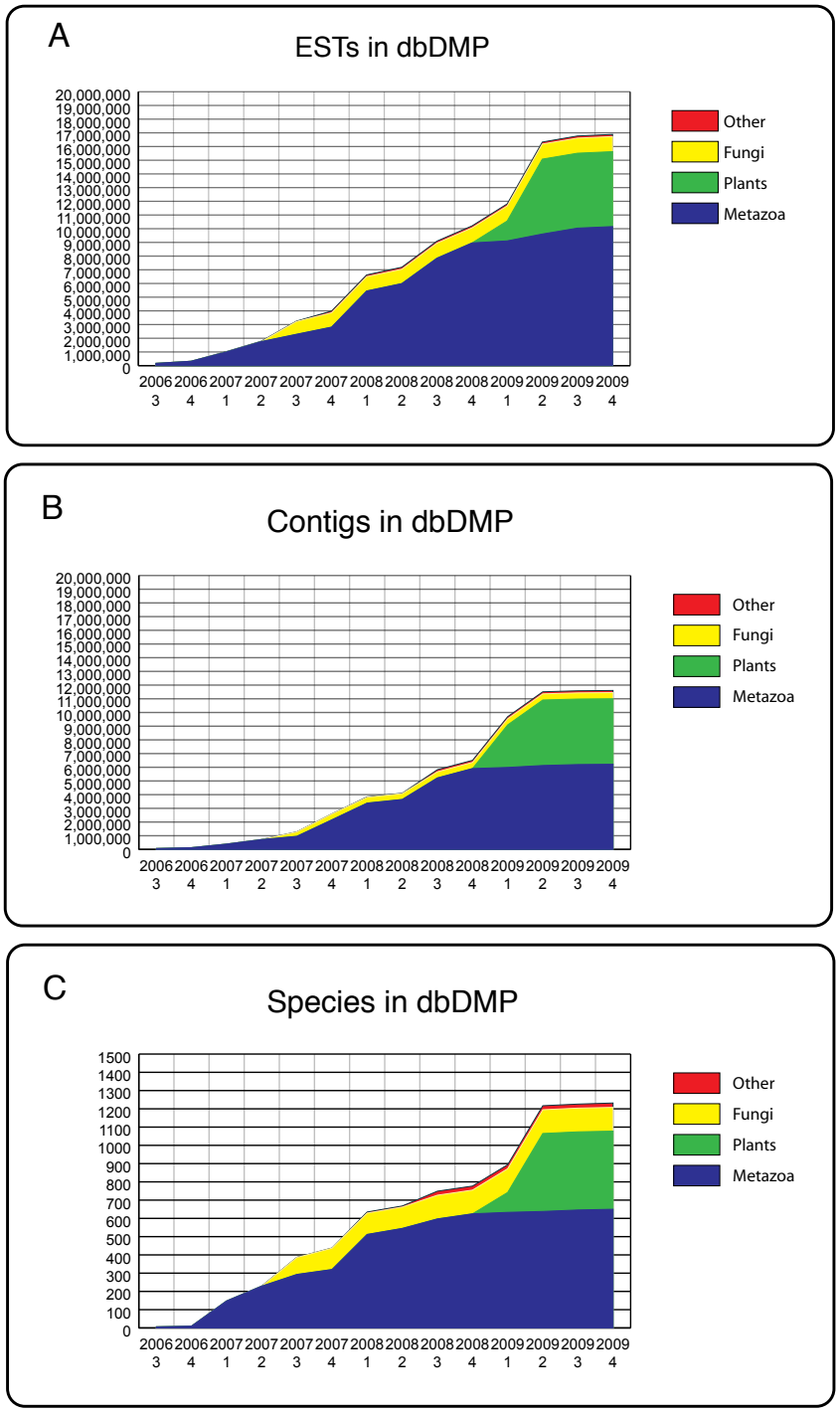


Figure 6.1: Growth of dbDMP The three plots illustrate the growth of dbDMP since its start in the third quarter of 2006. Plot **(A)** shows the number of ESTs for the three kingdoms metazoans, plants and fungi and 'Other'. The latter comprises lower eukaryotic species that belong to neither of the 3 kingdoms, such as protists. The values for each group have been stacked on each other so that the total height of the plot equals the total number of sequences. Plot **(B)** shows the number of EST contigs present in dbDMP, and **(C)** the increase of species represented by sequence data in dbDMP is plotted.

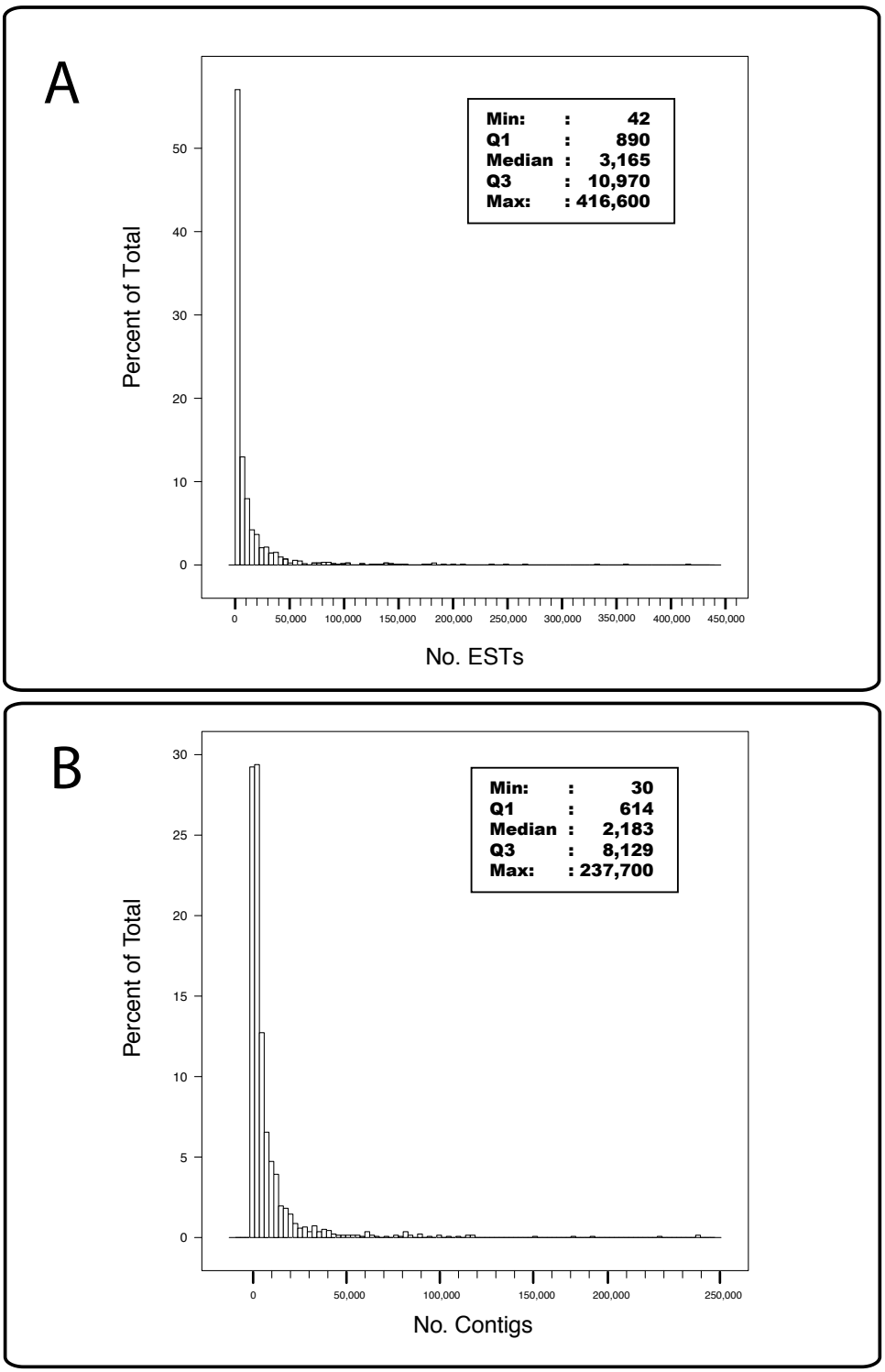


Figure 6.2: Histograms of project sizes The histograms show the size distributions of all EST projects (**(A)**; 1257 projects without Gene Indices) and clustering projects (**(B)**; 1375 projects, including Gene Indices). The total height of the bars in each plot sum up to 100%. The five number summaries (minimum, first quartile, median, third quartile and maximum) have been added to each plot to provide additional insights into the anatomy of these distributions.

6.3 Summary of the Clustering

By clustering the ESTs, we tried to achieve two goals. First, we wanted to reduce the redundancy found among ESTs and second, we wanted to obtain sequence data of higher quality. The redundancy among ESTs can be measured by dividing the number of contigs in a project by the number of ESTs the contigs are based on. This ratio (c/E ratio) will be close to 1 if only a few ESTs were overlapping and assembled into one contig. In contrast, projects that show high amounts of sequence redundancy will have smaller values, because here many ESTs occur in one contig and therefore the resulting number of contigs is substantially smaller than the number of ESTs. To measure the improvement of sequence quality, we calculated the difference between the average sequence length (ASL) before and after clustering for each project. We distinguished between projects based on unprocessed and preprocessed ESTs, because in the latter, low quality regions have already been externally removed and we usually skip the quality clipping step during clustering (see Section 4.2.3). This greatly influences the difference of the EST and the contig sequence length.
Figure 6.3 summarizes the results for all projects. In projects based on preprocessed ESTs, the increase of the ASL, the project size and the sequence redundancy are clearly positively correlated. In small projects, hardly any redundancy is found. Accordingly, most of the resulting contigs consist of only one EST and the ASL remains constant. With increasing project sizes, more redundancy is found. This causes an increase of the average number of ESTs each contig consists of, which in turn increases its sequence length; in some projects by more than 200 bases on average.
Projects based on unprocessed ESTs show a differing picture. Here, the ASL difference is often negative, which reflects the amount of low quality regions found in unprocessed ESTs. With increasing project sizes, the ASL difference shifts to values >0, but there are still some projects with high amounts of sequence redundancy in which the average sequence length of contigs is 200 bases shorter than that of the corresponding ESTs (Fig. 6.3).
This could lead to the impression that contigs based on unprocessed ESTs are in general shorter than those based on preprocessed ESTs, but so far we have not considered the total length of the sequences. Figure 6.4 shows the ASL of ESTs and contigs for each project. The average contig length of the majority of projects lays within the range from roughly 400 to 800 bases. This is true for both projects based on unprocessed and preprocessed ESTs. We therefore conclude that the resulting contigs are of equal quality, independent of whether the ESTs have been externally preprocessed or not. Furthermore, especially in larger projects, the clustering efficiently removes sequence redundancies while improving the overall sequence quality.

While re-inspecting Fig. 6.4, one particular project stands out. The contigs found in

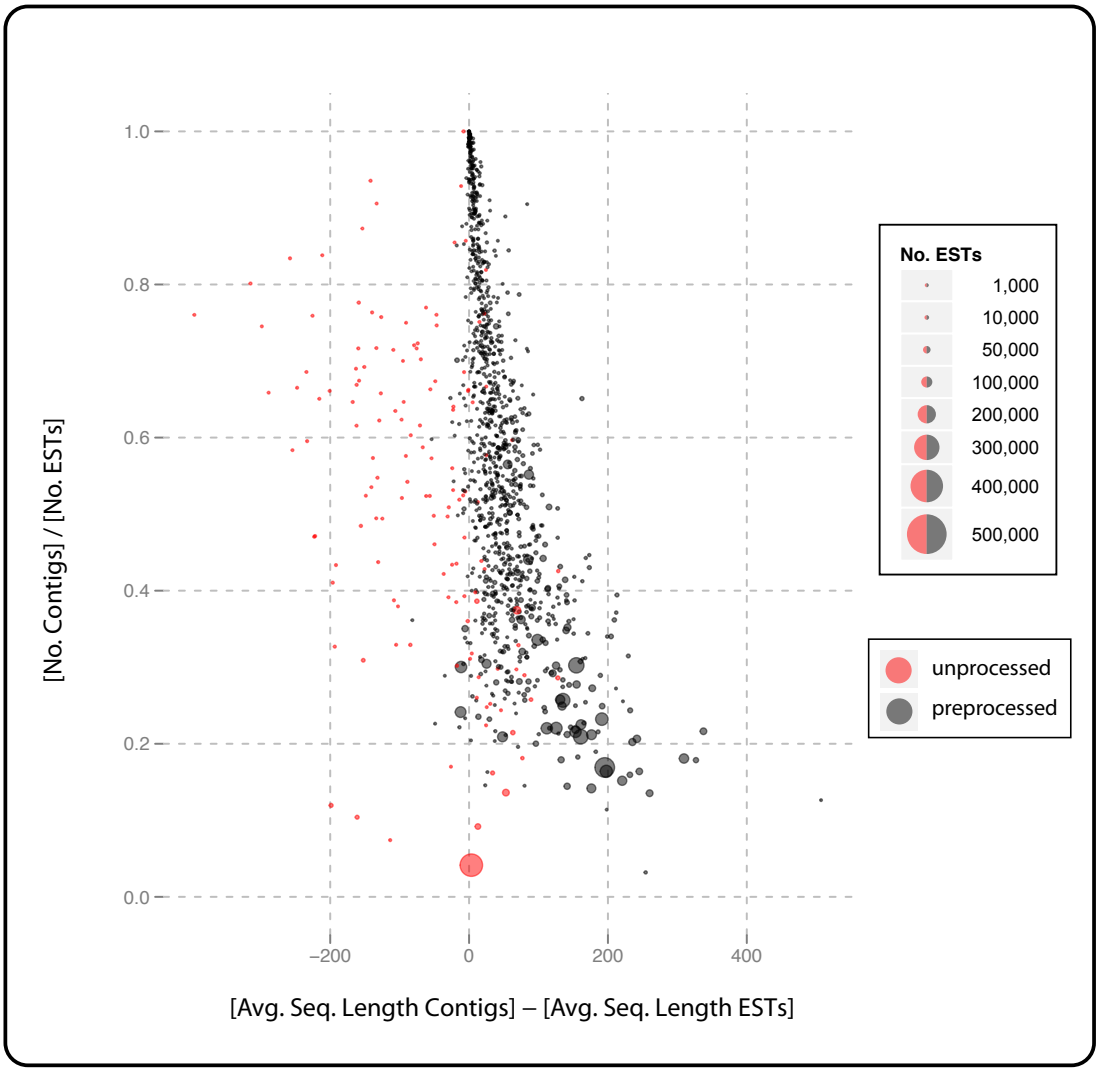


Figure 6.3: Summary of the clustering for each project Each dot represents one of the 1257 clustering projects (excluding Gene Indices). The size of each dot corresponds to the number of ESTs that clustering project is based on, as indicated by the legend on the right. Clustering projects that are based on preprocessed ESTs (1120 projects) are represented by black dots. Red dots mark projects based on unprocessed ESTs (136 projects). The y-axis gives the ratio of the numbers of sequences present in each project before (ESTs) and after (contigs) the clustering. Lower values correspond to higher amounts of redundancies among ESTs. On the x-axis the difference of average sequence length before and after clustering is plotted.

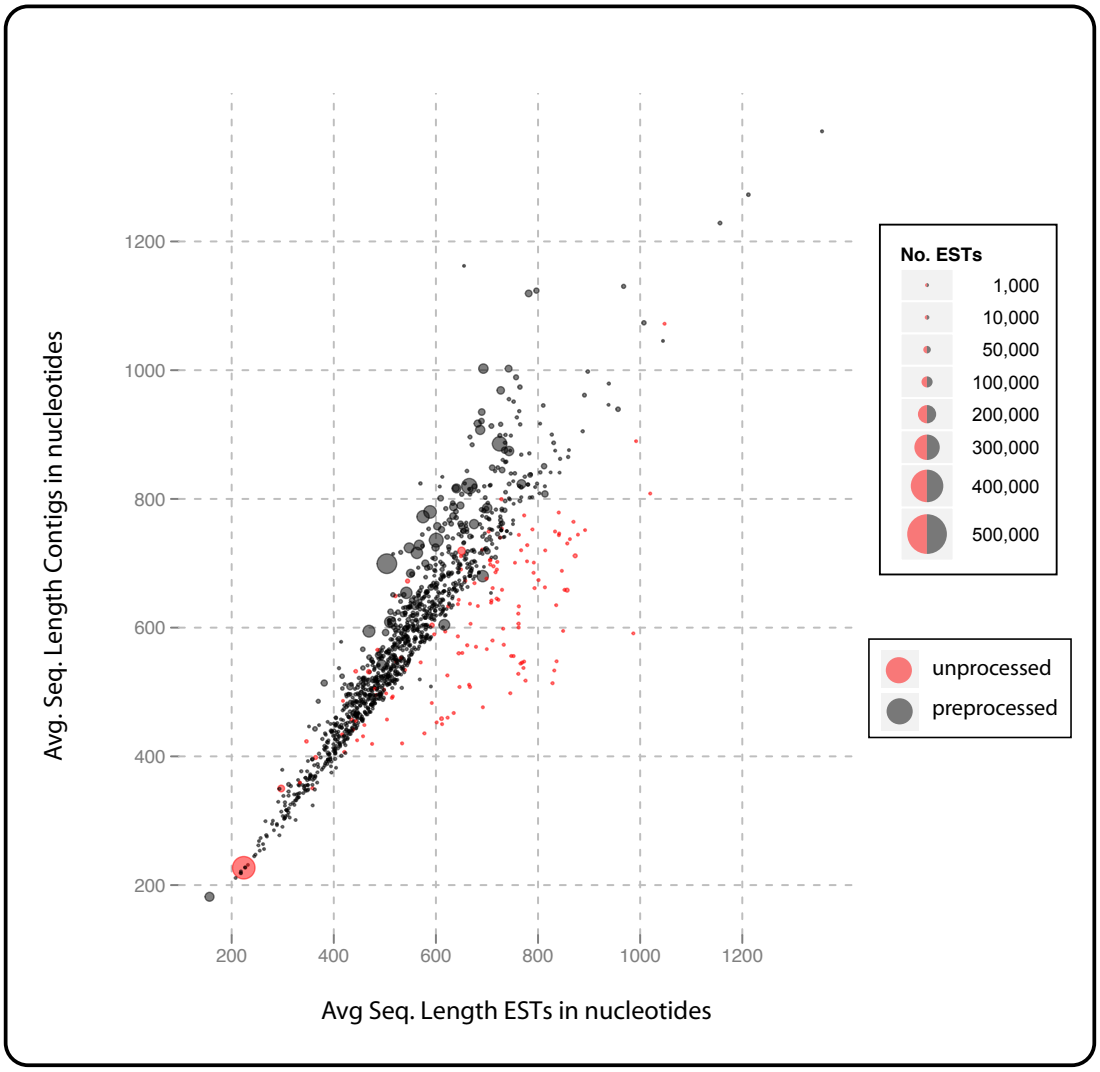


Figure 6.4: Sequence lengths before and after clustering Here we plotted the average sequence length of the ESTs (x-axis) against the average sequence length of contigs (y-axis). The coloring and sizes of the dots is analog to Fig. 6.3

the largest project in dbDMP (*Pandinus imperator*; ∼416,000 cleaned ESTs; located at the bottom left corner in Fig. 6.4) had an ASL of 227 nucleotides, which is only 4 nucleotides longer than the average length of the underlaying ESTs. This is interesting in so far as this particular EST project is one of the two published projects currently present in dbDMP that has been generated by the 454 sequencing[8] method instead of the traditional Sanger sequencing (see Section 3.3.3). On the one hand, shorter ESTs are expected for this project since generally, the 454 sequencing method yields shorter reads than the Sanger sequencing method (Schuster (2008)). On the other hand, this clustering project contains only approximately 17,000 contigs. Each contig therewith consists of roughly 25 ESTs which should intuitively yield much longer sequences.
The other project generated by 454 sequencing, *Phaseolus coccineus* (132,036 ESTs, 40,205 contigs), contains even shorter contigs (contigs: 181.85 nucleotides, ESTs: 156.63 nucleotides on average). To look further into that matter, we compared the contig lengths of all 1,257 projects. It turned out, that the two mentioned projects generated with 454 sequencing are among the five projects that yielded the shortest contigs (data not shown). The other projects that contained equally short contigs were based on less than 700 ESTs each and therefore less redundant. This suggests that ESTs generated with 454 sequencing are not only highly redundant in terms of genes they represent, which could be simply due to the large sizes of the projects, but also seem to cover the same part of a gene, so that the assembly of ESTs does not yield longer consensus sequences. However, such a small sample size is by no means enough to draw further conclusions from this observation. Future projects will shed more light on the quality of sequence data obtained by 454-sequenced ESTs.

6.4 Completeness of Data

To date, we have processed about 17 million ESTs (Fig. 6.1). While this is already an impressive collection, it is only one-fourth of the more than 63 million ESTs currently available from dbEST[9]. Thus, there are still tremendous amounts of public data not yet incorporated into our database. However, the vast majority of these sequences are from species for which already a complete genome sequence or a Gene Index was released. According to our hierarchy of data sources (Section 6.1), we usually do not process these data. This has to be taken into account when evaluating the completeness of data in dbDMP. Due to the large number of EST projects, it is arduous to determine the exact number of ESTs that need to be processed. In order to get a rough estimate, we checked the availability of a genome assembly or a Gene Index for the 40 species with the highest

[8] http://www.454.com
[9] http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

EST counts in dbEST (Table 6.1). Strikingly, for each species a genome assembly and/or a Gene Index is available and for only three of them, we downloaded and processed ESTs from dbEST. The total sum of ESTs of all 38 species listed in Tab. 6.1 which we have not integrated into dbDMP is ∼39 million. In other words, of the 63 million ESTs currently available from dbEST, about 39 million can be safely excluded from the processing, because sequence data of higher quality, such as derived from genome assemblies or Gene Indices, exist. Accounting for this, leaves approximately 7 million ESTs that need to be processed by us. However, this number is still overestimated, because we have integrated Gene Indices of a further 48 species which are not listed in Tab. 6.1 and for which we refrained from processing the underlying ESTs from dbEST. These 48 species account for in total 4 million ESTs (not shown). The remaining discrepancy of approximately 3 million ESTs is caused by EST projects of lower eukaryotes which we have yet to consider.
To conclude, for the time being we consider dbDMP to include the vast majority of EST data that are publicly available.

6.5 Error Sources

When working with data from such a broad variety of sources as presented here, errors in the data are almost unavoidable. Although we make every endeavor to produce data of the highest possible quality, due to the complexity of the data generation, we are not able to account for every possible scenario that could lead to unwanted results.
Here, we report two exemplary cases which led to the (temporary) hosting of faulty data in dbDMP.

6.5.1 Massive Vector Contaminations in EST Projects

dbDMP contains eight EST projects from species of the genus Citrus (citrus fruits) that share three properties. *(i)* All of them were downloaded from dbEST, *(ii)* all have been submitted to dbEST by the same sequencing group and *(iii)* all showed a remarkably high contig/EST ratio despite the fact that there are several thousand ESTs clustered in each project. (Table 6.2). A high contig/EST ratio could be caused by molecular methods applied during the creation of the corresponding cDNA libraries, which are suitable for equalizing the frequencies of cDNAs during the cloning process. These procedures, summarized by the term 'normalization' (Bonaldo *et al.* (1996)), reduce the redundancies among ESTs already on the molecular level and consequently there

[7] `http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html`
[8] `http://www.diark.org/diark/`

Table 6.1: The top 40 entries in dbEST in terms of available ESTs and their representation in dbDMP Column 2 gives the number of available ESTs from dbEST. An 'X' in the third column indicates, that at least a draft version of an annotated genome assembly is available. If a Gene Index has been released, the corresponding species is marked with an 'X' in the fourth column. The fifth column gives the number of ESTs present in dbDMP. The data for this table was extracted from the dbEST summary[7] (October 2009) and the DIARK genome database[8].

Species	No. ESTs in dbEST	Genome Assembly	Gene Index	No. ESTs in dbDMP
Homo sapiens	8,296,280	X	X	
Mus musculus + domesticus	4,852,144	X	X	
Zea mays	2,018,854	X	X	
Bos taurus	1,552,571	X	X	
Sus scrofa	1,536,375		X	
Arabidopsis thaliana	1,527,298	X	X	
Danio rerio	1,481,930	X	X	
Glycine max	1,422,604	X	X	
Xenopus tropicalis	1,271,375	X	X	
Oryza sativa	1,249,001	X	X	
Ciona intestinalis	1,205,674	X	X	
Triticum aestivum	1,067,285		X	
Rattus norvegicus + sp.	1,009,817	X	X	
Drosophila melanogaster	821,005	X	X	
Xenopus laevis	677,806		X	
Oryzias latipes	665,382	X	X	
Brassica napus	643,601		X	
Gallus gallus	600,075	X	X	
Hordeum vulgare + subsp. vulgare	501,614		X	
Salmo salar	494,392		X	
Panicum virgatum	436,535		X	
Phaseolus coccineus	391,138		X	
Canis lupus familiaris	365,909	X	X	
Physcomitrella patens subsp. patens	362,131	X	X	
Vitis vinifera	357,849	X	X	
Caenorhabditis elegans	355,217	X	X	
Ictalurus punctatus	354,434		X	
Branchiostoma floridae	334,502	X	X	334,502
Pinus taeda	328,628		X	
Malus x domestica	324,308		X	
Ovis aries	323,866		X	209,771
Nicotiana tabacum	317,190		X	
Aedes aegypti	301,596	X	X	
Picea glauca	299,455		X	
Solanum lycopersicum	296,848		X	
Oncorhynchus mykiss	287,928		X	
Neurospora crassa	277,147	X	X	
Gasterosteus aculeatus	276,992	X		
Medicago truncatula	269,237	X	X	
Gossypium hirsutum	268,786		X	267,774
Sum	**39,424,779**			**812,047**

Table 6.2: Conspicuous EST projects Listed are eight EST projects from species of the genus *Citrus* with a remarkably high contig/EST ratio (column 3). The second column gives the number of ESTs (after cleaning) and the third column the number of sequences contained in the corresponding clustering projects.

Species	No. ESTs	No. Contigs	contig/EST ratio
Citrus aurantifolia	8196	7007	0.85
Citrus aurantium	14555	11490	0.79
Citrus latifolia	8742	8128	0.93
Citrus limettioides	8181	7806	0.95
Citrus limon	1481	1423	0.96
Citrus reticulata	55886	41652	0.75
Citrus sunki	5200	4622	0.89
Citrus x limonia	10970	9265	0.84

is hardly any redundancy found during the clustering of ESTs. Alternatively, cDNA libraries can be established using subtraction methods (Bonaldo *et al.* (1996)). Here, those cDNA molecules that are already present in a previously created cDNA library are removed before cloning. One can speculate that subtracted cDNA libraries contain mainly clones of lowly expressed genes, because highly expressed genes are most likely included in the cDNA library that is used for the subtraction. Since highly expressed genes are the driving forces behind sequence redundancies among ESTs, we would expect a high contig/EST ratio in projects derived from subtracted cDNA libraries as well.
However, undetected vector contaminations can also lead to such a high ratio. Vector sequences located at the termini of ESTs cause incompatibilities between ESTs that originated from the same gene and prevent their clustering. Given the large number of ESTs in the eight projects, we decided to look further into that matter.
As described earlier, ESTs downloaded from dbEST should have been freed of vector contaminations by the submitting party. Nevertheless, we designed the processing pipeline in such a way that contaminations are still expected and each EST is screened for vector contaminations with multiple methods, all using the NCBI vector database UniVec as a reference (see Section 4.2.2.
In the following, we restrict our more detailed report to the EST project of *Citrus latifolia*, but the presented findings also hold true for the other seven EST projects listed in Tab. 6.2.
We started by inspecting the log files generated during the processing of the *Citrus latifolia* ESTs. In 374 sequences a vector contamination was detected. With less than 5% of all sequences included in that project, this is not alarming and could be due to false positives. Unfortunately, the corresponding entries in dbEST give no information regarding the cDNA library construction and used vector system. The entries are also not

linked to a publication that could provide these details. We thus searched the literature archives and found the paper in which the cDNA library construction and EST generation for these species are described (Targon *et al.* (2007)). It revealed that the corresponding cDNA library was neither normalized nor subtracted. To our surprise, the authors stated that their clustering efforts of that particular EST collection yielded 4,883 unique clusters and therewith only roughly half as much as our own clustering attempt. Alerted by this finding, we further investigated the cause of this discrepancy. It turned out, that the used vector system is not included in the UniVec database. Leftover vector sequences would therefore remain undetected by our processing pipeline. We manually added the vector sequence to our copy of the UniVec database and additionally set up a sequence file containing the specific splice sites as required by the program Lucy (see Section 4.2.2).
With the new parameter settings, our pipeline detected a vector contamination in almost 5,000 of all 8,742 ESTs. The clustering, however, was not improved at all and still yielded a contig/EST ratio of > 0.9. By querying the corresponding log files we found that the detected vector contaminations were almost exclusively located at the $5'$ end of the ESTs, but rarely at the $3'$ end. We therefore carefully inspected the sequence files. It turned out that especially at the $3'$ end, the quality of the sequence seemed to be rather poor, as we could observe multiple examples of mono-nucleotide repeats. We further found some sequences that contained continuous stretches of undetermined bases. These facts suggest that the ESTs are not only insufficiently cleaned from contaminations, but also not cleaned from low-quality regions before submission to dbEST. If this is indeed the case, then the amount of undetected vector contamination could still be substantial, because vector contaminations could be disguised by sequencing errors and therefore not detectable. However, since we are not in possession of the quality values this statement is rather speculative. Furthermore, without the quality values we cannot mark regions of lower quality which are consequently treated as high-quality regions by the clustering routine. In turn, this leads to incompatibilities among related ESTs which should be united in a contig. Eventually, ESTs remain as singletons as we can observe it here.
To conclude, with the currently provided informations, a proper processing of the affected Citrus EST project is not possible. Since we found similar contamination rates in the other projects submitted by the same sequencing group (Tab. 6.2), we have locked access to all the affected data in dbDMP until this issue is resolved.

6.5.2 Foreign Species Contaminations in EST Projects

In a downstream phylogenetic analysis that is not a subject of this thesis, we incorporated protein coding sequence data of the two annelids *Helobdella robusta* and *Capitella sp.* The set of sequences representing *Capitella sp.* were solely based on the available genome

assembly[10] (referred to as *Cap_Ge* hereafter). For *Helobdella robusta*, we used two independent sets of sequences. One is based on genes predicted in the genome assembly[11] (*Hel_Ge*) and the second one is comprised of clustered ESTs that have been downloaded from dbEST (*Hel_EST*). Unexpectedly, the resulting phylogenetic tree suggested that *Hel_EST* is more closely related to *Cap_Ge* than to *Hel_Ge* (not shown). A biological explanation for these results is rather difficult to find. In order to find the cause of this strange observation, we dissected our data set in which 47 genes are represented in each of the three sets (*Cap_Ge*, *Hel_Ge* and *Hel_EST*). For each of the 47 genes, we calculated an alignment with MAFFT (Katoh *et al.* (2005)). Afterwards we used TREE-PUZZLE (Schmidt *et al.* (2002)) to calculate the Maximum Likelihood distances (MLd) for all three possible sequence pairs in each alignment.
One would expect that the MLd between *Hel_EST* and *Hel_Ge* would be close to 0, since the sequence data in both sets originated from the very same species. Hence, the evolutionary distance between both sets should reflect only intraspecies diversity, such as different alleles in the sampled individuals.
However, only ten genes met our expectations and showed a very small MLd value between *Hel_EST* and *Hel_Ge* (Table 6.3). In the remaining 38 cases, the MLd between *Cap_Ge* and *Hel_EST* was the smallest. Moreover, in 36 of these 38 alignments, TREE-PUZZLE could not even detect a distance between the sequences from the *Cap_Ge* and the *Hel_EST* sets and hence, the sequences were identical. This implies that both subsets originated from the same species which proves that there are severe problems with the sequence data. In order to shed further light on this issue, we referred to the original ESTs the involved contigs are based on. Strikingly, all *Helobdella* contigs that showed a closer evolutionary relationship to the *Capitella* genome than to that of *Helobdella* could be traced back to a single cDNA library called `CAWY Helobdella robusta Subtracted Early Library`, which is the origin of 15,350 ESTs. This strongly indicates that this cDNA library is faulty. In order to determine whether the library was simply mislabeled as *Helobdella* library, although it contains exclusively *Capitella* clones, or whether genetic material from both species were mixed during the generation of the library, we aligned each EST with both genome assemblies using BLAT (Kent (2002)). Afterwards, we extracted the best hit for each EST and discarded ESTs that matched with less than 90% of their total length to one of the genome sequences. This filtering was necessary to account for ESTs that stem from genomic regions not covered by the draft genome assemblies. This allowed us to align 11,712 of the 15,350 ESTs to either the *Helobdella robusta* or the *Capitella sp.* genome. Of these, 4,011 matched best to the *Capitella* genome sequence. We therefore consider the CAWY library to be heavily contaminated with foreign genetic material.
dbDMP contains a further ∼86,000 *Helobdella robusta* ESTs from three other cDNA

[10] `http://genome.jgi-psf.org/Capca1/Capca1.home.html`
[11] `http://genome.jgi-psf.org/Helro1/Helro1.home.html`

libraries. They have all been generated by the same sequencing center as those derived from the contaminated library. We therefore repeated the analysis with the remaining ESTs. At least one further cDNA library seems to be contaminated with genetic material from *Capitella* (Tab. 6.4). The few suspicious ESTs found in the CAXZ and CAXA library are most probably false positives. They can be explained by highly conserved genes with a low sequence divergence between *Helobdella* and *Capitella*. If the corresponding gene is not included in the draft assembly of the *Helobdella* genome, the EST will match to the equivalent region in the *Capitella* genome and yet pass the filtering. However, with more than 800 suspicious ESTs found in the CAWX library, we would rather not trust this explanation in this case. We therefore locked access to the *Helobdella* EST project.

6.6 Application of the Data

Currently, access to the data hosted at dbDMP is only granted to the members of the Deep Metazoan Phylogeny project. Within this small community, dbDMP was already of great benefit and the included sequence data formed the foundation of multiple studies. Table 6.5 contains a small collection of publications that incorporated sequence data of dbDMP in particular. Further studies are already submitted for publication or are currently in progress and will be published in the near future.

Table 6.3: ML distances between the three data sets per gene The first column gives our internal gene ID. The second column gives the Maximum Likelihood distance (MLd) between the sequences of the *Helobdella* EST set and the *Helobdella* genome set. The third column contains the MLd between the sequences of the *Helobdella* ESTs and the *Capitella* genome. The last column gives the MLd between both genome sets.

Gene-ID	*Hel_EST* <-> *Hel_Ge*	*He_ESTs* <-> *Cap_Ge*	*Hel_Ge* <-> *Cap_Ge*
21884	0.00620	0.36059	0.35194
22001	0.36658	0	0.36658
22055	0.80213	0	0.80213
22083	0.00691	0.18248	0.20912
22285	0.00581	0.33577	0.38988
22296	0.08106	0	0.07928
22451	0.08648	0	0.08648
22468	0.27080	0	0.28054
22490	0	0.18226	0.18093
22551	0	0.23933	0.24622
22560	0	0.32084	0.31834
22568	0.39673	0	0.39673
22583	0.00373	0.30311	0.30964
22603	0.16587	0.00426	0.15984
22638	0.05049	0	0.05032
22664	0.15740	0	0.15740
22679	0.35132	0.00537	0.36461
22736	0.21343	0	0.21842
22853	0.11895	0	0.13804
22910	0.31946	0	0.31627
22979	0.25542	0	0.26230
23035	0.26165	0	0.26165
23170	0.35491	0	0.35330
23221	0.31637	0	0.32603
23273	0.15063	0	0.14985
23285	0.57582	0	0.56921
23290	0.06987	0	0.06987
23444	0.39401	0	0.39401
23477	0.20284	0	0.21544
23495	0.13392	0	0.13186
23513	0.46554	0	0.45867
23526	0.22674	0	0.22674
23553	0	0.09677	0.14896
23599	0.14366	0	0.14366
23680	0.35097	0	0.34088
23758	0.31041	0	0.30921
23824	0.49478	0	0.48811
23888	0.26257	0	0.26229
23909	0.25191	0	0.25796
23950	0.29353	0	0.30035
24038	0.37151	0	0.37784
24074	0.36823	0	0.37736
24115	0.00508	0.45738	0.48354
24116	0.11945	0	0.11945
24143	0.18201	0	0.17999
24170	0.50364	0	0.50364
24212	0.18550	0	0.18550

Table 6.4: Contamination in the *Helobdella* EST collections In column 1 we listed the names of the cDNA libraries that gave rise to all publicly available *Helobdella* ESTs. Column 2 contains the total number of ESTs from each cDNA library present in dbDMP. Column 3 gives the number of EST we could unequivocally assign to either *Helobdella robusta* or *Capitella sp.* via BLAT search against the genome sequences. The last column shows the number of ESTs which best BLAT hit was triggered by the *Capitella sp.* genome sequence and therefore are contaminations of the *Helobdella robusta* cDNA libraries.

cDNA Library	No. of total ESTs	Assignable	Best hit to *Capitella*
CAWX	33,118	26,817	802
CAWY	15,350	11,712	4,011
CAXZ	25,208	21,200	17
CAXA	27,683	22,587	8

Table 6.5: Studies based on data from dbDMP This table gives a brief overview on published studies that incorporated sequence data from dbDMP. The first column gives the reference, the second column the subject of the paper.

Reference	Topic
Simon *et al.* (2009)	Phylogeny of basal Pterygota (winged insects)
Ebersberger *et al.* (2009a)	Phylogeny of fungi
Witek *et al.* (2008)	Phylogeny of Syndermata (Rotifera and Acanthocephala)
Roeding *et al.* (2007)	Phylogeny of metazoa
Ebersberger *et al.* (2009b)	Phylogeny of fungi
Bleidorn *et al.* (2009)	Placement of Myzostomida within the metazoan species tree
Struck and Fisse (2008)	Placement of Nemertea within the metazoan species tree
Helmkampf *et al.* (2008)	Phylogeny of Lophotrochozoa
Philippe *et al.* (2009)	Phylogeny of basal metazoa
Hausdorf *et al.* (2007)	Placement of Bryozoa
Roeding *et al.* (2009)	Phylogeny of Arthropoda

7 Orthology Assignment

7.1 Introduction

The principle of reconstructing evolutionary relationships of species, their phylogeny, is based on a simple idea. Markers, such as morphological characters or DNA sequences, are used as representatives for whole species. The evolutionary history of these markers can be reconstructed by comparing the character states of each marker in different organisms in the light of a chosen model of evolution. To be able to infer the evolutionary history of species from the history of a marker, it is crucial that both histories are tied together. A split in the lineages of the markers must be coincide with a split in the lineages of the species (speciation). If this requisite is not met, false conclusion about the relationships of the species are drawn.
Genes which lineages split due to a speciation event are called *orthologs* (Fitch (1970)). Their evolutionary history is thus congruent to that of the species they are found in, and therefore their sequences can be used as markers for phylogeny reconstructions. In contrast, genes that arose by a gene duplication event within a common ancestor, called *paralogs* (Fitch (1970)), must not be used (Fig. 7.1). Identifying orthologs in EST data is challenging, because the generation of ESTs is not directed towards preselected genes, but a random process (see 3.3.2). By that, in most of the cases the sequences themselves are the only information available.
Several approaches have been developed that identify orthologs only with the information provided by the sequences themselves, for example by comparing pairwise sequence similarities. A widely used strategy is to perform a bidirectional BLAST search (Altschul *et al.* (1997)), also referred to as reciprocal BLAST. The best hit for every sequence of a species A in another species B is determined. Afterwards each best hit sequence is used as query for a BLAST search with species A as target. Sequence pairs that are each other's best BLAST hit are assumed to be orthologs. This strategy has been extended to deal with more complex gene families or to define groups of orthologous sequences from more than two species (e.g. Remm *et al.* (2001); Li *et al.* (2003)). However, all these methods require that sequence data is available for all genes in all species under consideration. Otherwise, the orthology of reciprocal best BLAST hits is not guaranteed (Fig. 7.2). Consequently, orthology prediction methods based on the reciprocal BLAST hit criterion should not be used for EST data.

Another approach to identify orthologs on the sequence level would be phylogeny-based methods (e.g. Zmasek and Eddy (2002)). Here, orthology of sequences is assumed when a phylogenetic tree of the considered sequences is congruent to the tree of the species the sequences are derived from. These methods can be applied even if the sequence data is only partial, but they require the knowledge of the true species tree. This automatically disqualifies them for an application in phylogenetic studies, since revealing the species tree is the aim of the analysis.
To be able to incorporate EST data for phylogeny reconstructions, we developed a method, HaMStR, that reliably identifies orthologous sequences to a predefined set of genes within an EST collection or a proteome. HaMStR has been described and evaluated in detail in Ebersberger *et al.* (2009b). In the following we explain the algorithm.

7.2 HaMStR Algorithm

7.2.1 Step 1: Defining a Gene Set for the Ortholog Search

7.2.1.1 Generation of Core-Orthologs

As input we introduce the primer-taxa (set) where each taxon is completely sequenced and where the phylogeny of the primer-taxa is undisputed. Standard orthology prediction tools such as InParanoid (Remm *et al.* (2001)) or OMA (Roth *et al.* (2008)) can be applied to identify genes with orthologs present in all primer taxa, the so called core-ortholog groups. We compute orthologs for each pair of taxa with InParanoid (Remm *et al.* (2001)) or used pre-compiled orthology assignments provided by the InParanoid database (Berglund *et al.* (2007)). The pairwise orthology predictions are subsequently extended to include all primer taxa by using a criterion of transitive closure (InParanoid-TC). This approach has the advantage of generating ortholog clusters, with only one sequence per taxon. Such clusters can then be directly used for downstream standard approaches of phylogeny reconstruction. In brief, we order the n primer-taxa such that taxon $i+1$ is the closest relative to taxon i in the species tree, for $i = 1, 2, ..., n-1$, where ties are broken randomly. For each protein in taxon 1 we carry out the following loop:

a We identify the corresponding InParanoid-orthologs in taxon 2. If more than one co-ortholog exists, we choose the one with the highest InParanoid-score.

b With the protein from taxon 2 we identify the ortholog with the highest InParanoid-score in taxon 3. This procedure continues until the ortholog-pair for taxon $n-1$ and n has been determined.

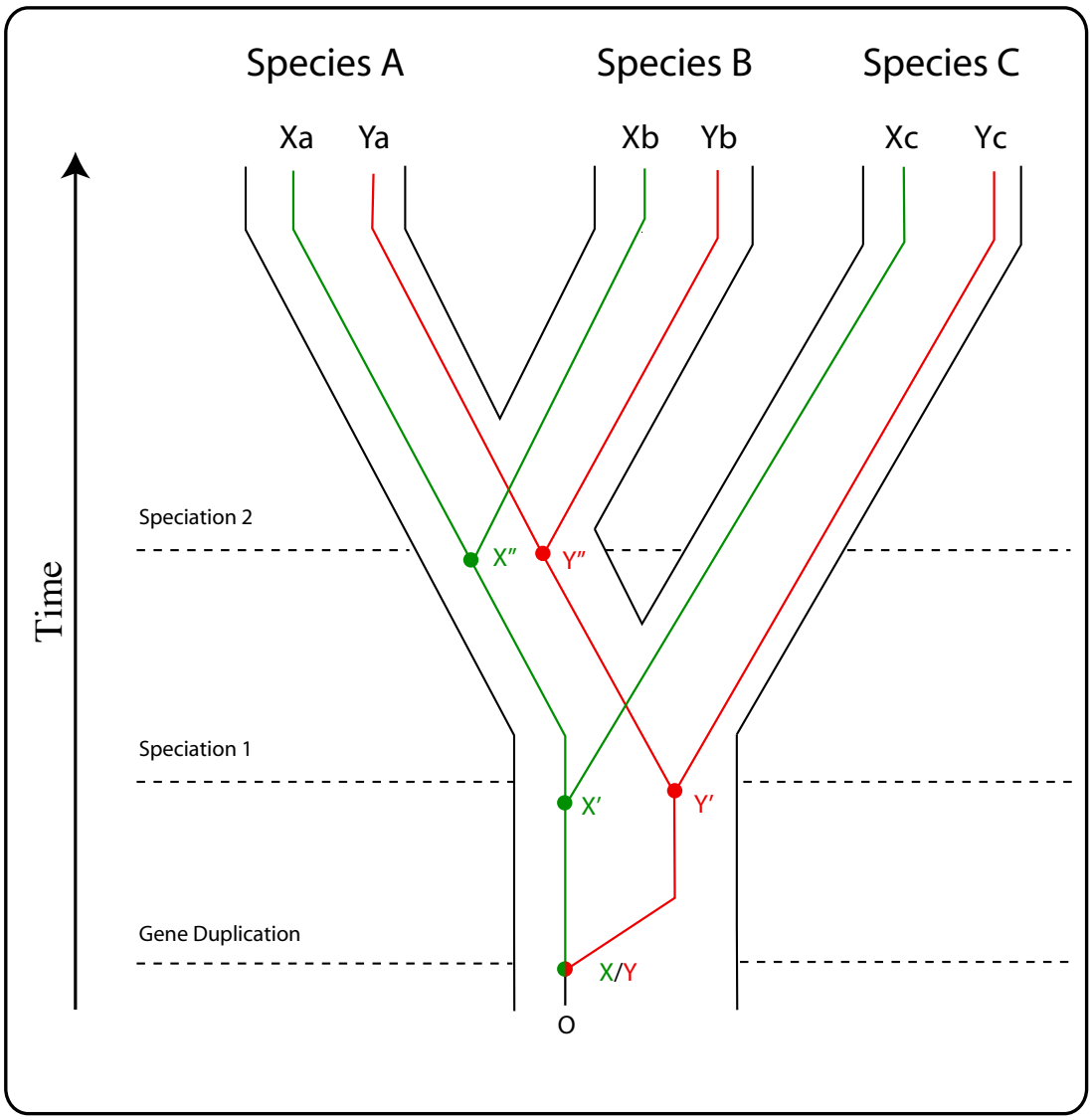


Figure 7.1: Concept of Orthology and Paralogy The black tree illustrates the evolutionary relationships of three species *A*, *B* and *C*. Within each species, two genes ((Xa,Ya), (Xb,Yb) and (Xc,Yc)) are present, which lineages can be traced back to a common origin (O) in the ancestor of all three species. All splits in the green gene tree, connecting Xa, Xb and Xc, are caused by speciation events. Xa, Xb and Xc are therefore called orthologs. The same holds true for the genes that are connected by the red gene tree (Ya, Yb and Yc), which are orthologs as well. In contrast, the shared origin of both gene groups is the gene duplication (X/Y). This makes each gene of the green lineage paralogous to each gene of the red lineage and *vice versa*. If genes from both groups are mixed for a tree reconstruction, the obtained tree topology may not be congruent to the species tree. For example a tree based on the sequences of Xa, Yb and Yc will support a grouping of species *B* and *C* to the exclusion of species *A*, because the evolutionary distance between Yb and Yc is shorter than between Yb and Xa.

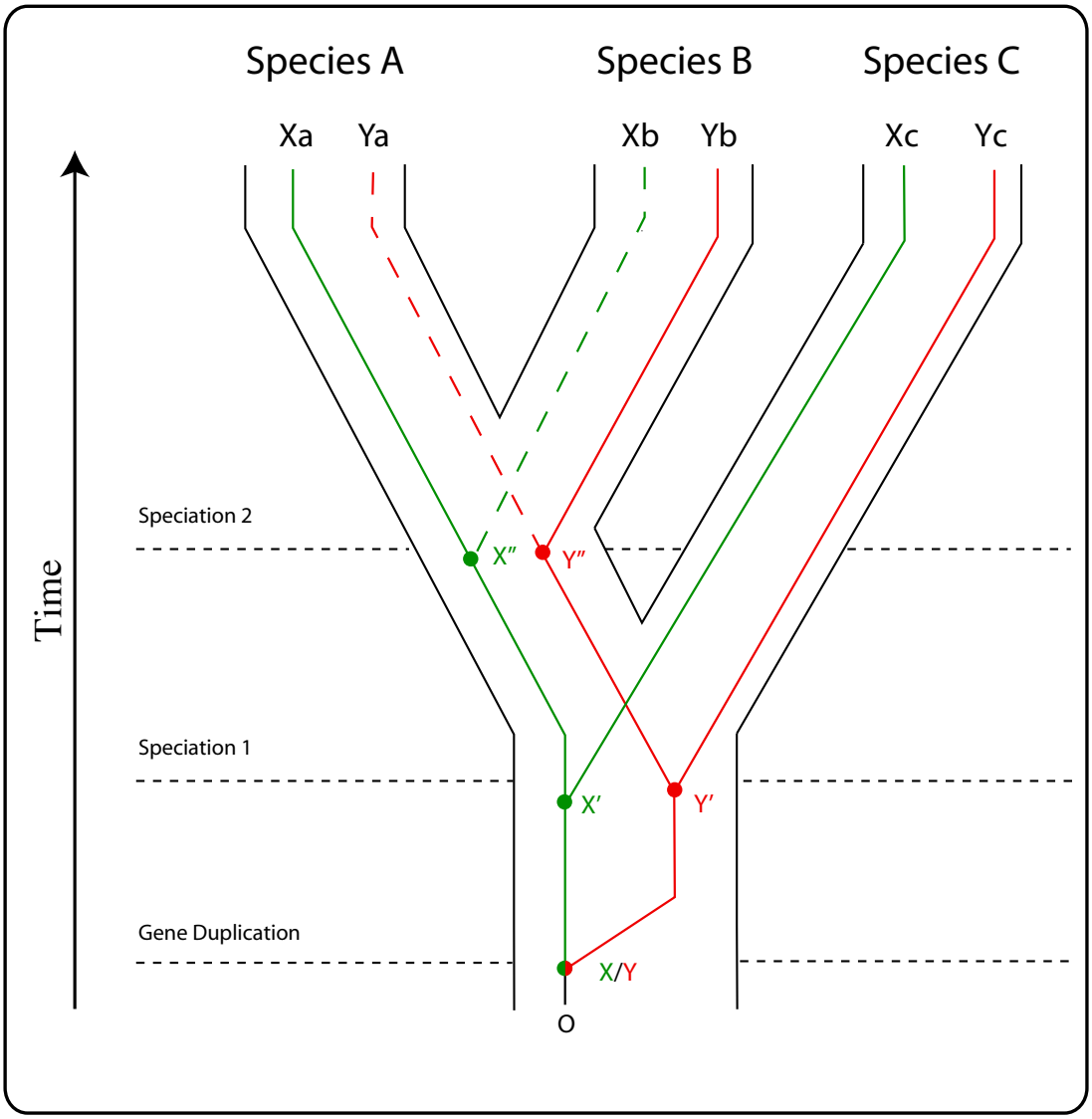


Figure 7.2: Orthology assignment with incomplete sequence data In this scenario, the sequence data for species A and B is incomplete. Only the genes connected by solid lines are available, because, for example, not sufficient ESTs have been generated yet, or the genome from which the individual genes have been predicted is still in draft status. A search for the most similar sequence to Xa in species B will result in Yb. A search for the most similar sequence to Yb in species A will confirm that Xa and Yb are the most similar sequences. This implies, that Xa and Yb are orthologs, although they are not. A similar situation occurs if the genomes of species A and B are completely sequenced, but genes Ya and Xb got lost during their evolution. In this case the reciprocal BLAST strategy will also wrongly predict Xa and Yb to be orthologs. This phenomenon is known as *hidden paralogy*.

c The circle is then closed by identifying the ortholog to the protein from taxon n in taxon 1.

d If the two proteins in taxon 1 at the beginning and the end of the round trip are identical, we keep the set of orthologs and call it core-ortholog. Else, we discard the proteins.

We end up with a collection of core-orthologs representing genes present in all primer-taxa. In each individual core-ortholog each taxon is represented exactly once. From the initial ordering of the primer-taxa it follows that the newly added sequence in each step is an ortholog to all other sequences already in the candidate cluster. The final closure step (**c**) excludes pathogenic cases resulting from hidden paralogy (Scannell *et al.* (2006), Fig. 7.2).
InParanoid-TC has one clear advantage over other orthology prediction programs. The pair-wise ortholog predictions from InParanoid can be pre-computed and stored for a set of taxa. From this data core-orthologs can be rapidly generated for any subset of primer-taxa without further computation.

7.2.1.2 Generation of Profile Hidden Markov Models

For each sequence cluster in the core-orthologs, the sequences are aligned using MAFFT Katoh *et al.* (2005) with the options `--maxiterate 1000` and `--localpair`. The resulting multiple sequence alignments, comprising the n sequences from the primer-taxa are then converted into a profile Hidden Markov Model (pHMM) (Durbin (1998b)). The programs HMMBUILD and HMMCALIBRATE from the HMMER package[1] are used for building, training and calibrating the pHMMs. Each core-ortholog is now represented by a pHMM.

7.2.2 Step 2 Extension of Core-Orthologs

We now extend the core-orthologs with data from additional taxa, the query-taxa. As data may serve translated ESTs, or protein sequences inferred from either complete or partially sequenced genomes.

7.2.2.1 pHMM Search

We use a fast implementation of the HMMSEARCH algorithm[2] to search protein sequence data from the query-taxon for matches to the individual pHMMs. If ESTs are the data

[1] `http://hmmer.janelia.org`
[2] `http://www.clccell.com`

for the query taxa, the individual ESTs are translated in all six reading frames prior to the search.

7.2.2.2 Re-BLAST and Orthology Prediction

To determine the orthology status of the hmmsearch hits, we use a reciprocity criterion (Fig. 7.3). Each hit is compared by BLASTP (Altschul *et al.* (1997)) to the proteome of one of the primer-taxa, the so-called reference-taxon (Proteome F in Fig. 7.3). Ideally, the reference-taxon should be the closest related primer-taxon to the query-taxon. If the protein of the reference taxon that contributed to the pHMM provides the best BLASTP hit, then the hmmsearch hit is added to the corresponding core-ortholog. Otherwise, it is discarded. We note that the reciprocity criterion is also fulfilled when the reference protein is among the lower ranking BLASTP hits, but has the same score as the top listed hit in the BLASTP output.

7.2.2.3 Post-processing of ESTs

To account for possible frame shifts caused by sequencing errors in ESTs we use GeneWise Birney *et al.* (2004) to generate a codon-alignment for the EST and the protein sequence of the reference-taxon. This alignment determines the coding part together with the reading frame in the EST.

7.3 Exploring the Potential of Compiling Phylogenetic Data Sets

To get robust results, phylogenetic analyses need a data basis of a certain size (Rokas *et al.* (2003b)). Furthermore, the data should be as complete as possible, i.e., orthologous sequences for the considered genes should be available for all species under consideration, although some missing data is tolerable (Philippe *et al.* (2004)). However, it is very unlikely, that every gene of a core-ortholog set will be found in every species, either because some genes have been lost during evolution, or because they are not covered by available sequence data. In this section, we explorer the potential of HaMStR and the sequence data stored in dbDMP to compile data sets for large-scale phylogeny reconstructions.

We have created four core-ortholog sets, using different primer-taxa. Table 7.1 lists the primer-taxa and the number of core-orthologs for each set. If available, the pairwise

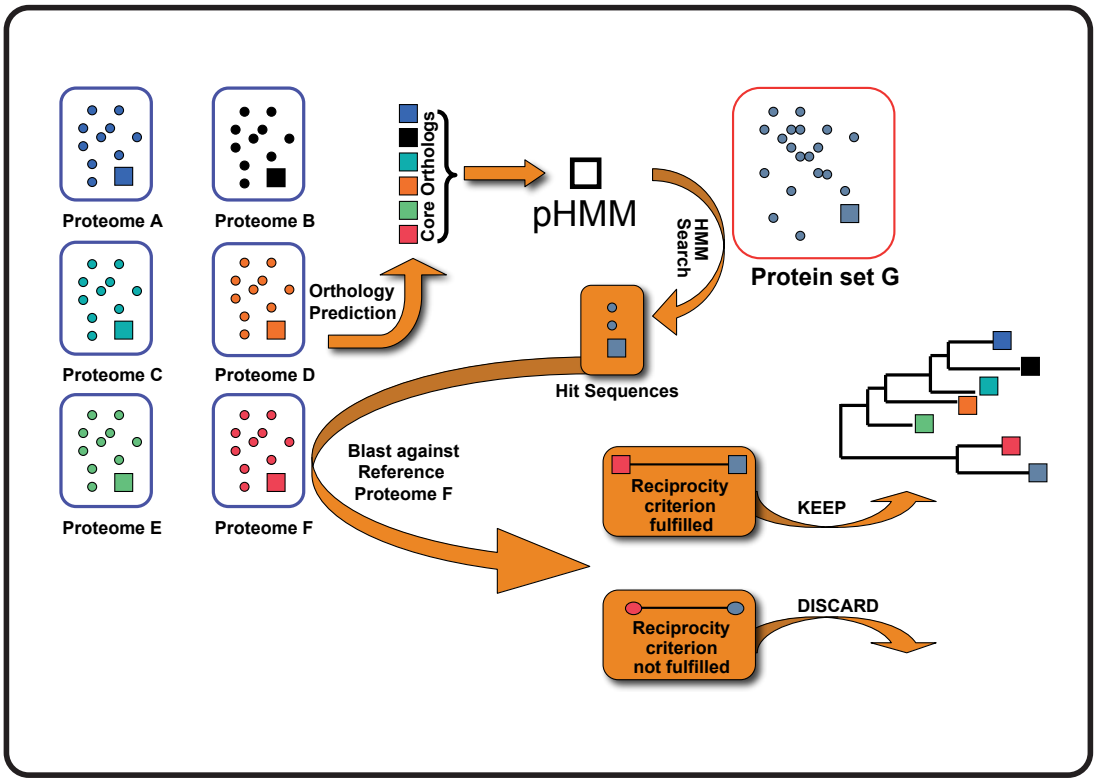


Figure 7.3: Workflow of the HaMStR Approach Standard orthology prediction tools are used to identify orthologous groups, the so called core-orthologs, for a set of completely sequenced primer taxa (Proteome A - F). The sequences in a core-ortholog are aligned and converted into a profile HMM (pHMM). A compilation of protein sequences or translated ESTs from a taxon not included in the primer-taxa (Protein set G) is searched for hits with the pHMM. The resulting candidates display features that are characteristic for the protein modeled by the pHMM. To determine the orthology status of the candidates, we introduce a reciprocity criterion. Each candidate is compared by BLASTP with the proteome of one of the primer-taxa, the so-called reference-taxon (Proteome F). If the best BLASTP hit sequence from the reference taxon corresponds to the protein that contributed to the pHMM, the candidate is called candidate-ortholog, else it is discarded.

orthologs have been extracted from the InParanoid database version 6 (Berglund *et al.* (2007)). Otherwise, we calculated them with the program InParanoid (Remm *et al.* (2001)).

The *Modelorganism set* is based on the proteomes of four metazoan species and the fungus *Saccharomyces cerevisiae* (Tab. 7.1). By including the fungus, we restricted the selection to genes that were already present in the last recent common ancestor of animals and fungi. Potentially, these genes can be found in any metazoan or fungal species and

hence, we can use the *Modelorganism set* to search for orthologs in the majority of all clustering projects present in dbDMP with the exception of plant projects.
The three other core-ortholog sets are designed to identify orthologs in particular groups within the Metazoa, namely lophotrochozoan, arthropods and chordates. Accordingly, the sets are called *Lophotrochozoa*, *Arthropoda* and *Chordata* set.
In contrast to the *Modelorganism set*, these sets include genes that are not necessarily present in every metazoan species. On the one hand, this restricts their use to scan for orthologs in clustering projects of species that belong to the same part of the species tree as the specific primer-taxa do. On the other hand, with these sets additional orthologs can be found, that are not included in the *Modelorganism set*, either because they are too diverged to infer orthology over large evolutionary distances, or because the presence of these genes is restricted to the individual groups. All four core-ortholog sets are also provided as downloads at the HaMStR download page[3].

We used HaMStR and the four core-ortholog sets to search for orthologs in hundreds of clustering projects in dbDMP (Tab. 7.1). Obviously, the number of genes for which orthologs can be identified in a certain clustering project depends on the number of genes that are tagged by the ESTs. To get an overview of how many orthologs can be expected, we plotted the number of genes found by HaMStR with the four core-ortholog sets per clustering project as function of the project size (Fig. 7.4). Since all ESTs that stem from one gene should be clustered in one contig, the number of contigs equals the number of genes, that are represented by the sequence data of a clustering project. For that reason we here use the number of contigs rather than the number of ESTs as measure of the project size.

The number of assigned genes steeply increases with growing project sizes for all four core-ortholog sets and finally converges at a maximum (Fig. 7.4). The maximum should equal the number of genes included in each core-ortholog set (3rd column in Tab. 7.1). Notably, the variance of detected genes given a certain size is also increased for larger projects. This makes accurate predictions about the expected number of genes to be found difficult. The values for the *Lophotrochozoa set* are clearly shifted upwards compared to those of the three other sets. For the first, up to 2,000 genes can be detected in a project, whereas the number of detected genes for the other sets never exceeds 1,000 (Fig. 7.4). This is probably only due to the different number of genes included in the four core-ortholog sets. The *Lophotrochozoa set* encompasses two to even three times as many genes than the other three sets (see 3rd column in Tab. 7.1). From this perspective it is expected that on average more genes of the *Lophotrochozoan set* can be assigned in a clustering project than genes of the other sets.
To account for the different core-ortholog set sizes, we normalized the values by dividing the number of assigned genes per project by the total number of genes included in each

[3] `http://www.deep-phylogeny.org/hamstr/download/datasets/`

Table 7.1: Core-Ortholog sets The first column gives the name of the set. In the second column the primer taxa of each set are listed in the order of the transitive closure. Species written in bold comprise the outgroup of a primer-taxa set. They have been included to ensure that all genes were present in the last recent common ancestor of the clade each core-ortholog set was designed for. In those instances the transitive closure was satisfied if an ortholog was found in at least one outgroup taxa of a primer-taxa set, indicated with a '/'. In the third column the number of genes included in each set is written. The fourth column contains the number of clustering projects present in dbDMP that have been screened with the corresponding set. The last column gives the number of searched projects for each core-ortholog set, that have a size of <2,000 ESTs.

Name	**Primer Taxa**	**No. of genes**	**Searched Projects**	**Searched Projects <2,000 ESTs**
Modelorganism	*Homo sapiens* *Ciona intestinalis* *Drosophila melanogaster* *Caenorhabditis elegans* ***Saccharomyces cerevisiae***	1035	606	321
Lophotrochozoa	*Lottia gigantea* *Helobdella robusta* *Capitella capitata* *Schistosoma mansoni* ***Apis mellifera / Daphnia pulex / Caenorhabditis elegans***	2340	112	63
Arthropoda	*Daphnia pulex* *Bombyx mori* *Tribolium castaneum* *Apis mellifera* *Drosophila melanogaster / Aedes aegypti* ***Lottia gigantea / Capitella spec.*** ***Caenorhabditis elegans / Caenorhabditis briggsae / Caenorhabditis remani*** ***Xenopus tropicalis / Tetraodon nigroviridis*** ***Homo sapiens***	776	236	146
Chordata	*Homo sapiens* *Mus musculus* *Canis familiaris* *Monodelphis domestica* *Xenopus tropicalis* *Gallus gallus* *Danio rerio* *Ciona intestinalis* ***Strongylocentrotus purpuratus*** ***Apis mellifera / Drosophila melanogaster / Caenorhabditis briggsae / Caenorhabditis remanei***	1000	137	65

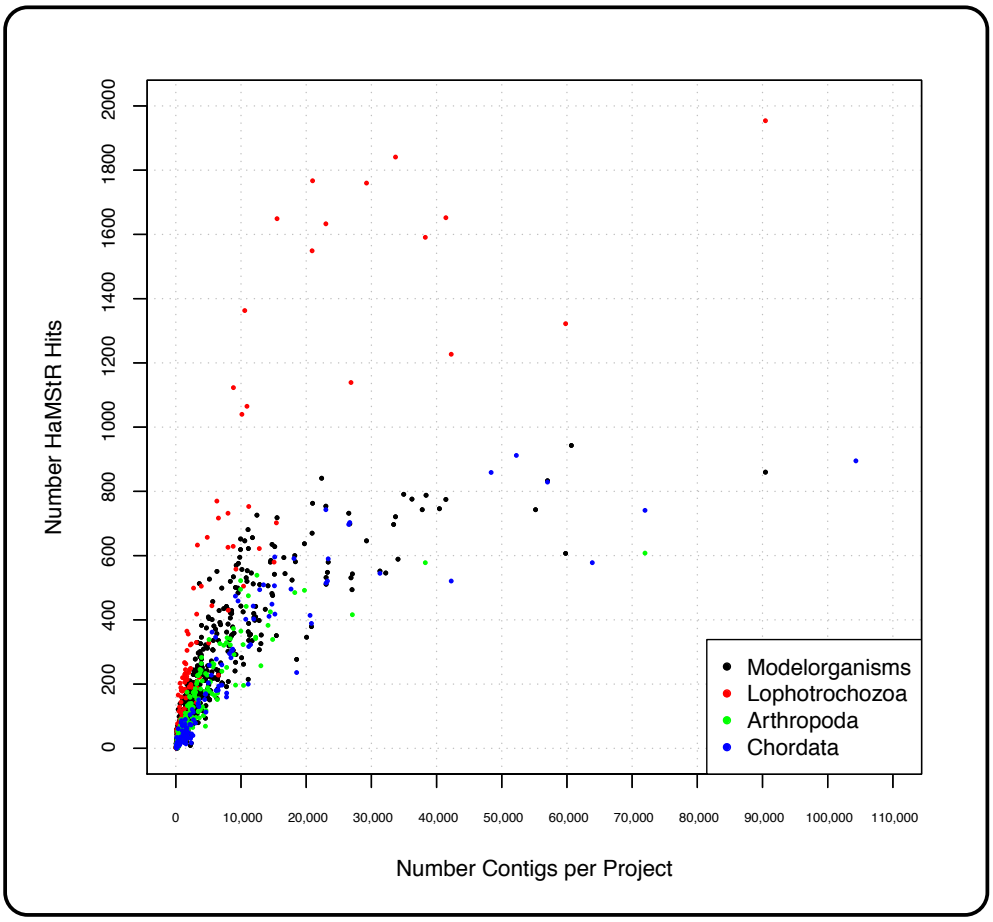


Figure 7.4: Relationship between project size and the number of detected genes Each dot represents a clustering project. The x-axis gives the number of contigs in each project, the y-axis gives the number of genes that have been found by HaMStR using the four core-ortholog sets.

core-ortholog set (Fig. 7.5).

No obvious differences between the four core-ortholog sets can be observed after normalization. The variance of the number of detected genes given a certain project size also becomes more clear. In projects with about 20,000 contigs, roughly between 25% and 80% of the genes in a core-ortholog set can be found. Furthermore, no clustering project contains sequences of all genes of a core-orthologs set. However, in very large projects ($> 30,000$ contigs), usually more than 50% of the genes are found. This should yield a sufficient overlap of sequence data for phylogeny reconstruction. Unfortunately, large projects are rare (see also Fig. 6.2). Thus, the scale of a phylogenetic analysis based

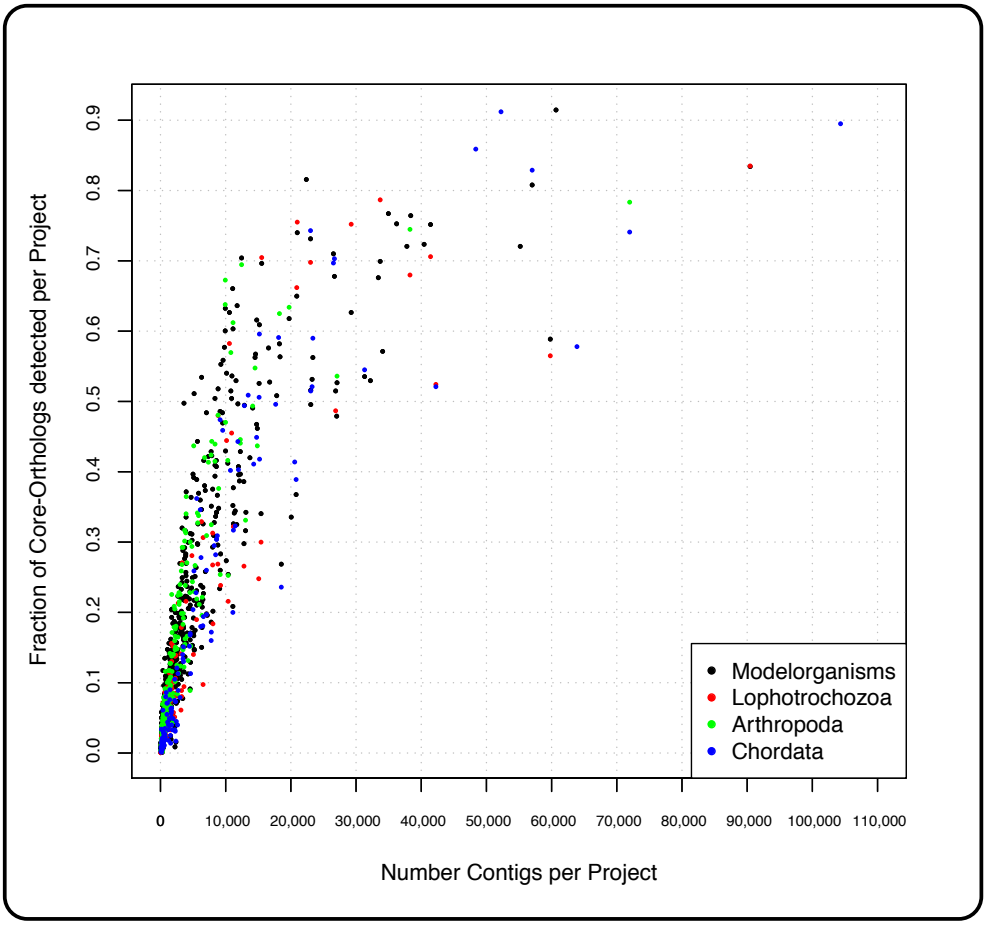


Figure 7.5: Relationship between project size and detected genes, normalized
This plot shows the same data as Fig. 7.4, but the numbers of assigned genes per project and core-ortholog set have been normalized to the interval $[0, 1]$. A value of 0 means, that none of the genes of the particular core-ortholog set has been found in the clustering project. A value of 1 means, that all genes of the core-ortholog set are represented by ESTs in that particular clustering project.

on ESTs heavily depends on how much overlapping sequence data between the smaller projects is present.
To investigate this, we re-plotted the normalized hits per clustering project, but only considered projects with not more than 2,000 contigs (Tab. 7.1). We further added regression lines to visualize the relationship between project size and found genes (Fig. 7.6). The slopes of the regression lines confirm, that the rates of detected genes explained by

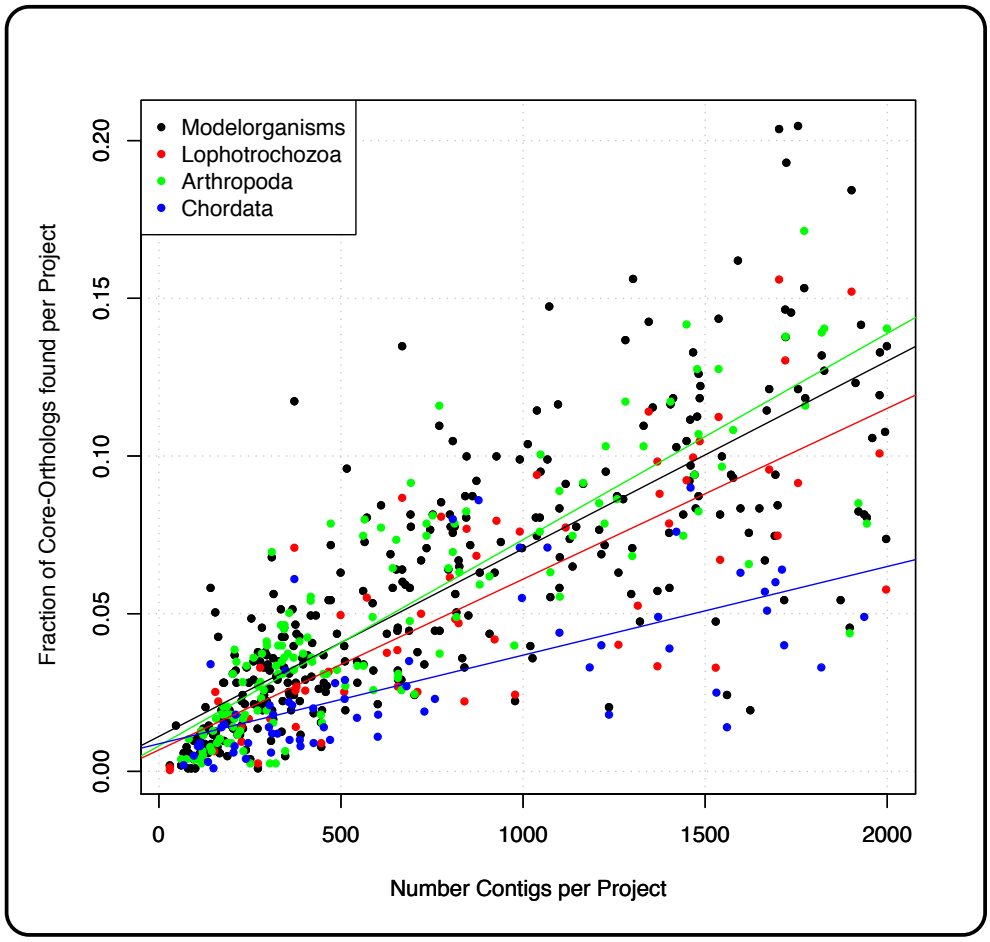


Figure 7.6: Relationship between project size and detected genes in smaller clustering projects, normalized This plot shows the ratio of project sizes and number of found genes in clustering projects with less than 2,000 contigs. The regression lines have been calculated with the least square method.

the project size differ only marginally between the individual core-ortholog sets. Only the *Chordata set* shows a lower discovery rate. Using the regression lines as estimates, between 6.5% (for the *Chordata set*) and 13.9% (for the *Arthropoda set*) of a core-ortholog set's genes can be expected to be represented in a clustering project with 2,000 contigs. If genes were sampled with equal probabilities during the EST generation process, the chances for a gene being found in two projects with 2,000 contigs each would be very low. To calculate the expected number of genes to be found in two different clustering projects of equal size, we can assume a hypergeometric distribution. Then, the probability p for k

genes detected in both projects is

$$p_k = \binom{M}{k} * \binom{N-M}{W-k} / \binom{N}{W}$$

with N being the total number of genes of a core-ortholog set, M the number of detected genes in the first project and W the number of genes sampled from the second. The expected overlap between two projects then is $W * M/N$. For example, the *Chordata set* contains 1,000 genes. According to the regression (Fig. 7.6), we can expect to find 6.5% (= 65) of these genes to be detected in a project of 2,000 contigs. The expected number of identical genes found in two different projects with a size of 2,000 contigs each would then be $(65 * 65)/1000 = 4.2$. If we add a third project of 2,000 contigs, the expected overlap is already < 1: $(4.2 * 65)/1000 = 0.273$. The compilation of a descent data set therefore seems to be very unlikely.
However, as already mentioned in section 3.4, genes differ in their expression levels. The mRNAs of highly expressed genes have an increased probability of being cloned due to their high-copy number. This in turn rises the chance that these genes will be tagged in an EST project. Although the expression level of a gene can differ substantially between different species, different developmental stages or even environmental conditions (Su *et al.* (2004)), there are genes, which are ubiquitously highly expressed (Zhang and Li (2004)). Such genes should be found in more EST projects than it would be expected by chance. Hence, these genes could form the foundation for phylogenetic data sets. We therefore counted in how many projects each gene of the four core-ortholog sets was found. To make the results comparable between the four core-ortholog sets, we calculated the discovery rate for each gene by diving the number of projects each gene was found in by the total number of projects screened with each set. The majority of the genes in each core-ortholog set is found in between 0% and 40% of all scanned clustering projects (Fig. 7.7A-D). However, each of the four sets include some genes, which show a discovery rate of more than 0.5. The *Lophotrochozoa set* even contains genes detected in more than 80% of all projects that were screened with this core-ortholog set (Fig. 7.7B). To see to what extend these genes are also found in smaller clustering projects, we restricted the analysis to clustering projects with less than 2,000 contigs (see Tab. 7.1). As expected, most of the genes are not found in any or in only a few of the smaller projects (Fig. 7.8) scanned with HaMStR. But each core-ortholog set contains at least some genes that are present in up to 70% of all considered projects (Fig. 7.8B). In order to specify these ubiquitous detected genes, we compiled a list for each core-ortholog set, containing the twenty genes found most frequently (Tables 7.2, 7.3, 7.4 and 7.5). The majority of these genes are ribosomal proteins. This observation is in line with recent studies that analyzed the composition of genes tagged by EST projects on a smaller scale (Hughes *et al.* (2006); Roeding *et al.* (2007)). The ubiquitous expression of genes encoding ribosomal proteins is not surprising, because they are involved in the protein synthesis and therewith in

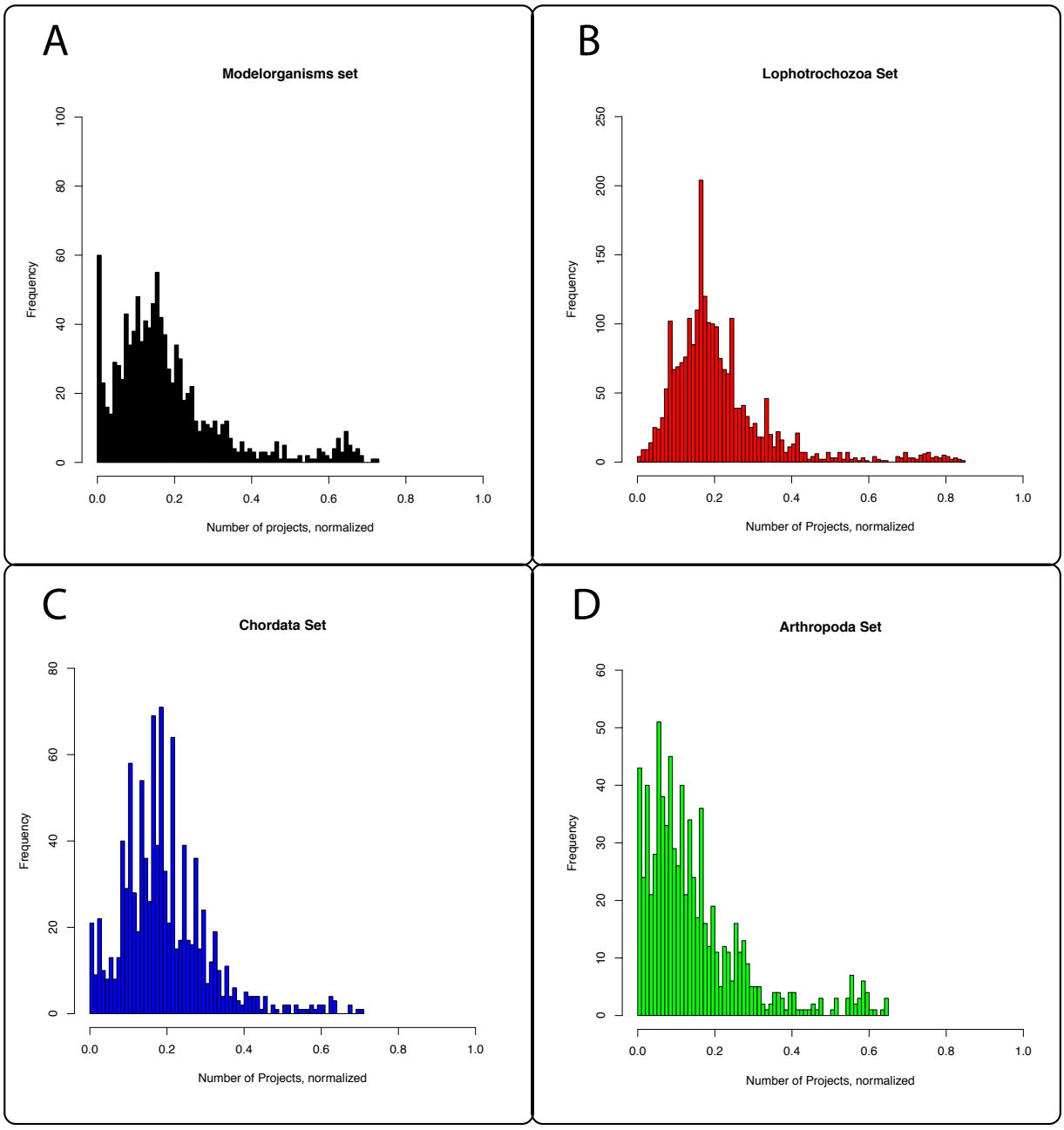


Figure 7.7: Frequencies of gene discovery The plots show how frequently each gene of the four core-ortholog sets was found in the scanned projects (Tab. 7.1, 4th column). The x-axis gives the normalized discovery rate. A value of 0 indicates, that the gene was not found, a value of 1 means that it was found in all scanned projects. The frequencies plotted on the y-axis sum up to the total size of each core-ortholog set (3rd column of Tab. 7.1).

Table 7.2: The genes most frequently found with the Modelorganism set In this table we listed the 20 genes of the Modelorganism set most frequently found in clustering projects <2,000 contigs. The first column gives the Ensembl protein ID of the human sequence that was incorporated into the core-ortholog. The second column gives the discovery rate for each gene and the last column contains the annotation, adopted from the human protein sequence.

Ensembl Protein ID	Discovery rate	Annotation
ENSP00000272317	0.58	Ubiquitin
ENSP00000349467	0.5	Calmodulin (CaM)
ENSP00000352469	0.5	Ubiquitin (outdated entry)
ENSP00000361288	0.49	40S ribosomal protein S8
ENSP00000296674	0.49	40S ribosomal protein S23
ENSP00000346050	0.49	40S ribosomal protein S3a
ENSP00000346018	0.49	60S ribosomal protein L7a
ENSP00000339027	0.48	60S acidic ribosomal protein P0
ENSP00000202773	0.47	60S ribosomal protein L6
ENSP00000339051	0.46	Translationally-controlled tumor protein
ENSP00000009589	0.46	40S ribosomal protein S20
ENSP00000369737	0.46	similar to ribosomal protein L10
ENSP00000233609	0.46	40S ribosomal protein S15
ENSP00000362744	0.46	40S ribosomal protein S4
ENSP00000368515	0.46	60S ribosomal protein L9
ENSP00000346015	0.46	60S ribosomal protein L27a
ENSP00000347049	0.46	Putative uncharacterized protein RPL17 (outdated entry)
ENSP00000359345	0.46	60S ribosomal protein L5
ENSP00000355258	0.45	60S ribosomal protein L10a
ENSP00000363676	0.45	60S ribosomal protein L11

one of the most fundamental mechanism found in organisms. These genes can form a foundation for phylogenetic data sets, which can be extended by further genes found to be present in the individual sequence data sets.

To conclude, the amount of overlapping genes is substantially higher than what would be expected by chance, which can be attributed to genes that are highly expressed in the majority of taxa under study. This allows to also include species for which only few ESTs have been generated so far. The large number of EST projects already processed by our pipeline in combination with the HaMStR approach harbor the potential to compile new data sets, which can help to solve the unanswered question of the relationships of species.

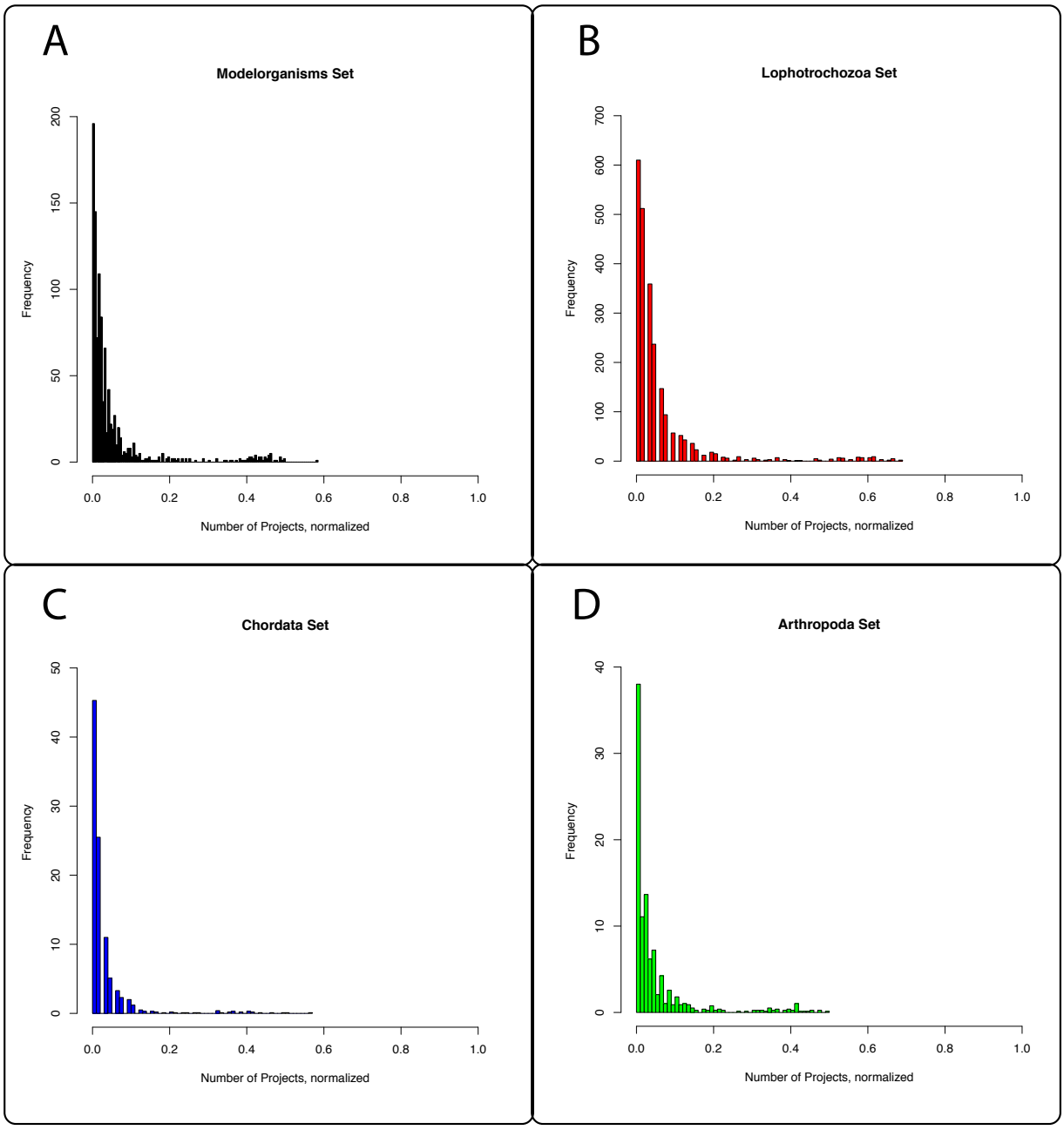


Figure 7.8: Frequencies of gene discovery in small clustering projects For these four plots, we only considered projects with less than 2,000 contigs (Tab. 7.1, 5th column). The x-axis gives the normalized discovery rate. A value of 0 indicates, that the gene was not found, a value of 1 means that it was found in all scanned projects. The frequencies plotted on the y-axis sum up to the total size of each core-ortholog set (3rd column of Tab. 7.1).

Table 7.3: The genes most frequently found with the Lophotrochozoa set In this table we listed the 20 genes of the Lophotrochozoa set most frequently found in clustering projects <2,000 contigs. The first column gives the Wormbase protein ID of the C. elegans sequence that was incorporated into the core-ortholog. The second column gives the discovery rate for each gene and the last column contains the annotation, adopted from Wormbase for the C. elegans protein sequence.

Ensembl Protein ID	Discovery rate	Annotation
WBGene00004421	0.68	large ribosomal subunit L10 protein
WBGene00004483	0.68	small ribosomal subunit S14 protein
WBGene00004417	0.67	large ribosomal subunit L6 protein
WBGene00004470	0.67	small ribosomal subunit S3A protein
WBGene00004472	0.67	small ribosomal subunit S3 protein
WBGene00004477	0.67	small ribosomal subunit S8 protein
WBGene00004496	0.67	small ribosomal subunit S27 protein
WBGene00004418	0.65	large ribosomal subunit L7 protein
WBGene00004420	0.65	large ribosomal subunit L9 protein
WBGene00004412	0.63	large ribosomal subunit L10a protein
WBGene00004481	0.63	small ribosomal subunit S12 protein
WBGene00004487	0.63	small ribosomal subunit S18 protein
WBGene00001168	0.62	elongation factor 1-alpha homolog
WBGene00004428	0.62	large ribosomal subunit L13a protein
WBGene00004429	0.62	large ribosomal subunit L17 protein
WBGene00004435	0.62	large ribosomal subunit L23 protein
WBGene00004436	0.62	large ribosomal subunit L24 protein
WBGene00004473	0.62	small ribosomal subunit S4 protein
WBGene00004478	0.62	small ribosomal subunit S9 protein
WBGene00012179	0.62	large ribosomal subunit L37e protein

Table 7.4: The genes most frequently found with the Chordata set In this table we listed the 20 genes of the Chordata set most frequently found in clustering projects <2,000 contigs. The first column gives the Ensembl protein ID of the human sequence that was incorporated into the core-ortholog. The second column gives the discovery rate for each gene and the last column contains the description, adopted from Ensembl for the human protein sequence.

Ensembl Protein ID	Discovery rate	Description
ENSP00000362744	0.57	40S ribosomal protein S4
ENSP00000346050	0.51	40S ribosomal protein S3a
ENSP00000339095	0.49	40S ribosomal protein S7
ENSP00000346001	0.46	60S ribosomal protein L3
ENSP00000346015	0.43	60S ribosomal protein L27a
ENSP00000374250	0.42	Guanine nucleotide-binding protein subunit beta-2-like 1
ENSP00000253788	0.42	60S ribosomal protein L27
ENSP00000225430	0.4	60S ribosomal protein L19
ENSP00000355258	0.4	60S ribosomal protein L10a
ENSP00000369757	0.4	40S ribosomal protein S6
ENSP00000350373	0.38	60S ribosomal protein L9
ENSP00000278572	0.38	40S ribosomal protein S3
ENSP00000222247	0.37	60S ribosomal protein L18a
ENSP00000252543	0.37	60S ribosomal protein L36
ENSP00000259469	0.37	60S ribosomal protein L35
ENSP00000318646	0.35	40S ribosomal protein S15a
ENSP00000356881	0.35	40S ribosomal protein S12
ENSP00000352402	0.34	Ribosomal protein L15
ENSP00000371103	0.32	60S ribosomal protein L37
ENSP00000363169	0.32	40S ribosomal protein S10

Table 7.5: The genes most frequently found with the Arthropoda set In this table we listed the 20 genes of the Arthropoda set most frequently found in clustering projects <2,000 contigs. The first column gives the Ensembl protein ID of the human sequence that was incorporated into the core-ortholog. The second column gives the discovery rate for each gene and the last column contains the description, adopted from the human protein sequence.

Ensembl Protein ID	Discovery rate	Description
ENSP00000361288	0.5	40S ribosomal protein S8
ENSP00000296674	0.48	40S ribosomal protein S23
ENSP00000369737	0.48	similar to ribosomal protein L10
ENSP00000301071	0.46	Tubulin alpha-1A chain
ENSP00000346037	0.46	60S acidic ribosomal protein P1
ENSP00000339027	0.45	60S acidic ribosomal protein P0
ENSP00000279259	0.44	Ubiquitin-like protein FUBI
ENSP00000196551	0.42	40S ribosomal protein S5
ENSP00000202773	0.42	60S ribosomal protein L6
ENSP00000346015	0.42	60S ribosomal protein L27a
ENSP00000233609	0.42	40S ribosomal protein S15
ENSP00000362744	0.42	40S ribosomal protein S4, X isoform
ENSP00000225430	0.42	60S ribosomal protein L19
ENSP00000222247	0.41	60S ribosomal protein L18a
ENSP00000344777	0.41	Putative uncharacterized protein (outdated entry)
ENSP00000253788	0.41	60S ribosomal protein L27
ENSP00000270634	0.4	60S ribosomal protein L13a
ENSP00000009589	0.4	40S ribosomal protein S20
ENSP00000262584	0.39	60S ribosomal protein L8
ENSP00000346050	0.39	40S ribosomal protein S3a

8 Application of EST-based Phylogenetics: Pterygota

In this chapter we will demonstrate the application of HaMStR and the data of dbDMP by addressing the evolution of winged insects —Pterygota. This study was published in Simon *et al.* (2009).

8.1 Background

Insects are the most diverse animal group on earth and dominate every ecosystem except the benthic zone (Grimaldi and Engel (2005)). The winged insects account for more than 98% of the class Insecta (Grimaldi and Engel (2005)). According to fossil records, flying insects originated in the Early Carboniferous period (approx. 320 MYA), whereas a DNA-based study suggested an origin in the mid-Devonian (approx. 387 MYA) (Gaunt and Miles (2002)). A recent analysis of Engel and Grimaldi (2004) suggested that the origin of insect wings occurred coincident with the development of arborescence, and agreed with the molecular estimates of Gaunt and Miles (2002). With the invention of the wings, insects were able to invade every ecosystem, escape predators, and exploit scattered resources, resulting in rapid radiations into vast numbers of species (Hennig (1969)). Considering the tremendous impact this change produced, the evolution of the flying insects is one of the most fascinating questions in evolutionary biology. Martynov (1925) was the first to distinguish two groups of winged insects based on wing function —Palaeoptera and Neoptera. He assumed the inability to fold back the wings, as seen in Ephemeroptera and Odonata, to be an ancestral condition and therefore called them Palaeoptera (old wings) in contrast to those with this ability, which he called Neoptera (new wings). The monophyly of Palaeoptera has been controversial ever since. In contrast to the accepted monophyly of Neoptera, the so-called "Palaeoptera Problem" is one of the unsolved mysteries in insect systematics.
Today, three hypotheses are proposed to explain the phylogenetic relationships of the basal winged insects: *(i)* the Palaeoptera scenario that supports a basal sister group position of Odonata and Ephemeroptera (Odonata+Ephemeroptera, Neoptera), *(ii)* the Metapterygota scenario (Ephemeroptera basal, Odonata+Neoptera), and *(iii)* the

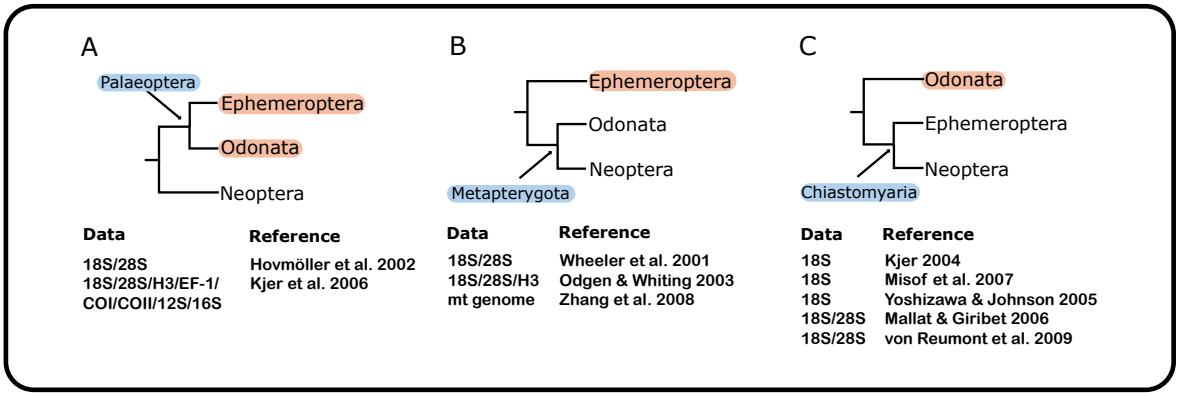


Figure 8.1: The three hypotheses at the base of the pterygotes (a) Palaeoptera (Ephemeroptera+Odonata, Neoptera), (b) Metapterygota (Ephemeroptera, Odonata+Neoptera), (c) Chiastomyaria (Odonata, Ephemeroptera+Neoptera). The sister group relationships are indicated in blue and the resulting basal pterygote order in red. Below are different molecular studies listed supporting one of the three hypotheses partly using the same genes.

Chiastomyaria scenario (Odonata basal, Ephemeroptera+Neoptera) (Whitfield and Kjer (2008)) (8.1). Each hypothesis is still considered viable and supported by morphological as well as molecular data. Moreover, some molecular data using the same genes support all three hypotheses depending on the analyses applied (e.g. Hovmöller *et al.* (2002); Ogden and Whiting (2003); Mallatt and Giribet (2006)).

The Palaeoptera are a morphologically well-supported group due to the fact that the Odonata and Ephemeroptera are unable to flex their wings back over the abdomen whereas members of the Neoptera harbor the necessary muscles and wing sclerites for this movement (Kukalová-Peck and Lawrence (2004)). Historically, the wing flexing mechanism (without backward folding) and the similar wing base sclerites seen in the Palaeoptera, were considered as an ancestral condition (e.g. Martynov (1925); Hennig (1969); Kukalova-Peck (1991)). Furthermore, the anal brace, the intercalary veins, and aquatic larvae are interpreted as plesiomorphic characters of the Ephemeroptera and Odonata (Kukalova-Peck (1991); Staniczek (2000); Bechly *et al.* (2001)). In contrast, the suppression of imaginal molts, the absence of the axillar-furcal muscle, the basalar-sternal muscles, and the missing terminal filum observed in the Odonata and Neoptera are possible synapomorphies supporting the Metapterygota scenario (e.g. Kristensen *et al.* (1991); Beutel and Gorb (2001); Grimaldi and Engel (2005); Willkommen and Hörnschemeyer (2007)). Alternately, the direct sperm transfer shared by the Ephemeroptera and Neoptera in contrast to the indirect sperm transfer in Odonata support the Chiastomyaria theory (Boudreaux (1979)). Moreover, the wing base structure of the Odonata and the remaining

pterygote orders show significant differences in appearance and function, for example wing flapping in Odonata is promoted by the direct flight muscles whereas in Ephemeroptera and Neoptera, it is promoted by indirect flight muscles (Ninomiya and Yoshizawa (2009)). The difficulties in establishing homology of the wing base structure between the Odonata and other Pterygota resulted in an extreme interpretation of Matsuda (1970); Matsuda (1981) and Greca (1980). They concluded that the wing base structure in odonates is so different that it cannot be homologized with that of Ephemeroptera and Neoptera. However, the monophyly of Pterygota is now well established through both morphology and molecular data (e.g. Kristensen *et al.* (1991); Wheeler *et al.* (2001); Grimaldi and Engel (2005); Kjer *et al.* (2006); von Reumont *et al.* (2009)). Recently Ninomiya and Yoshizawa (2009) established the homology of the wing base structures between the Odonata, Ephemeroptera and Neoptera. Based on wing base morphology, they almost unambiguously determined that there is a single origin of insect wings and flight, but they were not able to contribute further on the basal diversification of Pterygota.
Establishing a sound phylogenetic hypothesis for the origin of insect wings based on wing base structure and the wing folding mechanism remains crucial.

But why is the so-called Palaeoptera Problem not resolved despite the advances in molecular systematics? Whitfield and Kjer (2008) pointed out that the "ancient rapid radiation" is a major contributing factor in the inability to resolve insect relationships with molecular data. Due to short ancient internodes, connecting the taxonomic groups, inadequate molecular data sets, conflicting results within or among data sets, and an overall weak phylogenetic signal is observed in many pterygote phylogenetic studies (Wheeler *et al.* (2001); Ogden and Whiting (2003); Kjer *et al.* (2006); Misof *et al.* (2007); von Reumont *et al.* (2009)). In addition, one major challenge is to find useful molecular markers to accurately track these short ancient internodes. For the reconstruction of an "accurate" phylogeny, molecular marker systems are required which have kept pace with speciation but slow enough to have transferred the phylogenetic signal to the present (Regier and Shultz (1998)). Unfortunately, the rationale behind the selection of certain molecular markers is not always clear, and discrepancies and incongruence between individual gene trees may result in unresolved phylogenetic trees (Wheeler *et al.* (2001); Kjer *et al.* (2006)). Thus, phylogenetic analyses of single genes and even multiple marker systems have not yet conclusively resolved the basal pterygote diversification. It is therefore conceivable that resolution of these relationships may require not only large amounts of sequence data but also an assessment of data quality and quantity. Several studies have shown that analyzing a large number of genes simultaneously helps to infer unresolved issues in deep metazoan relationships (e.g. Philippe *et al.* (2005b); Savard *et al.* (2006); Roeding *et al.* (2007); Dunn *et al.* (2008)). Moreover, simulations and studies based on real data have shown that trees based on concatenated alignments provide better resolution for a particular topology than consensus gene trees —known

as "supertree" approaches (Rokas *et al.* (2003a); Gadagkar *et al.* (2005); Savard *et al.* (2006)). However, there is still a controversy about phylogenetic reconstructions derived from supertree versus "supermatrix" approaches (e.g. Gatesy *et al.* (2004); Wilkinson *et al.* (2007)). Both methods have demonstrated strengths and weaknesses, and some promising new approaches are addressing the existing problems. For example, for the supertree method, the recent proposal of a maximum likelihood (ML) approach forms an important idea for future phylogenetic inferences from genomic data (Steel and Rodrigo (2008); Cotton and Wilkinson (2009)). Also the implementation of new methods, for example Bayesian estimation of species trees (Edwards *et al.* (2007); Liu and Pearl (2007)) to simultaneously estimate gene trees and species trees from multilocus data using a coalescent framework has been shown to be very efficient in cases of recent speciation (Edwards *et al.* (2007); Belfiore *et al.* (2008); Wiens *et al.* (2008)). All these phylogenomic approaches have one problem in common; although the stochastic error is dramatically reduced by using a large number of data, they are not protected against systematic errors (Phillips *et al.* (2004); Delsuc *et al.* (2005)). Furthermore, systematic bias can be reinforced by increasing the number of characters resulting in a highly supported but incorrect tree (Felsenstein (1978); Jeffroy *et al.* (2006)). Long-branch attraction (LBA) coupled with taxon sampling, phylogenetic reconstruction methods and base composition bias are all factors that are known to cause systematic errors and to be potential pitfalls when attempting to recover "the true evolutionary history of species" (Zwickl and Hillis (2002); Phillips *et al.* (2004); Brinkmann *et al.* (2005); Delsuc *et al.* (2005); Philippe *et al.* (2005a)).

With the aim of addressing the origin of flying insects, we generated and analyzed expressed sequence tag (EST) data from the two basal orders of winged insects —from a mayfly (Ephemeroptera, *Baetis sp.*) and a damselfly (Odonata, *Ischnura elegans*). EST data provide a comprehensive random sample of protein-coding genes and an economic way to produce a large number of sequences for phylogenetic analysis of 'nonmodel' species, for which genome sequence projects are not yet available.

Although EST data collection is increasing due to the tremendous recent advances in sequencing technologies and as an optimal source for multigene approaches, ESTs from representatives of the basal winged insect orders are still scarce. Although ESTs are a promising tool to resolve deep phylogenetic questions, there are still necessary precautions to take when handling EST data sets. The complex nature of genome evolution including gene loss, duplications, expansion of gene families and functional diversification consequently requires assignment of gene orthology when using ESTs as a source for phylogenetic analyses (Hughes *et al.* (2006)). Furthermore ESTs represent a snapshot of gene expression within a given set of tissue, developmental stages and environmental conditions (Rudd (2003)), and the overlap of genes in the taxa may be very limited (Hughes *et al.* (2006)).

8.2 Compilation of the Data

8.2.1 Generation of Sequence Data

Specimens were stored in RNAlater (Qiagen) at $-80°C$ before RNA extraction. Total RNA of *Baetis sp.* was extracted four times from two larval specimens simultaneously using Qiagen RNeasy kits and pooled afterward. Total RNA of *I. elegans* was extracted from one adult specimen using Qiagen RNeasy kits. The two RNA samples were precipitated with 0.1Vol NaAC in diethylpyrocarbonate and 2.5Vol 100% ethanol for later construction of cDNA libraries. The Creator SMART cDNA Library Construction Kit (Clontech) and the Trimmer Kit (Evrogen) were used for the construction of the normalized cDNA libraries following the manufacturers' instructions. Modifications to the protocol were made concerning the cloning vector: pal32 (Evrogen) was used for directional cloning with insertion between two SfiI sites.
Plasmids were transferred via electroporation to Escherichia coli (strain DH10B, Invitrogen). Plasmids were isolated using the method of Hecht *et al.* (2006) and 5' end sequenced using BigDye V3 (ABI) and 3730XL capillary sequencer systems (ABI). By that, we obtained 4,225 *Baetis sp.* and 4,219 *Ischnura elegans* ESTs, which were subsequently processed and annotated with the pipeline described in Chapter 4.
After cleaning, 4,197 *Baetis sp.* of the initial 4,225 clones were left. For *Ischnura elegans* we obtained 4,217 from 4,219 cleaned ESTs. The clustering resulted in 3,035 contigs (635 contigs contain more than one EST, 2,400 singletons) for *Baetis sp.* and 3,194 (614 contigs contain more than one EST, 2,580 singletons) for *Ischnura elegans*.

The cleaned *Baetis sp.* ESTs have been deposited in the EMBL Nucleotide Sequence Database with Accession Nubers FN198828-FN203024 and *Ischnura elegans* ESTs with Accession Nos FN215340-FN219556.

8.2.2 Orthology Assignment

In order to compile a data set suited to address the phylogeny of Pterygota, we searched for orthologs with the HaMStR approach (see Section 7.2). To this end, we created a custom core-ortholog set, as depicted in Section 7.2.1.1. As primer-taxa we used *Anopheles gambiae*, *Apis melifera*, *Drosophila melanogaster*, *Homo sapiens* and *Aedes aegypti*. The pairwise orthologs were extracted from the InParanoid database version 6 (Berglund *et al.* (2007)). A successful transitive closure was achieved for 3,096 genes, which formed a core-ortholog set, referred to as *Insecta set* hereafter. The re-BLAST of the candidate

EST contigs (see Section 7.2.2.2) was performed against *Apis melifera* for all clustering projects.
Of the 3,096 core-orthologs, we could detect 436 in *Baetis sp.* and 527 in *Ischnura elegans*.

8.2.3 Extension of the Data Set with Public ESTs

We complemented our data set with clustering projects based on public ESTs for 25 pterygote and three apterygote species (Table C.1). Each project was screened for orthologs with HaMStR and the Insecta core-ortholog set.
Because not all of 3,096 genes were present in the EST contigs of each taxon (Tab. C.1 in the appendix), a concatenation of all gene alignments would have resulted in a substantial amount of missing data. We therefore used a PERL script (Ebersberger, unpublished) that automatically analyzes the amount of missing data for different combinations of taxa and genes. As selection criterion for the data sets, we imposed that *Baetis sp.*, *I. elegans*, and at least one apterygote taxon were present in each set. One data set (named *maxspe*) comprised 15 species and 125 genes with 18% missing data and a second (named *maxgen*) comprised 8 species, 150 genes, and 11% missing data (see Table C.2 in the appendix for a list of represented genes and the overlap between both sets). We decided to perform all analyses with both data sets to make our results more robust.

Sequences of both sets were aligned with MAFFT (Katoh *et al.* (2005)) using the options `--maxiterate 1000` and `--localpair`. Afterwards, we concatenated the alignments to generate one super alignment per data set. The *maxspe* set yielded an alignment length of 31,643aa. The *maxgen* alignment had a sequence length of 42,541aa.

8.3 Analyses

8.3.1 Phylogenetic Analyses of the Concatenated Data

Both alignments (*maxspe*, *maxgen*) were checked for putative randomly similar sections using ALISCORE (Misof and Misof (2009)). We applied a sliding-window size (w=6) with the BLOSUM62 matrix and function -e (option for EST data with lots of missing data section).
After the exclusion of putative randomly similar aligned sections, the data set *maxspe* comprised 26,152aa (initial 31,643aa, ∼18% randomly similar) and *maxgen* comprised 37,473aa (initial 42,541, ∼12% randomly similar). The final alignments have been deposited at TREEBASE[1] (study accession no. S2456).

[1] http://www.treebase.org

We then determined the best fitting model of protein sequence evolution with ProtTest 1.4 (Abascal *et al.* (2005)). The WAG (Whelan and Goldman (2001)) model of amino acid sequence evolution and a γ-model of rate heterogeneity (Gu *et al.* (1995)), with four classes of variable sites and one class of invariable sites (4Γ+I) was used in all subsequent phylogenetic analyses.
Maxspe and *maxgen* were treated equally in all following steps of phylogenetic and statistical analyses. Tests of the three alternative phylogenetic hypotheses at the base of the Pterygota were accomplished by using the approximately unbiased (AU) test, Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), weighted Kishino-Hasegawa (WKH) and weighted Shimodaira-Hasegawa (WSH) tests as implemented in CONSEL (Shimodaira and Hasegawa (2001)). First, alternative tree topologies were reconstructed by using GARLI 0.96b8 (Zwickl (2006)) under default parameters. Then, PAUP* (Swofford (2002)) was used to produce a file with the site wise log-likelihoods of alternative trees. The resulting files were summarized to a single file that served as input for CONSEL to calculate the p-value for each alternative phylogenetic hypothesis.
As the monophyly of the major groups was not disputed, we put a topological constraint according to the three phylogenetic hypotheses on the tree search to identify the highest likelihood topologies that satisfied a given hypothesis. In addition we constrained the monophyly of Paraneoptera and Holometabola in the *maxspe* data set and the monophyly of Holometabola in the *maxgen* data set (e.g. Hennig (1981); Yoshizawa and Saigusa (2001); Kaestner (2003); Beutel and Pohl (2006)).
Results of the hypotheses testing using heuristic search and incorporating topology constraints are summarized in Table 8.1. Based on the constrained analyses, the Chiastomyaria scenario (Odonata, Ephemeroptera+Neoptera) is significantly supported by all tests (AU, KH, SH, WKH and WSH) in the *maxspe* data sets while the *maxgen* alignment could not significantly reject the Metapterygota theory in the weighted SH test (WSH=0.062) using the 95 percent significance level.

In addition to the constrained analyses, searches in the absence of topological constraints were carried out. For this purpose, maximum likelihood analyses (ML) were performed with the Pthreads-parallelized version of RAxML 7.0.4 (Stamatakis (2006)) under a rapid bootstrap analysis (-f a) and the PROTMIXWAG model. The branching support was assessed by 1,000 bootstrap replicates. Bayesian inference (BI) analyses were performed using a compiled parallel version of MrBayes v3.1.2 (Huelsenbeck and Ronquist (2001); Ronquist and Huelsenbeck (2003); Altekar *et al.* (2004)) with two parallel runs under the WAG+4Γ+I model. Metropolis-coupled Markov chain Monte Carlo (MCMCMC) sampling was carried out with one cold and three heated chains starting from random starting trees and the program default prior probabilities on model parameters. The *maxspe* data were run for 3,000,000 generations (average standard deviation (SD) of split frequencies < 0.0078), and the *maxgen* data were run for 1,000,000 generations

Table 8.1: Statistical Confidence (P-Values) for alternative relationships at the base of the pterygotes AU: approximately unbiased test, KH: Kishino-Hasegawa test, SH: Shimodaira-Hasegawa test, WKH: weighted Kishino-Hasegawa test and WSH: weighted Shimodaira-Hasegawa test.

		P values				
Data set	Hypothesis	AU	KH	SH	WKH	WSH
maxgen	Palaeoptera	2e-04**	0.001*	0.001*	0.001*	0.001*
	Metapterygota	0.032*	0.035*	0.039*	0.035*	0.062
	Chiastomyaria	0.971	0.965	0.987	0.965	0.985
maxspe	Palaeoptera	7e-50***	0.002*	0.002*	0.001*	0.001*
	Metapterygota	0.029*	0.025*	0.025*	0.025*	0.048*
	Chiastomyaria	0.971	0.75	0.982	0.975	0.979

(average SD of split frequencies < 0.0000). For both data sets, samples of the Markov chain were taken every 100 generations giving a total sample of 30,000 trees (*maxspe*) or 10,000 trees (*maxgen*). Parameters were checked for stationarity with Tracer v1.4 (Rambaut and Drummond (2007)) and the first 10,000 trees were discarded as burn-in. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled after the initial burn-in period.

The reconstructed phylogenetic trees using both alignments (*maxspe*, *maxgen*) and both methods (ML, Bayesian) are shown in Figures 8.2 and 8.3, respectively. In all trees *Ischnura elegans* (Odonata) represent —with high bootstrap support/posterior probability (*maxspe*: 100%/100%; *maxgen*: 100%/100%)— the most basal winged insect specimens, supporting the Chiastomyaria theory. The topology generated from the *maxspe* alignment further supports the monophyly of Paraneoptera (*Aphis gossypii*, *Maconellicoccus hirsutus*) (98%/100%) and Holometabola (100%/100%), with a basal position of Hymenoptera within the Holometabola data set (Fig. 8.2). The relationships within the Lepidoptera were not well supported in the ML ($62 - 31\%$) and the BI ($32 - 28\%$) analyses based on the *maxspe* data set.
However, the tree based on the *maxgen* alignment is a true subtree of the *maxspe* tree. This indicates that the results are robust with respect to the number of species and genes. To further evaluate the quality of fit for the chosen model of evolution, we performed the test developed by Goldman (1993). The results (see Figures C.1 and C.2 in the appendix) support that the WAG model describes the data adequately.

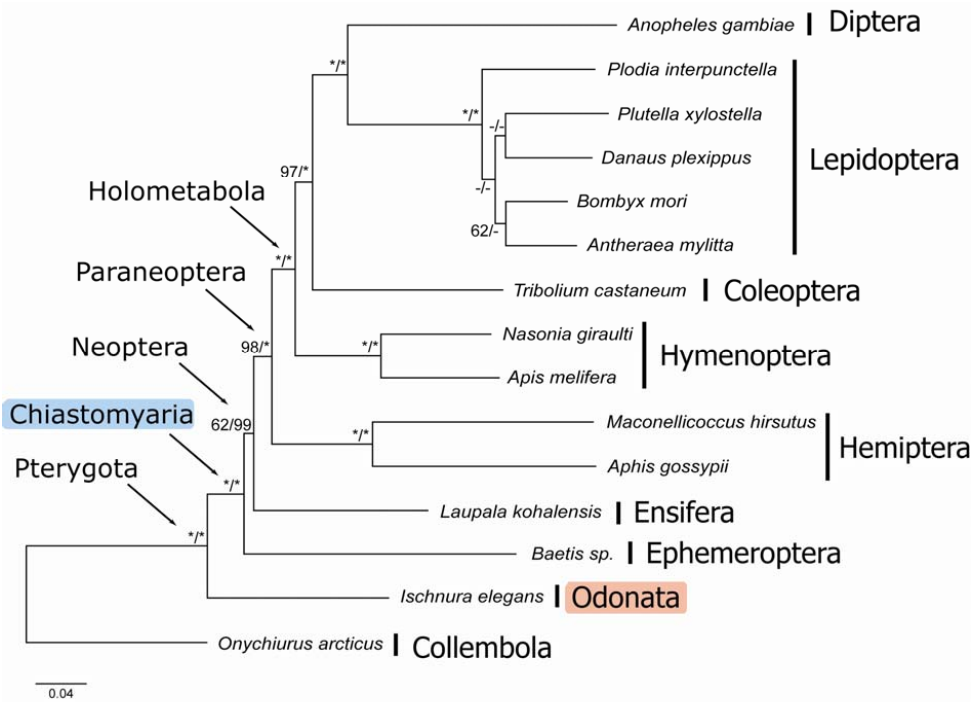


Figure 8.2: Maximum likelihood + Bayesian inference topology of *maxspe* Pterygote phylogenetic relationships based on 15 taxa and 125 genes data set (*maxspe*) showing a basal position of Odonata (*Ischnura elegans*), the monophyly of Paraneoptera and Holometabola. Branch lengths are from maximum likelihood trees. Bootstrap support values of maximum likelihood analysis and Bayesian posterior probabilities for each branch are indicated before and after a slash, respectively. Asterisk indicates 100% support value, hyphen indicates support value below 50%.

8.3.2 Phylogenetic Analyses of Single Alignments

Both data sets were scanned for individual genes represented in *Baetis sp.*, *Ischnura elegans*, and Onychiurus arcticus, as well as in at least one neopterous insect. In the *maxspe* alignment, we identified 39 genes and in the *maxgen* alignment 58 genes. Of these, 34 genes are present in both alignments. The function of these 63 genes was assessed through BLAST against the KOG (Eukaryotic Orthologous Groups) database[2] and assigned to the four major KOG categories: (1) cellular processes and signaling, (2) information storage and processing, (3) metabolism, and (4) poorly characterized (8.3).

We performed extended ML tree analyses of the individual *maxspe* (total 39) and *maxgen* (total 58) alignments to investigate the support of the three phylogenetic

[2]http://biotec.icb.ufmg.br/K-EST/begin.html

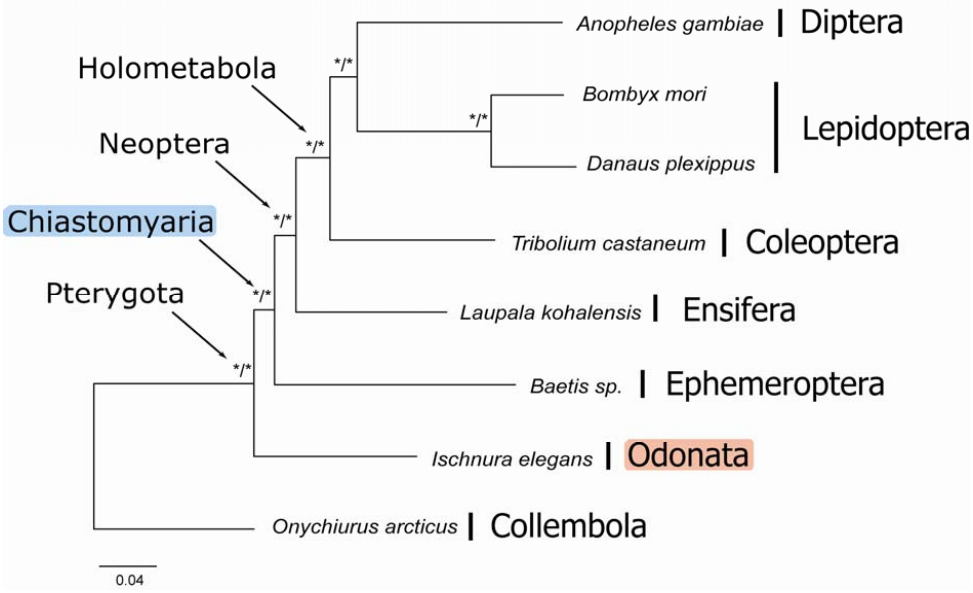


Figure 8.3: Maximum likelihood + Bayesian inference topology of maxgen Pterygote phylogenetic relationships based on 8 taxa and 150 genes data set (*maxgen*) showing a basal position of Odonata (*Ischnura elegans*). Branch lengths are from maximum likelihood. Bootstrap support values of maximum likelihood analysis and Bayesian posterior probabilities for each branch are indicated before and after a slash, respectively. Asterisk indicates 100% support value, hyphen indicates support value below 50%.

hypotheses by the individual genes. The log likelihood for each topology was calculated using TREE-PUZZLE 5.2 (Schmidt *et al.* (2002)). The topologies were considered as supported by the individual gene alignments if the P-SH < 0.05 and if the ΔlogL : S.E. ratio exceeded 0.5 (Table C.3 in the appendix). In addition, for each gene alignment of the *maxspe* (Table C.4(a)) and *maxgen* set (Table C.4(b)), that included a sequence of *Baetis sp.*, *Ischnura elegans*, *Onychiurus arcticus*, and at least one neopterous insect, an ML tree with 100 bootstrap replicates was calculated using RAxML. Within *maxgen*, based on the P-SH value and the $\Delta logL$: S.E. ratio, two genes (lethal (2) tumorous imaginal discs and Helicase at 25E) support the Metapterygota hypothesis and the gene Cysteine proteinase Cathepsin L (K-EST description) supports the Chiastomyaria hypothesis. The majority of the genes (55) represented in the *maxgen* set did not carry sufficient phylogenetic signal to distinguish between the three alternative topologies (Table C.3 in the appendix). In addition, the bootstrap analyses for each gene alignment did not provide significant support ($> 95\%$) for a single phylogenetic hypothesis (Table C.4(a)). To increase phylogenetic signal, the genes of the *maxgen* data set were concatenated according to their KOG category and subjected to ML tree analyses using the same

Table 8.2: Maximum likelihood support for the three different phylogenetic hypotheses of the concatenated alignments based on their KOG category. The favored topology of each KOG category is indicated in bold. The support is expressed as the ΔlogL : S.E. and the P-SH value. The $-\log L$ value of the best tree is written in square brackets.

		Cellular Processes and Signaling		**Information Storage and Processing**		**Metabolism**		**Poorly characterized**	
Data set	Hypothesis	ΔlogL : S.E.	P-SH	ΔlogL : S.E.	P-SH	ΔlogL : S.E.	P-SH	ΔlogL : S.E.	P-SH
maxgen	Palaeoptera	6.61	< 0.0000***	2.13	0.0570	0.87	0.204	1.06	0.204
	Metapterygota	6.64	< 0.0000***	**[31382.03]**	**1.00**	2.05	0.029	**[19394.10]**	**0.029**
	Chiastomyaria	**[47438.79]**	**1.00**	0.07	0.5860	**[24472.80]**	**1.00**	1.6	1.00
maxspe	Palaeoptera	6.79	< 0.0000***	1.64	0.0910	0.74	0.265	**[3228.31]**	**0.265**
	Metapterygota	6.77	< 0.0000***	0.37	0.4810	1.41	0.116	1.04	0.116
	Chiastomyaria	**[46215.35]**	**1.00**	**[43602.25]**	**1.00**	**[22195.41]**	**1.00**	1.00	1.00

methods as in the individual gene analyses. Table 8.2 summarizes the support for the three phylogenetic hypotheses as recoded for analyses based on the functional classification using the statistical methods. The proteins involved in cellular processes and signaling (concatenated = 5,285aa) gave the strongest support for the Chiastomyaria hypothesis and rejected significantly both other topologies. The proteins contained in the metabolism category (concatenated = 3,143aa) also favor the Chiastomyaria hypothesis but did not significantly reject the Metapterygota hypothesis. Proteins classified as information storage and processing proteins (concatenated = 4,697aa) favor the Metapterygota hypothesis but did not reject the Chiastomyaria hypothesis. The poorly characterized proteins (1,179aa) identified the Metapterygota topology as the best but again did not reject the remaining hypothesis.

None of the individual *maxspe* alignments, which were also subjected to extended ML tree analysis using TREE-PUZZLE and RAXML, provide significant support for one of the phylogenetic hypotheses (see Table C.3 and C.4(a) in the appendix). To increase the phylogenetic signal we also concatenated the individual *maxspe* alignments based on their KOG category assignment (cellular processes and signaling (3,511aa), information storage and processing (4,245aa), metabolism (1,551aa) and poorly characterized (329aa)). Three of the four KOG category derived *maxspe* alignments identified the Chiastomyaria phylogeny as the best ML tree, but the two alternative topologies could not be rejected by the proteins involved in information storage and processing + metabolism, whereas the genes involved in cellular processes and signaling significantly support the Chiastomyaria theory. Proteins categorized as poorly characterized identified the Palaeoptera topology as the best tree but not significantly (summarized as Table 8.2).

8.4 Discussion

The question of the first winged insect order has been dominated by the analyses of morphological characters and nuclear rRNA data (18S and 28S). Recently, Zhang *et al.* (2008) published the first mitochondrial genome of an Ephemeropteran. The analysis used the mitogenomic approach and supported the Metapterygota hypothesis. Despite numerous studies concerning the phylogenetic relationships at the base of pterygotes, the so-called Palaeoptera problem is still not solved and results are often conflicting. A combined analysis of nuclear rRNA (18S and 28S) and 275 morphological characters supported the Metapterygota hypothesis (Wheeler *et al.* (2001)) as did a combined analysis of 18S+28S rRNA, the protein-coding gene Histone 3, and morphology data (Ogden and Whiting (2003)). This hypothesis is supported by some diagnostic morphological characters connecting Ephemeroptera with the apterygote hexapods, such as molting, muscle structure in the tracheal system, and the caudal filament (Kristensen *et al.* (1991)).
However, different analyses of nuclear rRNA data by different authors support each of the three phylogenetic hypotheses depending on the phylogenetic inference method used for example combined 18S and 28S supports the Metapterygota hypothesis (Wheeler *et al.* (2001); Ogden and Whiting (2003)), the Palaeoptera hypothesis (Hovmöller *et al.* (2002)) and the Chiastomyaria hypothesis (Mallatt and Giribet (2006); von Reumont *et al.* (2009)). The longest standing hypothesis and the traditional textbook scenario based on morphological characters is the Palaeoptera hypothesis. It is supported by the inability of the Ephemeroptera and Odonata to fold their wings over the abdomen (Hennig (1969); Kukalova-Peck (1991)), the intercalary veins in the wings, the fusion of the galea and lacinia in the larval maxillae, and the aquatic larvae (Hennig (1981)). Kjer *et al.* (2006) also supported this hypothesis using 9 genes and 170 morphological characters. However, a strong argument for the third hypothesis —the Chiastomyaria hypothesis— is the indirect sperm transfer mechanism linking the Odonata to the apterygote insects (Boudreaux (1979)) and the direct flight muscles that are a unique character of Odonata. This phylogenetic hypothesis is further supported by several molecular studies (Kjer (2004); Yoshizawa and Johnson (2005); Misof *et al.* (2007)).
All studies clearly illustrate that basal pterygote divergence is difficult to unveil, despite the use of various morphological characters and molecular markers. One major problem is certainly the fast evolution of the pterygotes and the enormous diversity within this group. Furthermore, the preserved ancient characters in some taxa and the rate heterogeneity among orders lead to confusion among phylogeneticists. For example, Kjer *et al.* (2006) observed excessive substitution rate acceleration for Diptera and Diplura, whereas Odonata and Mantodea seem to almost ”stand still”. Finding appropriate molecular markers with phylogenetically informative sites tracking the narrow window, within which the divergence and origin of winged insects took place, is the major challenge. In this

study we included two crucial new basal winged insect EST data sets (representing the Odonata and Ephemeroptera), adopted a multigene approach, and evaluated the support of different classes of functional protein coding-genes for each of the three hypotheses.
Protein-coding sequences obtained by EST sequencing represent a valuable and relatively inexpensive possibility for resolving long outstanding deep phylogenetic relationships. The conserved nature of the housekeeping genes makes studies of divergences that took place millions of years ago possible. Thus, phylogenetic trees inferred from multi-gene approaches using ESTs have become a popular method to resolve long outstanding questions in deep metazoan relationships. Dunn *et al.* (2008) for example, improve the resolution of the animal tree of life using a concatenated alignment of 150 genes, Philippe *et al.* (2004) concatenated 129 orthologous proteins for eukaryotic species, and Savard *et al.* (2006) assembled 185 genes to resolve the radiation of Holometabolous insects. The advantages of a multigene approach instead of a single gene or few genes are numerous. Rokas *et al.* (2003b) pointed out that the biological process of a gene as influenced by natural selection or genetic drift may cause the history of the genes under analysis to obscure the history of the taxa. Issues such as gene duplication and lineage sorting may contribute to varying degrees of discordance between gene tree and species tree. Therefore, conflicting topologies are often seen in analyses of a single or small numbers of concatenated genes. Furthermore, the use of one or a few genes is known to be insufficient for the resolution of many clades (Bapteste *et al.* (2002); Rokas *et al.* (2003a); Rokas *et al.* (2003b)), whereas larger amounts of data and the increasing number of phylogenetic informative positions robustly resolve the topology (Philippe *et al.* (2004)).
However, is a multi-gene approach really a panacea for the accurate resolution of a species tree? A study by Gadagkar *et al.* (2005) indicates that this may not be the case, by showing that weak phylogenetic signals can be substantially reinforced when sequences are concatenated, but in the worst case it can also enhance support for the erroneous inferences, leading to very high bootstrap support for incorrect clades. In other words, the multi-gene approach does not necessarily lead to the correct topology because adding of new genes does not increase the accuracy of the topology in the presence of a bias. Various studies have shown that the consistency of tree reconstruction in phylogenomic studies is sensitive to the model of sequence evolution (Phillips *et al.* (2004); Jeffroy *et al.* (2006)) and to taxon sampling (Hillis *et al.* (2003); Brinkmann *et al.* (2005)), both potential sources of LBA artifacts. Subsequently, the detection and avoidance of LBA artifacts remain the most important challenge for phylogenomic studies. One strategy to reduce the impact of systematic bias would be to apply probabilistic methods that take into account variable evolutionary rates over sites and lineages (Kolaczkowski and Thornton (2004); Brinkmann *et al.* (2005)). Unfortunately, no current model covers the full complexity of biological history that can minimize the inconsistency of methods caused by model misspecification (Steel (2005)).
In this study, we have attempted to identify the impact of systematic bias in our phy-

logenetic analyses by applying suitable methods of analysis to better match the data, and did not detect any severe model violations (see Figures C.1, C.2, C.3 and C.4 in the appendix). Adequate taxon sampling remains the other crucial factor in phylogenomic studies to avoid LBA artifacts. Increasing the number of ingroup taxa from 7 (*maxgen*) to 14 (*maxspe*) resulted in a congruent topology and support for the Chiastomyaria hypothesis, that is, a basal position of Odonata. However, given the existing data we are not in the position to significantly enlarge taxon sampling. At this time the Chiastomyaria hypothesis is well supported, but we are aware that possible pitfalls (LBA, wrong model of sequence evolution, gene sampling) exist. Thus, future extended analyses are necessary to finally confirm the Chiastomyaria hypothesis.

On the other hand, not only is the phylogenomic methodology or taxon sampling important but the genes/proteins to which it is applied are also of relevance. The evolutionary history of the genes that compose the data sets may have a direct impact on the reconstructed phylogeny (Comas *et al.* (2007)). The phylogenetic signal of a gene is likely to be related to its evolutionary constraint and it has been suggested that a polytomy can be resolved by using genes that evolve at the optimal rate in the relevant timescale (Townsend (2007)). We therefore assessed the biological function of the represented genes and concatenated them according to their functional classification with the assumption that they harbor the same evolutionary history along the branches of the organismal phylogeny. It has been known that different evolutionary signals are a result of the different evolutionary processes that act upon the genes and that the functional role of these genes in the cell is important for the phylogenetic signal they carry (Graur and Li (2000)).

The statistical tests of concatenated alignments based on their functional classification showed that proteins belonging to the cellular processes and signaling category seem to harbor the strongest phylogenetic signal for resolving deep phylogenetic relationships. Our results are congruent with a phylogenetic study of the fungal kingdom (Kuramae *et al.* (2007)). These authors evaluated phylogenetically informative proteins for the fungal Tree of Life and identified proteins involved in cellular processes and signaling as phylogenetically more informative than the others.

Nevertheless, the large data set based on KOG (Eukaryotic Orthologous Groups) categories (*maxgen*: cellular processes and signaling = 5,285aa, information storage and processing = 4,697aa, metabolism = 3,143aa, and poorly recognized proteins = 1,179aa; *maxspe*: cellular processes and signaling = 3,511aa, information storage and processing = 4,245aa, metabolism = 1,551aa, and poorly recognized proteins = 329aa), gave in the majority of analyses no strong statistical support for any one hypothesis. There are several explanations for this observation. First of all, multiple substitutions at the same positions are expected to be frequent because the speciation event occurred millions of years ago. The saturation of the molecular markers will certainly reduce the phylogenetic signal and consequently the resolution. To investigate this, we conducted ML analyses for

each protein separately using Tree-Puzzle (WAG+4Γ+I) and RAxML (PROTMIXWAG). As expected, due to the limited number of alignment positions, the analyses from the individual alignments have shown that one gene did not harbor enough phylogenetic signal to unequivocally resolve the "Palaeoptera problem". Although the conserved nature of housekeeping genes is beneficial to track Mesozoic divergences, the phylogenetic content of single genes is too low, whereas concatenation seems to compensate for this fact.
It appears that the ancient rapid radiation that took place with the transition from nonwinged to winged insects represents one of the major obstacles for insect systematics. As we have shown for one of the major questions in insect phylogeny, molecular phylogenetics may overcome this hurdle by closing the gaps of genetic information from key orders, carefully applying multigene approaches and assessing the data quality.

Table 8.3: These genes were assembled in the four major KOG (Eukaryotic Orthologous Groups) categories: (1) cellular processes and signaling, (2) information storage and processing, (3) metabolism and (4) poorly characterized. ID number –the numerical identifier assigned to the gene during the HaMStR process, FlyBaseID/gene name– the corresponding ID number/gene name of the Drosophila melanogaster genome database (`http://flybase.org/`). maxspe/maxgen –genes represented in the alignments. These genes were also selected for the extended ML analyses of individual alignments.

KOG cat.	ID	FlyBaseID	gene name (FlyBase)	*maxspe*	*maxgen*
(1) cellular processes and signaling	6936	FBgn0038166	CG9588		+
	7538	FBgn0034709	CG3074	+	+
	7640	FBgn0015282	Proteasome 26S subunit 4 ATPase	+	+
	8073	FBgn0023174	Proteasome β2 subunit	+	+
	8075	FBgn0003150	Proteasome 29kD subunit	+	+
	8671	FBgn0033663	ERp60	+	+
	9489	FBgn0002174	lethal (2) tumorous imaginal discs		+
	8547	FBgn0010638	Sec61β	+	
	8032	FBgn0010226	Glutathione S transferase S1	+	+
	8784	FBgn0011217	effete	+	+
	8782	FBgn0010602	lesswright		+
	9616	FBgn0037756	CG8507		+
	9827	FBgn0025637	skpA	+	+
	7864	FBgn0036928	Translocase of outer membrane 20		+
	7902	FBgn0037231	CG9779		+
	8323	FBgn0024833	AP-47		+
	9169	FBgn0021814	Vps28		+
	7720	FBgn0025700	CG5885	+	+
	9562	FBgn0028985	Serine protease inhibitor 4	+	+
	7339	FBgn0011760	cut up	+	+
	9414	FBgn0052672	Autophagy-specific gene 8a	+	+
(2) information storage and processing	9511	FBgn0014189	Helicase at 25E		+
	7970	FBgn0001197	Histone H2A variant	+	+
	7512	FBgn0037346	extra bases	+	+
	6671	FBgn0029897	Ribosomal protein L17	+	+
	6790	FBgn0001942	Eukaryotic initiation factor 4a	+	+
	6906	FBgn0034967	eIF-5A	+	+
	6927	FBgn0037351	Ribosomal protein L13A	+	+
	7007	FBgn0010265	Ribosomal protein S13	+	+
	7098	FBgn0036213	Ribosomal protein L10Ab	+	+
	7316	FBgn0034743	Ribosomal protein S16	+	+
	7606	FBgn0005593	Ribosomal protein L7	+	+
	7883	FBgn0039713	Ribosomal protein S8	+	+
	7950	FBgn0034751	Ribosomal protein S24	+	+

KOG cat.	ID	FlyBaseID	gene name (FlyBase)	*maxspe*	*maxgen*
	8013	FBgn0002590	Ribosomal protein S5a	+	+
	8023	FBgn0010409	Ribosomal protein L18A	+	+
	8456	FBgn0034138	Ribosomal protein S15	+	
	8732	FBgn0064225	Ribosomal protein L5	+	+
	8997	FBgn0039129	Ribosomal protein S19b	+	+
	9404	FBgn0031980	Ribosomal protein L36A	+	
	9821	FBgn0036825	Ribosomal protein L26	+	
	6715	FBgn0024558	Diphthamide methyltransferase		+
	9590	FBgn0028737	Elongation factor 1 β	+	+
	7383	FBgn0023211	Elongin C		+
	7771	FBgn0023212	Elongin B		+
(3) metabolism	6637	FBgn0014028	Succinate dehydrogenase B		+
	9384	FBgn0011361	mitochondrial acyl carrier protein 1		+
	9813	FBgn0031436	CG3214		+
	9569	FBgn0028662	VhaPPA1-1		+
	7434	FBgn0039697	CG7834	+	+
	7214	FBgn0000116	Arginine kinase	+	+
	9594	FBgn0250814	CG4169		+
	6958	FBgn0036580	PDCD-5	+	+
	9095	FBgn0028833	Dak1		+
	9007	FBgn0250837	Deoxyuridine triphosphatase		+
	7631	FBgn0032192	CG5731	+	+
	8076	FBgn0033879	CG6543	+	+
(4) poorly characterized	8942	FBgn0024188	separation anxiety		+
	7015	FBgn0086254	CG6084	+	+
	7736	FBgn0035528	CG15012		+
	7742	FBgn0038739	CG4686		+
	8092	FBgn0030724	Nipsnap		+

9 Aspects of EST-based Phylogenetics

9.1 Introduction

Public EST databases are growing constantly. Data for an increasing number of species appear in these databases and ESTs for already present species are continuously added. With the appropriate strategies, these tremendous amounts of sequence data can be utilized to compile data sets for phylogeny reconstruction of unprecedented sizes. However, the resulting trees, although in general good resolved, still contain parts of uncertainties. For example, in Chapter 8 we employed an EST-based data sets containing sequence data from 125 genes to recover the splitting order of the three major groups found in winged insects. The resulting phylogenetic tree was robust, as we have proven with various tests (see Sec. 8.3.1). But the tree also contained splits, not relevant for our main conclusions, that are not fully resolved (c.f. Fig. 8.2), despite the fact that only 18% of the underlying data were missing.
This observation is not limited to our own analyses, but a common phenomenon. For example, Philippe *et al.* (2004) used 129 genes to successfully reconstruct the evolutionary relationships of selected eukaryotes in general, but failed to resolve the splits separating arthropods, deuterostomes, nematodes, and platyhelminths in particular. Similarly, Dunn *et al.* (2008) employed 150 genes to assemble an overall well supported tree of 77 animal species, which, however, showed only weak support for the inferred topology of chordates and other clades.
In Rokas and Carroll (2006), the authors discuss that these hard to resolve clades might have evolved with a specific pattern. They argue that if two splits in a lineage of species occurred in fast succession, only few changes could have been accumulated between these splitting events, yielding a short internal branch in a phylogenetic tree. If external branches, leading to the tips of the tree, are long, the weak phylogenetic signal generated by a short internal branch might be completely obscured. Then, the splitting order of the lineages cannot be recovered, independent of the amount of sequence data that is incorporated.
Here, we want to discuss an aspect of EST-based phylogeny reconstruction that is related to this matter, but has never been considered before.

The generation of ESTs is not directed towards certain genes, but a random process.

Consequently, each EST project contains sequences of a unique subset of expressed genes. Hence, if the data is considered as a matrix in which the rows represent all taxa and the columns all genes with available sequence data in any taxon (*taxa-gene matrix*), there will be empty cells. To compile an informative data set for phylogeny reconstruction, the taxa-gene matrix should be condensed to minimize the number of empty cells, while maximizing the number of genes and taxa considered. Finding an optimal solution, i.e., removing as few taxa/genes as necessary to obtain an as complete as possible matrix is mathematical not trivial (van Uitert *et al.* (2008)). For that reason *ad hoc* solutions are often applied. For example Dunn *et al.* (2008) excluded all genes from their initial data set, which were represented in less than 25 of the 77 taxa they considered. They additionally used a criterion that chooses preferentially genes also present in relatively small EST projects to ensure a large taxon sampling. Philippe *et al.* (2004) on the other hand, manually chose 174 genes that were "showing a reasonable taxonomic distribution". As we have demonstrated in section 7.3, there are genes, such as those encoding ribosomal proteins, which are ubiquitously highly expressed and thus can be found in the majority of EST projects. By condensing the taxa-gene matrix, the data set will be presumably enriched by such genes that are in general highly expressed.
Drummond *et al.* (2005) showed that in yeast there is a clear negative correlation between the average expression level of a gene and its evolutionary rate. They found that genes which are in general highly expressed, evolve slower than lower expressed ones. It was proposed that a selection pressure towards robustly folding amino acid sequence is causing the reduced mutation rates in highly expressed genes: The translation of mRNAs into amino acid sequences is not 100% accurate. Amino acid sequences that robustly fold into a functional protein structure despite some translation errors provide an advantage for the cell as less resources are wasted by synthesizing non-functional proteins. Once such a robust sequence is established, mutations yielding less robust folding polypeptides will be removed by purifying selection. Since the majority of all possible mutations will reduce the robustness of the correct folding and consequently be removed, the rate of manifested mutations is decreased in such genes.
This implies that the selection for genes found in many EST projects can introduce a biased towards slowly evolving genes. Here, we analyze whether the compilation of EST-based data sets indeed introduces a bias towards slowly evolving genes, and the effects such a bias has on the phylogenetic tree reconstruction. As test scenario we compiled an EST-based sequence set suited to examine the evolution of chordates.

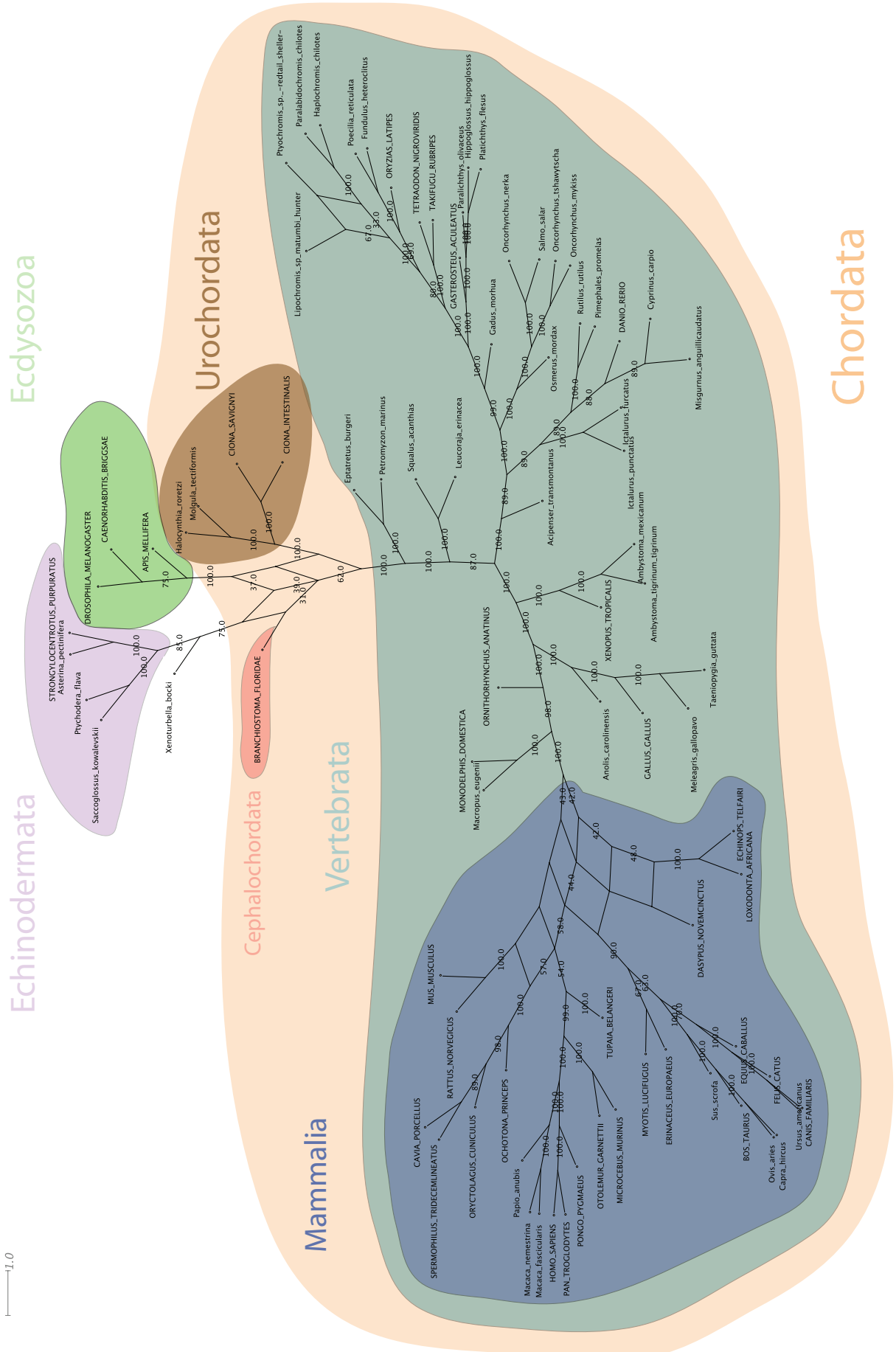
9.2 Compilation of Test Data

We screened 123 taxa belonging to the chordates, hemi- and urchordates or echinodermates with HaMStR (Chapter 7) to identify orthologs to the genes of the *Chordata set* (see

section 7.3 for details). This set contains 1000 profile Hidden Markov Models (pHMMs, Durbin (1998a)) trained with orthologous protein sequences from eight chordates, one echinodermate, two arthropods and two nematodes proteomes.
The coding sequence in putative orthologous EST contigs was determined and translated into the amino acid sequence with GeneWise (Birney *et al.* (2004)) using the presumably evolutionary closest protein sequence included in the pHMM as guideline for a codon alignment.
To increase the taxon sampling, we further searched for orthologs in 25 taxa with completely sequence genomes (see Table D.1 in appendix for a complete taxon list).
For preparation of the final data set, we excluded taxa with less than 50 genes to increase the completeness of the data matrix. Afterwards, a PERL script (courtesy of Ingo Ebersberger, unpublished) was used that heuristically tries different combinations of genes and taxa and reports the amount of missing data for each combination. The user can choose a taxa/genes combination that seems appropriate. Here, we decided to use 102 genes and 82 taxa (37 proteomes, 45 ESTs) resulting in a data matrix with only 17% of missing data. Such a low amount of missing data indicates, that all considered genes can be found in the majority of EST projects and therefore are likely to be highly expressed in general.
The 102 orthologous groups consisting of the translated EST contigs and the protein sequences from the fully sequenced taxa were individually aligned with mafft (Katoh *et al.* (2005)) with the parameters `--localpair` and `--maxiterate 1000`. Subsequently, we concatenated the 102 gene alignments resulting in a superalignment of 42,561 amino acid positions. Missing data in the alignment were represented by an 'X'.
Alignment columns with more than 50% gaps or missing data were removed, leaving 31,719 positions. A maximum likelihood tree was calculated with RAxML (Stamatakis (2006)) using the WAG model of amino acid substitutions (Whelan and Goldman (2001)). 100 bootstrap replicates were sampled to assess the support values for individual splits.

9.3 The EST-based Chordata Phylogeny

The commonly used tree representation with bootstrap support values is a convenient and easy to interpret way to visualize the relationships between taxa. However, this representation hides a lot of information, because there might be a substantial amount of support for alternative topologies not shown in the tree. A consensus network reveals such alternative topologies. It represents all splits with a support of a specified bootstrap value. This can result in alternative paths connecting taxa for which the phylogenetic signal in the data is ambiguous. Figure 9.1 shows the consensus network of the bootstrap replicates determined with SplitsTree (Huson and Bryant (2006)) in which splits that are present in at least 30% of all bootstrap replicates are drawn. This reveals that


Chordata
Urochordata
Cephalochordata
Vertebrata
Mammalia
Ecdysozoa
Echinodermata
1.0
Ptychromis_sp._-redtail_sheller-
Paralabidochromis_chilotes
Haplochromis_chilotes
Poecilia_reticulata
Fundulus_heteroclitus
ORYZIAS_LATIPES
TETRAODON_NIGROVIRIDIS
TAKIFUGU_RUBRIPES
GASTEROSTEUS_ACULEATUS
Paralichthys_olivaceus
Hippoglossus_hippoglossus
Platichthys_flesus
Lipochromis_sp._matumbi_hunter
Gadus_morhua
Oncorhynchus_nerka
Salmo_salar
Oncorhynchus_tshawytscha
Oncorhynchus_mykiss
Osmerus_mordax
Rutilus_rutilus
Pimephales_promelas
DANIO_RERIO
Cyprinus_carpio
Ictalurus_furcatus
Misgurnus_anguillicaudatus
Ictalurus_punctatus
Ambystoma_mexicanum
Ambystoma_tigrinum_tigrinum
Acipenser_transmontanus
Petromyzon_marinus
Epatretus_burgeri
Squalus_acanthias
Leucoraja_erinacea
CIONA_SAVIGNYI
CIONA_INTESTINALIS
Molgula_tectiformis
Halocynthia_roretzi
BRANCHIOSTOMA_FLORIDAE
DROSOPHILA_MELANOGASTER
CAENORHABDITIS_BRIGGSAE
APIS_MELLIFERA
STRONGYLOCENTROTUS_PURPURATUS
Asterina_pectinifera
Ptychodera_flava
Saccoglossus_kowalevskii
Xenoturbella_bocki
ORNITHORHYNCHUS_ANATINUS
MONODELPHIS_DOMESTICA
Macropus_eugenii
XENOPUS_TROPICALIS
Anolis_carolinensis
GALLUS_GALLUS
Taeniopygia_guttata
Meleagris_gallopavo
DASYPUS_NOVEMCINCTUS
ECHINOPS_TELFAIRI
LOXODONTA_AFRICANA
MUS_MUSCULUS
RATTUS_NORVEGICUS
CAVIA_PORCELLUS
SPERMOPHILUS_TRIDECEMLINEATUS
ORYCTOLAGUS_CUNICULUS
OCHOTONA_PRINCEPS
Papio_anubis
Macaca_nemestrina
Macaca_fascicularis
HOMO_SAPIENS
PAN_TROGLODYTES
PONGO_PYGMAEUS
OTOLEMUR_GARNETTII
MICROCEBUS_MURINUS
TUPAIA_BELANGERI
MYOTIS_LUCIFUGUS
ERINACEUS_EUROPAEUS
Sus_scrofa
EQUUS_CABALLUS
BOS_TAURUS
Ovis_aries
Capra_hircus
FELIS_CATUS
Ursus_americanus
CANIS_FAMILIARIS

the tree is well resolved for most of its parts. There are only two areas in which severe conflicts can be observed: At the base of mammals and at the base of chordates, between cephalochordates, urochordates and the outgroups. The latter conflict will be examined more in detail, especially if it is caused by the choice of genes.

9.4 Background Information about the early Evolution of Chordata

The chordates are distinguished into three major groups, whose evolutionary relationships were subject of constant argument. Traditionally, the cephalocordates, represented by *Branchiostoma floridae*, were considered as sister group to the vertebrates. Both taxa form a clade called *Euchordata*. The urochordates (also known as tunicates) were considered as the most basal chordate group (Zeng and Swalla (2005)). This topology is weakly supported by morphological characters as well as some molecular markers (Winchell *et al.* (2002)). In contrast, recent analyses based on large-scale molecular data suggest that the cephalocordates form the earliest branching chordate group, while urochordates and vertebrates diverged later (a clade termed *Olfactores*) (Putnam *et al.* (2008)). This topology is also referred to as *new chordate phylogeny* and nowadays widely accepted.

9.5 Analyses of Conflicts

9.5.1 Compilation of a Data Set

To focus on the gist of this investigation, we are considering only one representative for each of the three chordate groups: *Branchiostoma floridae* for the *Cephalochordata*, *Ciona intestinalis* for the *Urochordata* and *Homo sapiens* as representative of the vertebrates. We further added the sea urchin (*Strongylocentrotus purpuratus*) as outgroup to our taxon set. We have chosen these species, because for each at least a draft version of the genome is available. This allows us to compare genes over their entire length and not only the part that is covered by an EST. By that, potential artifacts are avoided.

Figure 9.1 ***(facing page)*:** The network is based on 100 ML bootstrap replicates of 102 concatenated gene alignments from 82 taxa. It was calculated with SplitsTree, considering only splits that occurred in at least 30% of the bootstrap replicates. The support for each individual split is given by the number at each branch. Taxa entirely written in capital letters are represented by proteomic data, while for all other EST data was used.

With the exception of *Branchiostoma floridae* all taxa were used to compile the *Chordata set* for the HaMStR search. Thus, orthologs for each gene contained in the *Chordata set* have been already determined in these three species. To complete our data, we performed a HaMStR search with the *Chordata set* in the proteome of *Branchiostoma floridae*. For 980 genes an ortholog could be assigned.
Those 980 genes were split into two subsets: the 102 genes we used to reconstruct the chordate network shown in Figure 9.1, called *tree subset* hereafter, and the remaining 878 genes, named *complement subset*.

9.5.2 Likelihood Mapping

To get an impression of the phylogenetic signal present in the alignment of every participating gene, we used likelihood mapping (LM, Strimmer and von Haeseler (1997)). This method provides an easy to interpret visualization of a gene set regarding its phylogenetic information content. Mapping the complete set of 980 genes (Figure 9.2) shows that the majority of genes (41.5%) support a grouping of *Homo sapiens* with *Ciona intestinalis* and hence the new chordate phylogeny. Approximately one quarter (24%) supports the traditional topology. The third possible topology, grouping *Homo sapiens* with the outgroup (*Strongylocentrotus purpuratus*), is supported by ∼12%. The remaining genes (22.8%) cannot be unequivocally assigned to one topology, but contain contradicting or only little signal. We next repeated the analysis for the two subsets and indeed they differ regarding their phylogenetic signal. The LM plot of the *complement subset* differs only slightly from that of the total set (Fig. 9.3A). In contrast, the LM plot for the *tree subset* shows a completely different picture (Fig. 9.3B). The amount of genes supporting the new chordate phylogeny dropped by 13.5% and is not supported by the majority of genes anymore. Each of the two other topologies gained about 10% of support. This provides an explanation why the *tree subset* fails to resolve the base of chordates (c.f. Figure 9.1). Our analysis clearly shows that the *tree subset* does not equal a random sample of the total gene set, because otherwise one would expect the percentages of the LM plot to stay roughly within the range of that of the total set, as it can be seen for the *complement subset*.

9.5.3 Maximum Likelihood Tree Analysis

LM analyses the phylogenetic signal of individual genes. However, the Chordata network shown in Fig. 9.1 is based on a concatenated alignment. It has been repeatedly shown, that concatenation of single gene alignments effects better resolution (e.g. Philippe *et al.* (2004)). In order to further validate the differing signal strength of both subsets, we concatenated the gene alignments of each subset. Subsequently, we calculated a maximum

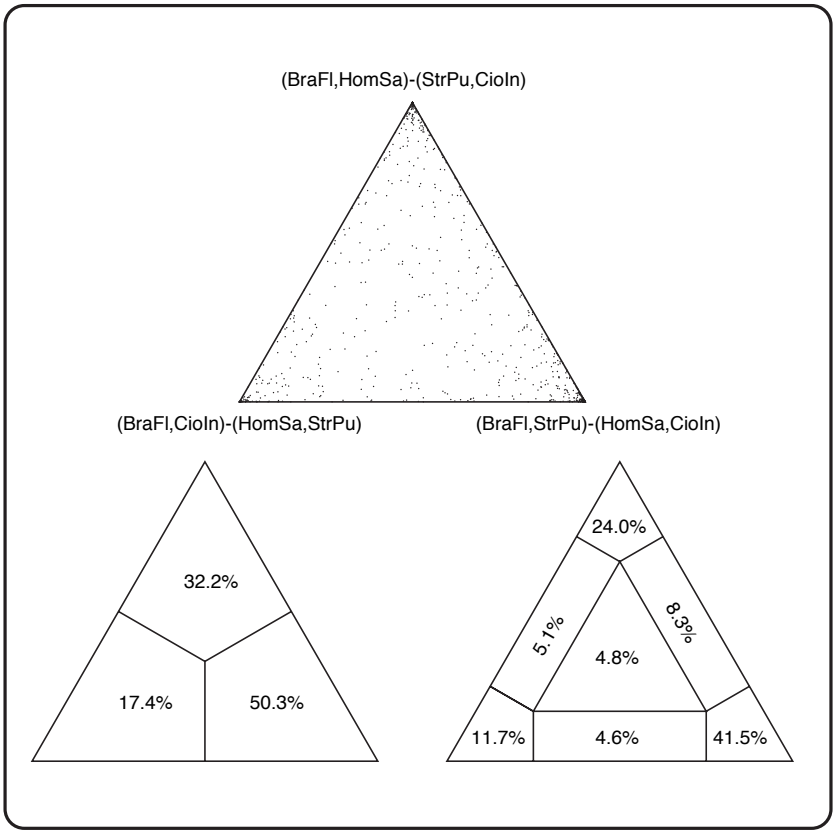


Figure 9.2: Likelihood Mapping of the Chordata gene set Every dot in the top triangle represents one of the 980 genes in our total gene set. The three possible evolutionary relationships between the four taxa are given at the three corners of the triangles. BraFl = Branchiostoma floridae, HomSa=Homo sapiens, CioIn=Ciona intestinalis and StrPu=Strongylocentrotus pupuratus. The bottom right corner corresponds to the new chordate phylogeny (Olfactores), while the top corner complies to the traditional view (Euchordata). The localization of dots indicates the phylogenetic signal contained in the alignments. The closer a dot is drawn to one of the corners, the stronger the alignment supports the corresponding tree topology. Alignments placed in the center area do not support a particular tree. The bottom triangles give the percentages of genes located in the defined sections.

likelihood tree with bootstrap support for the inner branch, using the same parameter as mentioned earlier. Again, both subsets deliver contradicting results (Figure 9.4). The *complement subset* contains a strong signal yielding a fully resolved tree that is congruent to the new chordate phylogeny. On the contrary, the *tree subset* supports the traditional grouping, although the support is weak. Furthermore, tree based on the *tree subset* has shorter branches, which indicates that the genes included in this set might evolve slower on average than the remaining genes of the *Chordata set*. However, the weak signal could

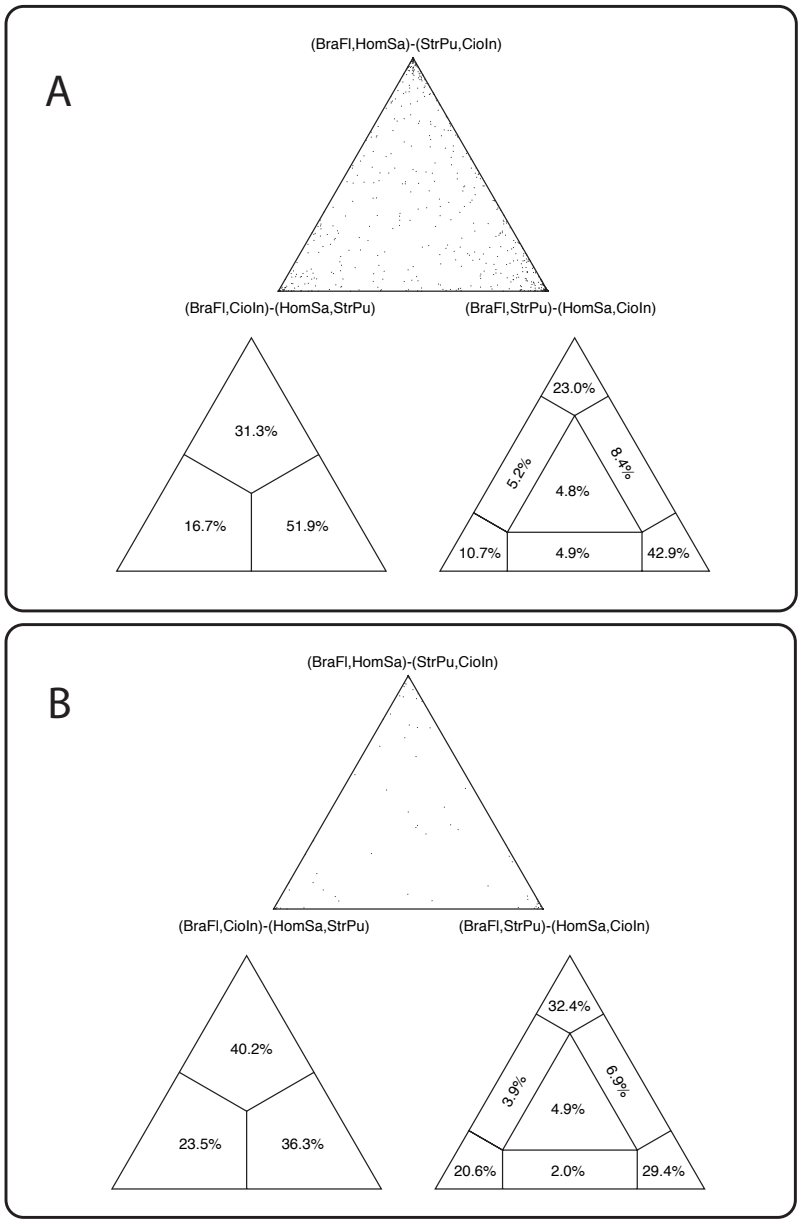


Figure 9.3: Likelihood mapping of the Chordata subsets
(A) Likelihood mapping of the 878 genes of the *complement subset*.
(B) Likelihood mapping of the 102 genes of the *tree subset*. The arrangement of the three possible tree topologies at the three corners is analog to Fig. 9.2

be due to the fact, that the *tree subset* contains much fewer genes than the *complement subset* (but see Section 9.5.5).

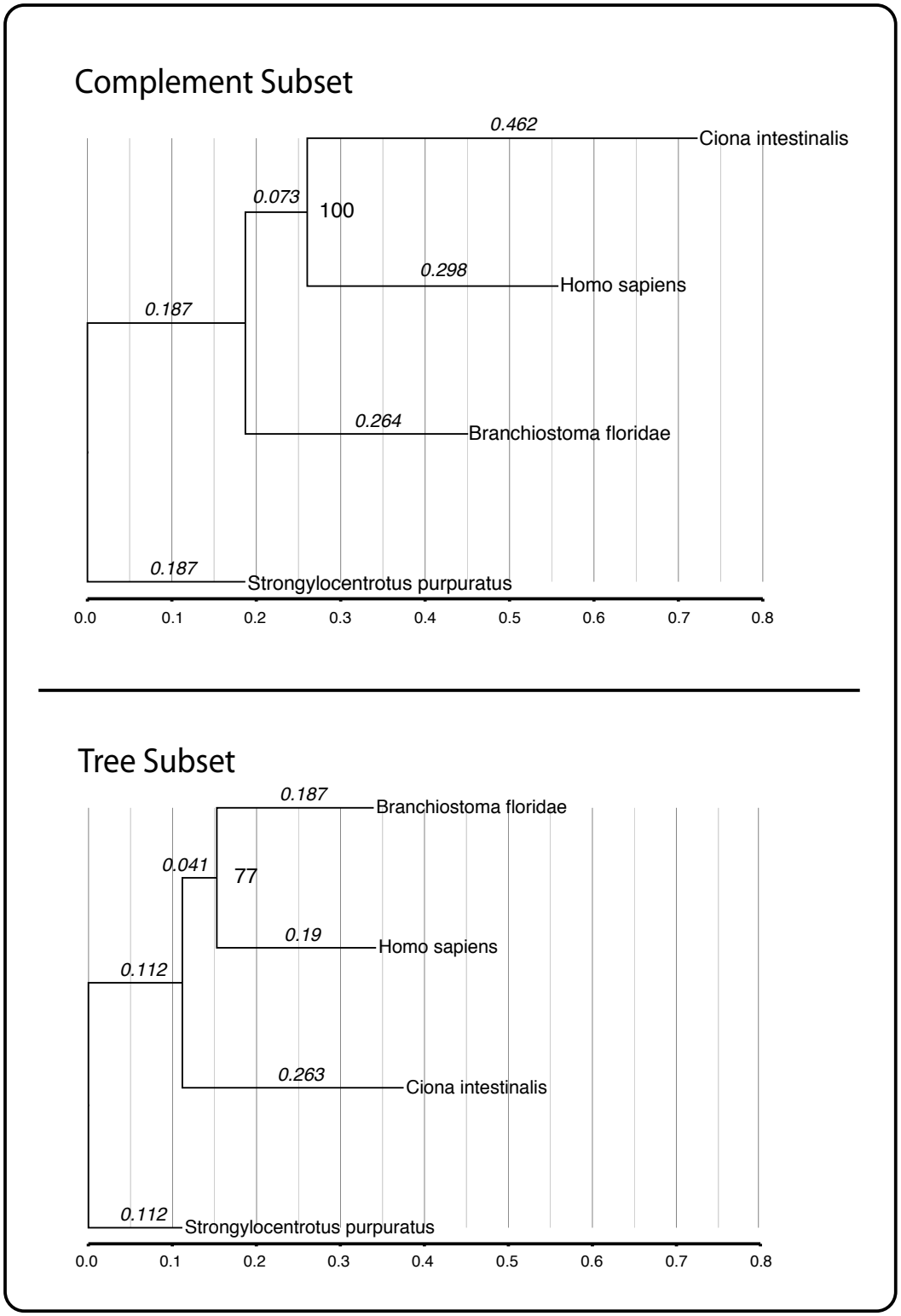


Figure 9.4: Maximum Likelihood Trees of gene subsets Both shown trees have been calculated with RAxML and the WAG model, with 100 bootstrap replicates. The tree in the top section is based on the *complement subset*, the tree in the bottom section is based on the *tree subset*. The individual branch lengths (substitutions per alignment position) are written in italic at each branch. The bootstrap support for each tree is written at the inner node. Both trees are shown with equal scaling to illustrate the different branch lengths.

9.5.4 Determination of the Evolutionary Rates

To shed more light on whether the weaker signal in the *tree subset* is related to the evolutionary rates of its genes, we calculated the pairwise maximum likelihood distances (MLd) between each sequence pair in each alignment. For this purpose we used TREE-PUZZLE v. 5.2 (Schmidt *et al.* (2002)) and the WAG substitution model. Subsequently, we averaged over all six values per alignment. This gives us an estimate of the evolutionary rate of each gene independent of a tree topology. Fast evolving genes will have accumulated more mutations on all branches, which will result in a greater average pairwise distance between the sequences.
The mean of all average distances of the *tree subset* is 0.43 mutations per site, for the *complement subset* it is 0.7. We sought to substantiate this difference statistically and compared the obtained distributions of averaged distances for both subsets with a one-sided Kolmogorov-Smirnov test. This test evaluates if two groups of values are samples from the same distribution or stem from different ones. The null hypothesis, both subsets stem from the same distribution, was rejected with a p-value $< 3.3 * 10^{-16}$. To conclude, the sequences of the *tree subset* show significantly shorter distances than the *complement subset* and thus evolved significantly slower.

9.5.5 Compilation of Random Gene Sets

To exclude the possibility that the weak phylogenetic signal contained in the *tree subset* is just a coincidence and not due the specific selected genes, we empirically determined the probability of composing a gene set with a similar phylogenetic signal as found in the *tree subset*. To this end, we generated 100 random sets with a PERL script, each containing 102 from the 980 genes in the total set. The alignments of each sample were concatenated and we executed a maximum likelihood tree reconstruction with bootstrap support as described in section 9.2. This should also reveal, whether the lack of resolution when using the *tree subset* is only due to its smaller size compared to the *complement subset*. The results, see Figure 9.5, indicate that the choice of the specific genes included in the *tree subset* is responsible for the weak phylogenetic signal. The vast majority (98%) of samples support the new chordate phylogeny, in general with good bootstrap support (in 80% of all sampled sets with a bootstrap value > 95). In fact, only 2% of all samples support the traditional grouping of vertebrates with cephalocordates. The third possible tree topology is not recovered at all. We conclude that, given the data, compiling a gene set that supports the traditional chordate phylogeny by chance is very unlikely. This suggests that selecting those genes of the *Chordata* core-ortholog set which are most frequently found in EST projects, introduces a bias towards slowly evolving genes. The consequence of this bias is, in this case, the wrong inference of the evolution of chordates.

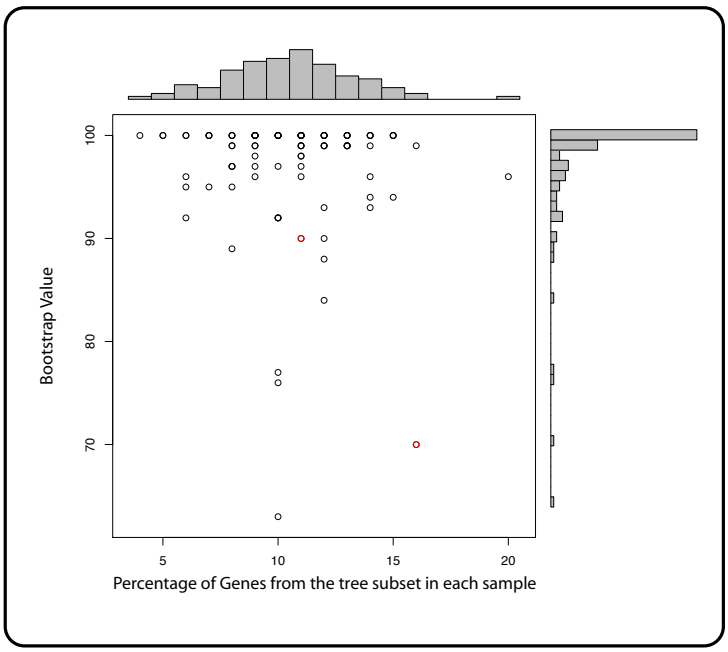


Figure 9.5: Results of ML tree reconstructions based 100 random gene samples Each circle represents a randomly chosen subset of 102 genes. The gene alignments of each subset have been concatenated and a maximum likelihood tree with 100 bootstrap replicates was calculated. On the x-axis the percentages of genes from the *tree subset* contained in each sample are shown. The y-axis gives the bootstrap value for the internal branch of the four taxa Maximum Likelihood (ML) tree. As circles can be stacked on each other, marginal histograms are added to illustrate the distribution. Only two subsets (highlighted in red) resulted in a ML tree that shows the traditionally assumed monophyly of *Branchiostoma floridae* and *Homo sapiens*. In all other 98 subsets, the obtained topology is congruent to the new chordate phylogeny, i.e. *Branchiostoma floridae* is grouped with the echinodermate.

9.5.6 Gene Expression

So far, we have shown that the *tree subset* does contain mainly slowly evolving genes. We assumed that this is due to our gene selection criterion that favors highly expressed genes. Next, we want to explore, whether the gene selection is indeed connected with gene expression levels.

The Genomic institute of the Novatis Research Foundation (GNF) provides an extensive data set[1] containing gene expression values of 12,605 human genes measured in 79 different tissues with Affymetrix chips (Su *et al.* (2004)). Of the 980 genes considered in our analysis, we could extract expression values for 815 (98 of the *tree subset*, 717 of

[1] http://wombat.gnf.org/index.html

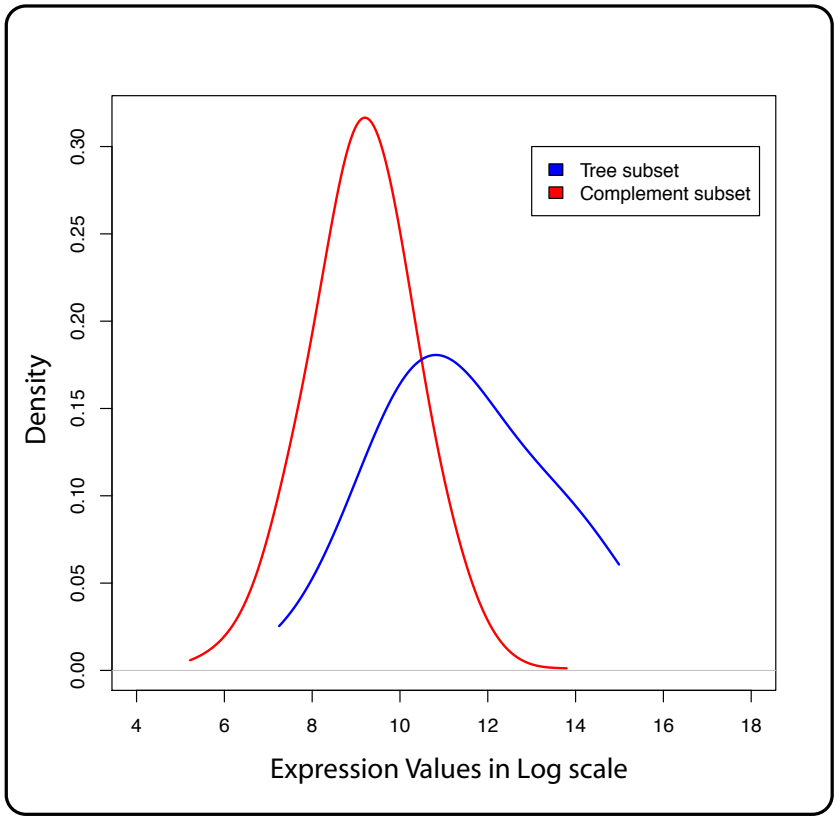


Figure 9.6: Gene expression values in human This plot shows the distributions of gene expression values averaged over 79 human tissues for the two subsets (*tree subset* in blue, *complement subset* in red). Expression values on the x-axis are given in logarithmic scale.

the *complement subset*). Plotting the distributions of average expression values for both subsets reveals that on average the genes of the *tree subset* are higher expressed than those of the *complement subset* (Fig. 9.6). The difference between the two distributions is significant (Kolmogorov-Smirnov test, p-value $< 2.2 * 10^{-16}$). Indeed, the *tree subset* is enriched with higher expressed genes.

9.5.7 Correlation between Evolutionary Rate and Discovery Rate

Of course, the analysis of the gene expression levels is restricted to the human genes and does not allow to draw conclusions on the expression levels of these genes in general. In order to get a more generalized view on the correlation between evolutionary rate and the frequency a gene is found with in an EST project, we extended the analysis depicted in Section 7.3: We counted in how many clustering projects each gene of a core-ortholog

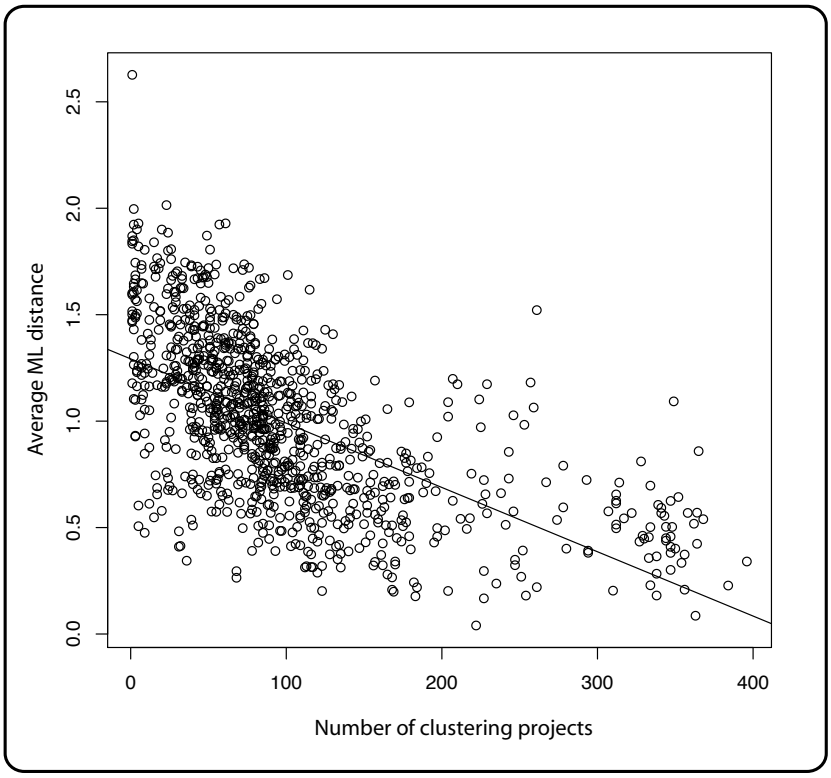


Figure 9.7: Correlation between discovery and mutation rates Each circle represents one of the 1035 genes included in the *Modelorganism set* that was found by HaMStR in at least one of the 606 screened clustering projects. On the x-axis the number of clustering projects each gene was in is plotted against the average maximum likelihood distance between primer taxa given on the y-axis.

set was found by HaMStR (c.f. Fig. 7.7). Now, to additionally measure the evolutionary rate of each gene, we calculated the average maximum likelihood distance between all possible sequence pairs of the primer-taxa. with Tree-Puzzle. Instead of the *Chordata set* we here use the *Modelorgansim set*, because the specific selection of primer-taxa of the latter allows to screen more clustering projects (see Section 7.3). Overall, we screened for the presence of 1035 genes in 606 clustering projects, covering the entire metazoan species tree.

Although the linear regression in Figure 9.7 does not fit well, a clear trend is observable. The majority of genes with a large average maximum likelihood distance is found in a few clustering projects only. On the contrary, genes found to be present in more than 300 clustering projects have a rather slow evolutionary rate. These findings support our hypothesis, that preferably selecting genes present in many clustering projects will result

in a gene set consisting of mainly slowly evolving genes.

9.6 Discussion

We have demonstrated that the common way of selecting genes in EST-based data sets for phylogeny reconstruction introduces a bias towards highly expressed genes, which tend to evolve with a below-average mutation rate. For the reconstruction of deep splits, i.e., speciation events, that took place hundreds of millions of years ago, choosing slow evolving genes is mandatory. If genes evolve too fast, every sequence position might have experienced several mutations, which obscures the phylogenetic signal. In that case, similarities between sequences are purely random. Such a state is called mutational saturated (Meyer *et al.* (1986)).
Furthermore, Philippe (2000) showed that slowly evolving genes help to avoid methodological artifacts such as long branch attraction, which leads to a grouping of taxa with long external branches regardless of the true phylogeny (Felsenstein (1978)). Slowly evolving genes are therefore favorable in that case and are frequently used for reconstructing deep metazoan phylogeny (e.g. Ruiz-Trillo *et al.* (2002)). From this point of view, EST driven phylogenies should result in robust trees and many examples of their successful application in phylogenetic studies exist (de la Torre *et al.* (2006); Dunn *et al.* (2008); Bourlat *et al.* (2006)).
In stark contrast, problems might arise for resolving splits that occurred in fast succession, called radiations. To resolve these splits, it is essential that sufficient mutations have accumulated in the short time between two speciation events. Otherwise, the phylogenetic signal is too weak and the order in which the lineages split cannot be reconstructed. Slower evolving genes fulfill this requirement less likely than faster evolving ones, due to their reduced mutation rates. Hence, gene selection strategies commonly used in EST-based frameworks which, as we have shown, restrict the selection of genes to slower evolving ones, appear suboptimal.

The base of the chordates seems to be an example of a radiation. Using the *tree subset*, the inferred length of the internal branch separating the two splits that gave rise to the three main chordate lineages, is only 0.041 (Fig. 9.4). This indicates that only few mutations occurred between the two splits, leaving only a weak phylogenetic signal. However, as we have demonstrated in section 9.5.5, by altering the gene set and by that including a higher number of faster evolving genes, a recovery of the splitting order of the chordate lineages is possible. Strikingly, literally any other gene set of the same size seemed to be better suited to resolve the basal chordate phylogeny than the gene set we obtained by enforcing an as complete as possible data matrix in the first place. But please keep in mind, that any other gene set would have also probably decreased

the taxon sampling for a large-scale analysis, because the faster evolving genes can be found less frequently in EST projects than the slowly evolving ones. Based on our results we suggest that if EST-based phylogenetic reconstructions fail to resolve all splits, one should consider the presents of radiation events. In these cases the choice of genes should not be driven by the number of includable taxa and the amount of missing data in the data matrix. We rather suggest to compile additional sets with a reduced taxon sampling to specifically address the relationships of the problematic taxa. By focusing on only a few species, the number of overlapping genes to choose from might increase, which in turn allows to include rather lowly expressed genes too. This would weaken the bias towards highly expressed and thus slowly evolving genes and may help to better resolve these splits. Of course, this heavily depends on the data at hand and might not be feasible if taxa are involved for which only a limited number of ESTs is available. Quite recently the second generation sequencing methods have been proven useful for EST sequencing (Roeding *et al.* (2009); Gibbons *et al.* (2009)). The further drop-off in prices for sequencing will hopefully encourage the community to ramp up the number of ESTs for taxa hitherto represented by small EST projects. In the future, a broad variety of sequence data concerning their evolutionary rate could be available for many taxa. This would allow to select sequence data that evolved with exact the right rate —not showing saturation effects, but enough changes to resolve fast occurring splits. Hopefully, this will help to finally determine the evolutionary relationships of taxa, that is still unclear despite the incorporation of sequence data from hundreds of genes.

Part II

TonB dependent transporter

Nothing in Nature is random. ... A thing appears random only through the incompleteness of our knowledge.

Benedict Spinoza

10 TonB-dependent Transporters

10.1 Introduction

In the previous part of this thesis we described our infrastructure to process the tremendous amounts of ESTs nowadays available. This allows us to utilize these sequences for compiling high-quality data for sound phylogeny reconstructions.
The permanent growing of sequence databases is, however, not limited to ESTs. The number of publicly available genome assemblies rises likewise. This is especially true for prokaryotic genomes. Due to their relatively small sizes of only a few mega bases, genomes from more than 1000 species have been already completely sequenced and annotated[1] (see Fig. 1.2). In contrast to EST data, full genomic data contains the complete inventory of genes present in an organism. This enables us to study the evolution of biological systems shared between individual species.
In the following we describe an exhaustive investigation in which we streamlined the analysis of several hundred eubacterial genomes. Our aim was to explore a protein family involved in nutrition transportation in Gram-negative bacteria. The outcome of this study has been published in Mirus, Strauss *et al.* (2009).

10.2 Background

Filamentous cyanobacteria contain molecular machines for oxygenic photosynthesis under all growth conditions (Adams and Duggan (1999)). These machines, as well as those involved in respiration and nitrogen metabolism, depend on non-proteinaceous cofactors such as iron (Kustka *et al.* (2002), Shcolnick and Keren (2006)). The level of iron found in cyanobacteria is generally one order of magnitude higher than in non-photosynthetic bacteria (Keren *et al.* (2004)) and accounts for about 0.1% of their biomass (Roger *et al.* (1986)). Even though iron and copper are required for the function of respiratory and photosynthetic complexes, their intracellular level has to be tightly controlled as these ions pose a risk of oxidation (Shcolnick and Keren (2006)). Therefore, the uptake of iron is highly regulated in order to avoid intoxication. On the other hand, it is

[1] `http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`

hypothesized that iron limitation might have been one of the selective forces in the evolution of cyanobacteria (Ting *et al.* (2002)), and one might speculate that those cyanobacteria with the most efficient iron uptake systems might have had an evolutionary advantage. To enhance iron uptake, eubacteria secrete low-molecular-weight iron chelators (siderophores) under iron-limiting conditions to complex environmental iron (Ferreira and Straus (1994)). The siderophore-iron complexes are bound by receptor proteins (TonB-dependent transporters, TBDTs) in the outer membrane which are composed of a transmembrane β-barrel domain, a so-called plug domain and a periplasmic exposed TonB box. The siderophore-iron is subsequently transferred to the cytoplasm by transport proteins in the cytoplasmic membrane (Wandersman and Delepelaire (2004), Miethke and Marahiel (2007)). This process is dependent on TonB which provides the energy required for the translocation of siderophore-iron complexes across the outer membrane (Andrews *et al.* (2003)). In order to facilitate this translocation, the periplasmic domain of TonB interacts with the TonB box of the loaded TBDT. It is proposed that TonB exerts a pulling force on the TonB box and, thereby, partially unfolds the plug domain enabling the translocation of the siderophore into the periplasmic space (Gumbart *et al.* (2007)). Several TBDTs have been identified. Beside the ones for iron transport (e.g. Clarke *et al.* (2001), Lee (1995)), TBDTs for e.g. nickel (Schauer *et al.* (2007)), disaccharides (for sucrose SuxA; Blanvillain *et al.* (2007), for maltose MalA; Neugebauer *et al.* (2005)), oligo- (CsuF; Cheng *et al.* (1995)), polysaccharides (SusC; Reeves *et al.* (1996)) or large degradation products of proteins (RagA; Nagano *et al.* (2007)) are described. The most intensively studied function of TBDTs is the iron uptake in Gram-negative bacteria. Three large classes are defined, namely transferrin-/lactoferrin-binding proteins, porphyrin and siderophore transporters (Braun and Killmann (1999)). In addition to the transport of iron across the outer membrane by TBDTs, an additional ferric iron uptake system is postulated, but the corresponding outer membrane receptor has not yet been identified (Cartron *et al.* (2006)). The TBDTs TbpA (transferring-binding protein A) and LbpA (lactoferrin-binding protein A) facilitate the uptake of iron from transferrin/lactoferrin, respectively; the uptake is also assisted by the lipoproteins TbpB and LbpB which face the extracellular side (Perkins-Balding *et al.* (2004)). The porphyrin-transporting TBDTs include HasR, HgbA, HmbR (heme; Clarke *et al.* (2001), Perkins-Balding *et al.* (2004)) and BtuB which transports the cobalt-complexing vitamin B12 (cobalamin; Ferguson and Deisenhofer (2002)). Heme uptake is especially important in bacterial pathogens, where various heme-containing compounds are utilized (Lee (1995)).

The siderophore TBDTs are further sub-classified according to their substrate –that is the chemical nature of the siderophore they bind. Siderophores belong inter alia to hydroxamates, catecholates, phenolates, citrates or combinations thereof (Miethke and Marahiel (2007)). For example, the siderophore transporters FepA, ViuA and IroN recognize catecholates, FhuA, FoxA and FhuE hydroxamate and FecA citrate.

The iron uptake system in cyanobacteria is not well understood. For the non-filamentous cyanobacterium *Synechocystis* sp. PCC 6803 the TBDTs encoded by *sll1206*, *sll1406*, *sll1409* and *slr1490* were partially characterized (Katoh *et al.* (2001), Singh *et al.* (2003)). For filamentous cyanobacteria such as *Anabaena* sp. PCC 7120 (also termed *Nostoc* sp. PCC 7120) only siderophore secretion (Simpson and Neilands (1976), Goldman *et al.* (1983), Clarke *et al.* (1987)) and the influence of enhanced or reduced iron levels on the growth (Massalski *et al.* (1981), Guikema and Sherman (1983), Hutber *et al.* (1977), Latifi *et al.* (2005)) were investigated. *Anabaena* sp. PCC 7120 secretes the hydroxamate-type siderophore schizokinen, allegedly the only siderophore secreted (Simpson and Neilands (1976), Goldman *et al.* (1983)). Only recently, a TBDT encoded by *schT* (*alr0397*) involved in the uptake of schizokinen was identified. The expression of the gene *schT* (*alr0397*) was mildly increased under a shortage of Fe3+. A *schT* knock-out mutant showed a moderate phenotype of iron starvation and the characterization of its siderophore-dependent iron uptake demonstrated the function of *schT* as a TonB-dependent schizokinen transporter (Nicolaisen *et al.* (2008)).

To learn more about iron transport systems in general and in cyanobacteria particularly we searched for genes coding for TBDTs based on previously experimentally characterized TBDTs. Subsequently, we assigned putative substrates for so far uncharacterized TBDTs according to their sequence similarity to already known TBDTs. We observed a substantial difference in the number of TBDT genes in the analyzed cyanobacteria.

10.3 Compilation of the Data

10.3.1 Literature Search for characterized TonB-dependent Transporters

By extensive literature search we obtained the GenBank IDs for 98 TBDT sequences, which were subsequently extracted from the NCBI database (Tab. E.1 in the appendix). For 67 of these transporters, experimental data about their substrate is available. The substrates of further 27 TBDTs have been predicted. These predictions are based on co-localization with genes of a specific metabolic pathway or on metabolic-specific co-regulation by either transcription factors or a riboswitch (Schauer *et al.* (2008)). The substrates of the remaining four transporters are still unknown.

10.3.2 Identification of TonB-dependent Transporters

In order to complement our data set with yet uncharacterized TBDTs, we downloaded 686 completely sequenced eubacterial genomes that were available in June 2008 from the NCBI ftp server[2], together with the corresponding protein sequences. To identify putative TBDTs encoded in the genomes, we used a sequence similarity driven approach: As described on page 114, all known TBDTs show the presence of a β-barrel domain, a TonB box and the plug domain. These domains are in part highly conserved. The PFAM database[3] (Finn *et al.* (2008)) provides profile hidden Markov models (pHHMs) (Durbin (1998b)) for the conserved fragments. PFAM entry PF00593 covers part of the β-barrel domain. The conserved part of the plug domain is represented by the entry PF07715. The TonB box is a short motif of about 8 amino acids and therefore not covered by a pHMM. Figure 10.1 exemplarily shows the localization of sequence parts of the *E. coli* TBDT FepA that match to the two mentioned pHMMs.

With each of the pHMMs we searched in all proteomes using the program HMMSEARCH from the hmmer package[4]. Since a functional TBDT should contain all conserved parts, we were only interested in those proteins that triggered a hit with both pHMMs. To identify such sequences, we wrote a PERL script (PARSE_HMMSEARCH.PL) that parses two hmmsearch outputs and reports protein sequences that appeared as hits in both of them. The user can set an E-value threshold to discard protein sequences that yielded only insignificant hits with one or both searches. Optionally, the PERL script downloads annotation data for each candidate from GenBank.

The chosen E-value threshold heavily influences the number of resulting candidates. A more relaxed cutoff yields more sequences, but also increases the risk of adding sequences which are no TBDTs. To evaluate what cutoff gives reasonable results, we tried different values and noted the total number of sequences that fulfilled our requirements together with the number of species for which at least one candidate TBDT was found (Fig. 10.2).

As expected, a more stringent (lower) E-value cutoff reduces both the total number of candidate proteins and the number of species, which are represented by at least one putative TBDT. However, the reduction is not linear. Lowering the threshold from 10^{-1} to 10^{-10} results in an only slight decrease of TBDT candidates (Fig. 10.2A). The number of species remains constant (Fig. 10.2B). Setting the cutoff to any value below 10^{-10} leads to a dramatic decline of the data. We decided to use a rather conservative setting, but at the same time tried to include as many species as possible. We therefore proceeded with our analyses considering only hits with an E-value $\leq 10^{-10}$ in both searches, which yielded 4,586 putative TBDTs. 36 of these sequences we had already found by our literature search and we subsequently removed them. Including the 98 TBDTs from the literature,

[2] `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/`
[3] `http://pfam.janelia.org/`
[4] `http://hmmer.janelia.org/`

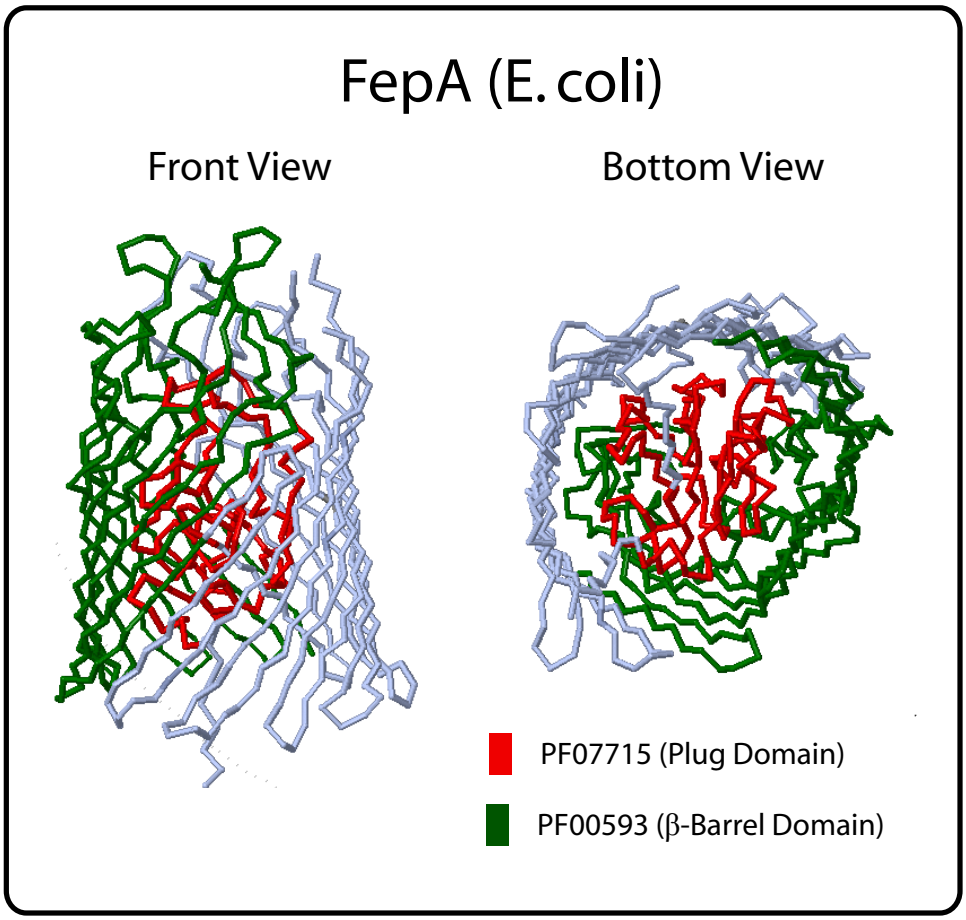


Figure 10.1: Three-dimensional structure of FepA Shown is the three dimensional structure of the *E. coli* TBDT FepA (GenBank ID: 6730010). Highlighted are the parts of the protein that match to the two PFAM pHMMs PF07715 (red) and PF00593 (green). The structural information was extracted from the protein data bank[4] (pdb) and rendered with RasMol[5].

---

our data set encompassed in total 4,648 proteins from 347 species (see Tab. E.2 in the appendix).

Compared to previously published bioinformatic approaches to study TBDTs (Blanvillain *et al.* (2007), Koebnik (2005)) our more stringent selection criterion led to the identification of fewer candidate sequences in the species which had been analyzed before (not shown). More specifically, within the species analysed by Koebnik, we selected seven sequences not previously identified, but did not consider 103 sequences (Koebnik (2005)). A similar ratio (+22/-142) was found when comparing the number of sequences selected by us found in species previously analyzed by Blanvillain *et al.* (2007), who selected 3020

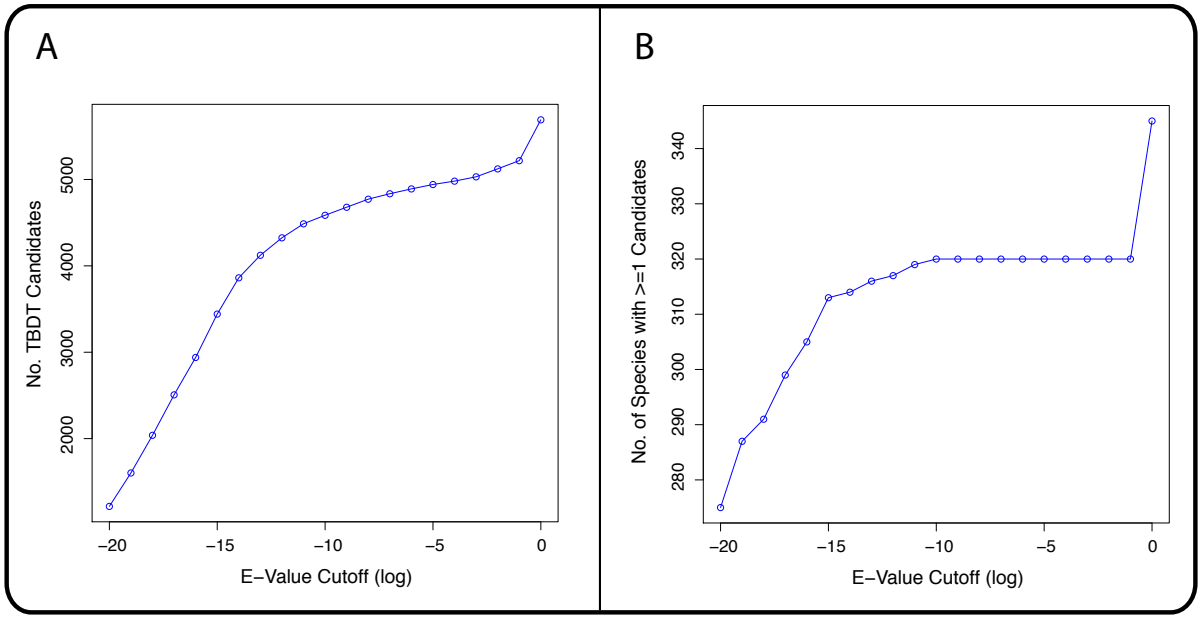


Figure 10.2: Ratio of E-value cutoff and data size (A) shows the relation between the E-value cutoff for the hmmsearch, given on the x-axis in logarithmic scale, and the total number of resulting putative TBDTs on the y-axis. Plot **(B)** shows for different E-value cutoffs the number of species which are represented by at least one candidate TBDT.

sequences which resulted in a discrepancy of about 5%.

10.4 Analyses

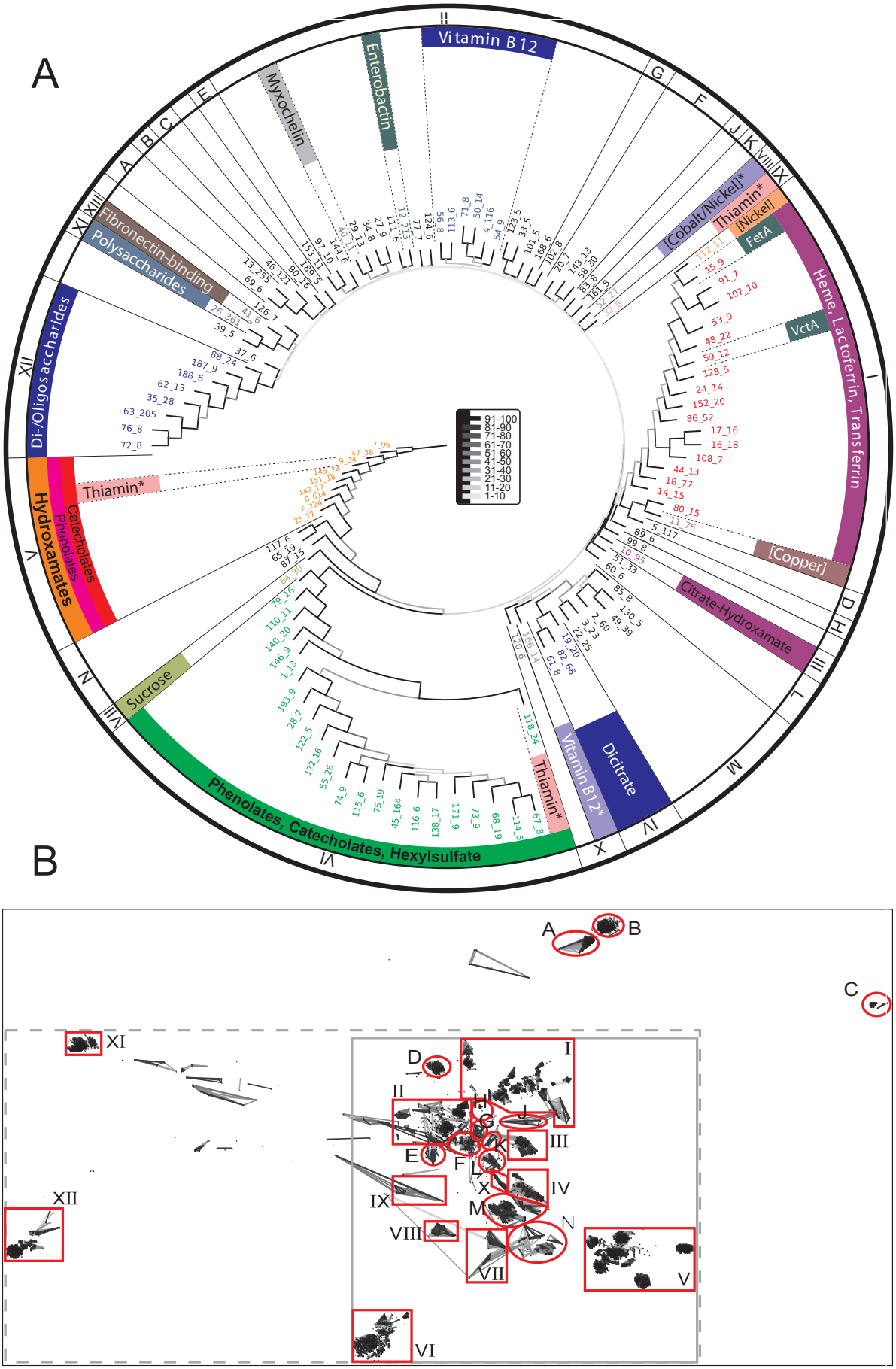
10.4.1 Clustering

In order to classify the candidate TBDTs with yet unknown substrates, we first performed a cluster analysis of the identified putative TBDTs, including the 98 published sequences. For this task we used the program CLANS (Frickey and Lupas (2004)), an implementation of an intuitive clustering algorithm. In brief, the program first performs a pairwise BLAST search with each sequence against all other sequences in a given set. In the second step, CLANS creates a three dimensional space, in which each sequence is represented by a dot. The arrangement of the dots is initially random. In an iterative procedure, two kinds of forces are then applied. First, dots are attracted by each other with a force

[4] `http://www.pdb.org`
[5] `http://rasmol.org/`

that is proportional to the negative logarithm of the pairwise BLAST P-value of the corresponding sequences. Thus, dots representing very similar sequences attract each other strongly, while dots of dissimilar sequences show no attraction at all. Second, a mild repulsive force is applied to all dots which is anti-proportional to the distance between dots in the 3d space. By that, a collapsing of all dots representing similar sequences into one place is prevented.
In each iteration, the force vectors for all dots are calculated and the dots are moved accordingly by a defined amount of space. Then, the next iteration is started and all force vectors are re-calculated, given the new positions of all dots. Afterwards, the dots are moved again and the next iteration begins.
As soon as the movement of dots becomes negligible in subsequent iterations, the program can be stopped. The resulting arrangement of all dots visualizes the degree of similarity of the corresponding sequences.
For our analysis, we set the cutoff such that only P-values $< 10^{-10}$ in pairwise BLAST searches were for the calculation of the attraction force during the CLANS-clustering. We further used the BLAST P-values to define clusters. Within a cluster, each dot is connected to at least one other dot of the same cluster with an edge that corresponds to a BLAST P-value $< 10^{-90}$. This criterion led to 195 clusters with at least two elements.

Figure 10.3 *(facing page)*: Clustering of the putative TonB-dependent transporters (TBDTs) The sequences found by the described genome wide searches were analyzed by CLANs as described. **(A)** shows the consensus tree of the pairwise mean cluster distances. The branches are colored according to their respective bootstrap value in shades of grey as indicated by the legend in the middle of the tree. The numbers at each leaf are of the format 'x_y', where 'x' is the cluster number and 'y' the number of sequences belonging to this cluster. We have further indicated the transported substrates. An asterisk marks predicted substrates. Brackets indicate that the metal ion is known, but the metallophore has not yet been identified. The regions as shown in **(B)** are marked by I to XII and A to N. **(B)** shows a two-dimensional transformation of the three-dimensional CLANS clustering. The regions from Figure **(A)** are marked by red polygons (containing at least a single exp/pTBDT) and red circles (no functionally characterized TBDT). Sequences with a high similarity (P-value $< 10^{-90}$) form a cluster. This is illustrated by lines colored in shades of grey (the darker the smaller the P-value) connecting the dots. The regions shown in Figure 10.4 (grey dashed line) and Figure 10.6 (solid grey line) are highlighted.


A
Vitamin B12
Enterobactin
Myxochelin
Fibronectin-binding
Polysaccharides
Di-/Oligos accharides
Thiamin*
Hydroxamates
Catecholates
Phenolates
Sucrose
Phenolates, Catecholates, Hexylsulfate
Thiamin*
Vitamin B12*
Dicitrate
Citrate-Hydroxamate
[Copper]
Heme, Lactoferrin, Transferrin
VctA
FetA
[Nickel]
Thiamin*
[Cobalt/Nickel]*
91-100
81-90
71-80
61-70
51-60
41-50
31-40
21-30
11-20
1-10
B
A
B
C
D
I
II
G
J
K
III
E
F
L
X
IV
IX
M
N
VIII
VII
V
VI
XI
XII

10.4.2 Phylogenetic Analysis of Clusters

To further elucidate the relationship of the 195 clusters, we performed a bootstrap analysis. To this end, we ran CLANS 100 times with a random initial configuration of the sequences in 3d space. After each run, we determined the cluster centers and computed pair-wise distances between centers. With the PHYLIP package v3.68 (Felsenstein (1989)) we constructed a neighbor-joining tree for the resulting 100 distance matrices and we inferred the majority rule consensus tree with support values for the splits in the consensus tree. Clades in the consensus tree (Fig. 10.3A) were translated into 'regions' on the two-dimensional sequence landscape (Fig. 10.3B). A region is marked by roman numerals if the substrate for at least one TBDT in this region has been experimentally verified (expTBDTs) or predicted (pTBDTs), and marked by upper case letters if no substrate TBDT in the region is known. Figure 10.3B shows the expTBDT regions I-VII, XI, XII and XIII and the pTBDTs regions VIII, IX and X together with the uncharacterized regions A-N. Figure 10.4 shows an enlarged version of the dashed rectangle in Figure 10.3B. The colors describe the substrate that binds to the corresponding TBDTs. Figure 10.4 (bottom) shows an enlargement of the expTBDTs regions, where the numbers refer to sequences with a known substrate (see Tab. E.1 in the appendix).

10.4.3 Classification of TonB-dependent Transporters

Based on the CLANS clustering, we assigned functions to our candidate TBDTs. We assumed that all transporters grouped within a cluster recognized the same substrate. Consequently, we adopted the confirmed functions of expTBDTs and pTBDTs for all candidate TBDTs that are found in the same cluster. In the following we describe these assignments more in detail.

Region I This region consists of 19 clusters with at least five sequences per cluster. Ten of these clusters contain sequences with a confirmed function such as porphyrin, lacto-/transferrin and nickel transporters (Figure 10.4). Cluster 11 contains the copper chelate binding protein OprC (sequence 1, for references see Tab. E.1 in the appendix). A group of clusters (15, 17, 18, 59, 86, and 107; sequences 11-28) are comprised of heme-transporting (HmbR) proteins. Remarkably, the two enterobactin (catecholate; see Tab. E.1 in the appendix) transporters VctA (cluster 59; sequence 28) and FetA (cluster 15; sequence 27) are located within the porphyrin group. This finding corroborates the observation that VctA and FetA are, supposedly, involved in transporting porphyrin (Beucher and Sparling (1995), Mey and Payne (2001)). In cluster 16, the lactoferrin-binding LbpA or transferrin-binding TbpA proteins are found (sequences 29 and 30). There is also a small

I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
A
B
C
D
E
F
G
H
J
K
L
M
N
Ni
Cu (OprC)
lacto-/transferrin (LbpA/TbpA)
porphyrin (e.g. HemR, HasR)
Catecholate (enterobactin/salmochelin)
cobalamin (BtuB)
cobalamin (predicted)
catecholate
catecholate (myxochelin)
citrate-hydroxamate (Aerobactin/Rhizobactin)
citrate
hydroxamate (FhuA)
catecholate-hydroxamate
citrate-hydroxamate
citrate-catecholate-hydroxamate
phenolate
thiamin (predicted)
cobalamin (predicted)
unknown substrate
phenolate
thiamin (predicted)
catechol
hexylsulfate
sucrose (predicted)
sucrose (SuxA)
Ni (predicted)
Co (predicted)
thiamin (predicted)
cobalamin (predicted)
digested proteins (RagA)
malto-oligosaccharides/ starch (SusC)
chondroitin sulfate/ hyaluronic acid (CsuF)
OmpW
maltose (MalA)
chito-oligosaccharides
pectin/Xylan
fibronectin-binding

cluster (number 112), which contains a single nickel-transporting expTBDT (sequence 31, FrpB4).

Region II Region II contains 20 clusters, three of them contain expTBDTs (cluster 4, 12, 40). Cluster 4 contains the experimentally confirmed cobalt-complexing vitamin B12 transporter BtuB (sequence 32). Moreover, predicted BtuBs were identified in the same cluster, and in clusters 160 and 165 (sequences 33-39). Cluster 12 contains IrgA, BfrA and IroN sequences (No. 2-10) transporting enterobactin, DHBS (catecholate) or salmochelin (glycosylated catecholate). In addition, a myxochelin (catecholate) transporter (sequence 40, cluster 40) occurs in region II.

Region III Cluster 10 is characterized by sequences of the aerobactin/rhizobactin (Citrate-hydroxamate; see Tab. E.1 in the appendix) transporters IutA and RhtA (sequence 41 and 42).

Region IV The largest cluster in region IV (No. 82) contains the sequences of the ferric rhizoferrin (carboxylate) transporter RumA and the diferric dicitrate transporter FecA (sequences 44 and 45).

Region V This region consists of nine clusters with three of them (clusters 0, 6 and 7) encompassing expTBDTs and two pTBDTs (clusters 9 and 25). The most frequently found transporters in this region are for hydroxamate-type siderophores, such as desferrioxamine (hydroxamate; cluster 0, sequences 50, 51, 56), ferrichrome (hydroxamate; cluster 0, sequence 55), pseudobactin A (citrate-catecholate-hydroxamate; cluster 6, sequences 62 and 63), pyochelin (phenolate; cluster 6, sequence 66), or anguibactin (catecholate-hydroxamate; cluster 7, sequences 69-71). The proteins for which sequences are found in cluster 9 (sequences 67 and 68) are predicted to transport thiamin. Interestingly, proteins

Figure 10.4 *(facing page)*: Distribution of characterized (experimental/predicted) TBDTs Shown is a blow-up of the dashed frame from Figure 10.3B. It includes all regions containing expTBDTs and pTBDTs, which are marked by colored symbols and a number. The numbers correspond to the numbering in column 1 of Table E.1 in the appendix. The dashed frames are shown in a magnified view on the bottom. Circles define regions without functionally characterized TBDTs. For regions XI and XII the substrates are indicated on the left. The region numbering is explained in the text.

46 and 47 (cluster 25) are predicted to transport vitamin B12. This appears to be a false prediction as judged from the large distance to cluster 4 (region II) containing the experimentally confirmed BtuB. Setting an even lower P-value (10^{-100} instead of 10^{-90}) as the threshold for defining the clusters in CLANS, leads to cluster 0 splitting up in the upper part with all hydroxamate-type TBDTs (including all cyanobacterial TBDTs of cluster 0) and the lower part containing phenolate-transporting TBDTs and VciA, which has been shown to transport neither heme, vibriobactin, enterobactin, ferrioxamine B, aerobactin nor shizokinen (Mey *et al.* (2008)).

Region VI This region represents transporters for phenolates, catecholates or hexylsulfate and contains several clusters. A hexylsulfate transporting TBDT (sequence 77) can be found in cluster 45, a vibriobactin (catecholate) transporter (sequence 74) in cluster 140, and proteins transporting yersiniabactin (phenolate; sequences 72 and 73) in cluster 79. As already observed in region V, we also detected two sequences (75 and 76) in cluster 118 that are putative thiamin transporters.

Region VII Cluster 67 contains the sequence SuxA (sequence 78) an experimentally verified sucrose transporter. Please note, that sequence 79 has been predicted to transport sucrose (Blanvillain *et al.* (2007)). The prediction was based on the co-localization in the genome of the corresponding gene with the transcriptional regulator ScrR. Thus, our bioinformatic analysis provides additional evidence for this functional characterization of sequence 79.

Region VIII This region contains predicted nickel and cobalt TBDTs with unknown metallophore specificity and no representative of the expTBDTs.

Region IX Region IX consists of eight sequences in one cluster (No. 32), where two sequences are putative thiamin transporters. However, proteins assigned as thiamin transporters were also found in regions V (sequences 67 and 68, cluster 7) and VI (sequences 75 and 76, cluster 118). Their genes are co-localized in the genome with a cytoplasmic membrane transporter for thiamin (PnuT, Schauer *et al.* (2008)), however, the functional assignment remains to be proven.

Region X This region contains a TBDT predicted to transport cobalt-complexing vitamin B12 (sequence 43, cluster 166). However, it is far away from the BtuB cluster in region II (Figure 10.4). Hence, the assigned function should be taken with a grain of salt.

Region XI This region is clearly separated from the rest and contains cluster 26. The experimentally characterized TBDTs include oligosaccharide (CsuF, sequence 88), polysaccharide transporters (SusC, sequence 87) and transporters for degradation products of proteins (RagA, sequences 85-86). While many taxa are represented by sequences in the region I-X, region XI consists almost exclusively of species of the phylum Bacteroidetes with the exception of a single δ-proteobacterial sequence (`gi|108757959`, *Myxococcus xanthus*). Thus, sequences in this region may represent a particular function reflecting an adaptation of these organisms to their environment. Bacteroidetes are involved in food digestion in the intestinal tract of mammals. Hence, a specific TBDT class for the uptake of substrates provided by the host seems plausible.

Region XII Similar to region XI, region XII also shows a great distance to the other regions (Figure 10.4). It contains eight clusters (35, 62, 63, 72, 76, 88, 187 and 188) and only one expTBDT (MalA; sequence 90, cluster 63) that transports maltodextrin. Seven pTBDTs are suggested to transport xylan, pectin or chito-oligosaccharides (No. 91-97; for references see Tab. E.1 in the appendix), where six of them (sequences 91-96) belong to cluster 63 and the remaining pTBDT (sequence 97) to cluster 72. It appears that this region is composed of di- and oligosaccharide transporters. This is in line with this observation, that TBDTs found in species of the order Myxococcales (*Myxococcus xanthus*, *Sorangium cellulosum*), are located in this region. These species are found on decaying plant material consuming their saccharides.
In contrast to the so far described regions, region XII shows a homogenous species composition. Most of the sequences stem from α- and γ-proteobacteria (18.4%, 76%) and a few bacteroidetes, δ- and β-proteobacteria taxa.

Region XIII Positioned between region XI and the crowded area on the right side, this region is defined by a fibronectin-binding TBDT (sequence 98, cluster 41). An interaction of this TBDT with a glycoprotein has been observed. This observation is consistent with the close proximity of regions XI and XIII, because glycoproteins contain oligosaccharides, which in turn are substrates of some of the TBDTs found in region XI.
Similar to region XI, the sequences of this region consist almost exclusively of bacteroidetes.

Other regions For regions I to XIII we could infer at least putative functions for ∼3,700 sequences. However, from the ∼4,600 sequences we identified as putative TBDTs, ∼900 sequences remain in regions A-N in which no TBDT with experimentally confirmed or predicted substrate is located. Consequently, we were unable to assign any function (Figure 10.3). While we cannot discuss potential substrates for clusters in regions A-N, we can at least point to some regions that show a peculiar taxonomic composition. In

regions A and B mostly sequences from γ- (74%) and α-proteobacteria (19%), but also a few β-proteobacterial (5%) and bacteroidetes (1.5%), are present. Region C contains exclusively γ-proteobacterial sequences.

10.4.4 Setup of a TBDT Sequence Database

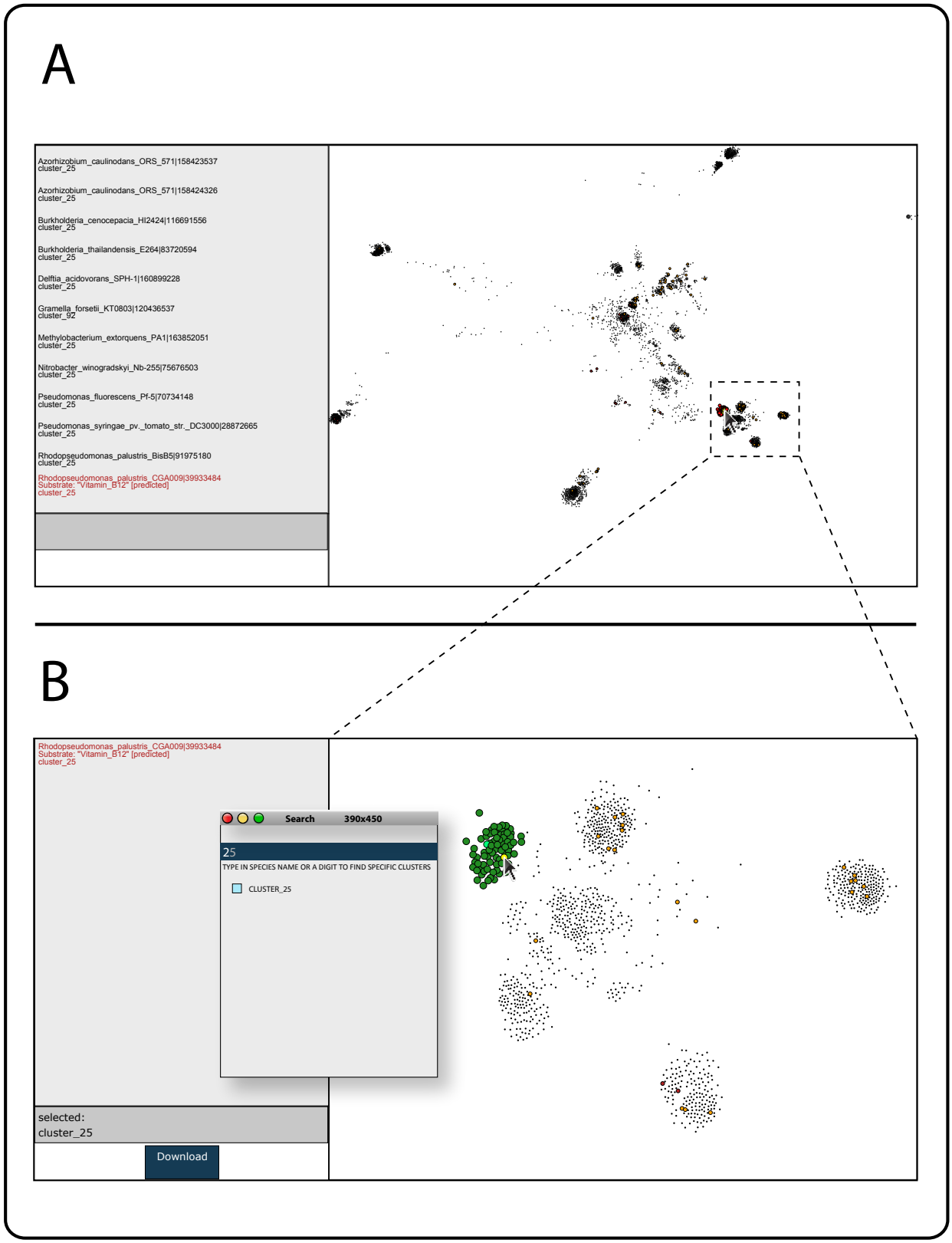
We have setup a sequence database that contains all protein sequences included in our analyses. It is accessible at `http://www.cibiv.at/TBDT`. Instead of providing a traditional text-oriented user interface, we implemented a more intuitive graphical user interface (GUI), called TBDT explorer (Fig. 10.5). For the GUI, we mimicked the graphical representation of the CLANS clustering as shown in Figures 10.3B. The user can navigate through the visualization of the clustering by zooming in and out and moving the viewing area. By that, the clustering can be explored on a fine scale. The viewing area is mouse sensitive. If the user places the mouse pointer on one of the dots representing a TBDT, information for that particular sequence (species name, GenBank ID and annotation) are shown in a dedicated field. Additionally, the ID of the cluster that sequence was assigned to (see Fig. 10.4.1) is displayed (Fig. 10.5A). Simultaneously, the dots representing the remaining sequences of the same cluster are highlighted. Dots that represent TBDTs with experimentally verified or predicted substrates (see Tab. E.1 in the appendix) are marked with different colors.
We provide two methods to obtain the sequence data. First, clicking on one of the dots selects the corresponding cluster and a download buttons appears, which is linked to a multi-fasta file containing all sequences of that specific cluster. Second, we implemented a search function that enables the user to select specific clusters by typing in the cluster ID as used in Figure 10.3A (Fig. 10.5B). Alternatively, the user can type in a species name, which results in the selection of all TBDTs from that species.
The GUI was programmed with Processing[5], an extension of the well-known programming language Java[6]. It will run within any modern web browser that supports Java, independent of the operating system of the user. Processing was especially designed to develop applications with focus on graphical output. This allows us to render the TBDT explorer with up to 15 frames per second which makes it very responsive to the actions of the user.

[5] `http://www.processing.org`
[6] `www.java.com`

10.4.5 Classification of TonB-dependent Transporters in Cyanobacteria

One of our aims was the identification and classification of cyanobacterial TBDTs. Hence we searched for sequences of putative TBDTs in 32 cyanobacterial genomes (proteins listed according to their accession code (Tab. 10.1, column 1). We additionally extracted the automated annotation from GenBank (Tab. 10.1, column 2). At present, this annotation is mostly limited to CirA, FhuE or BtuB. Hence, we analyzed the location of cyanobacterial sequences on the CLANS plot (Figure 10.6 shows the section of Figure 10.3B indicated by a grey box). All cyanobacterial TBDTs belong to regions with experimentally characterized TBDTs (see Figure 10.4 and dashed frames in Figure 10.6). To further confirm the classification determined with CLANS, we additionally constructed a phylogenetic tree

Figure 10.5 ***(facing page)*****: TBDT Explorer** These two screenshots illustrate the main features of the TBDT explorer. Each black dot represents a TBDT with yet undetermined function. TBDTs with experimentally confirmed (expTBDTs) or predicted (pTBDTs) substrates (see Tab. E.1 in the appendix) are represented by orange and red dots, respectively. The positions of all dots were extracted from the CLANS output. In **(A)** the total view on all ∼4600 TBDTs is shown. To demonstrate the mouse sensitivity, the mouse pointer was placed on the aggregation of dots in the top left section of the dashed frame. Dots underneath the mouse pointer are highlighted by an increased size and yellow color. Additionally, the species name, GenBank ID and our cluster ID are presented in a dedicated section on the left. The information for expTBDTs or pTBDTs is written in red color and extended by the corresponding substrate. Dots that represent TBDTs belonging to the same cluster as those underneath the mouse pointer are drawn in red. **(B)** shows a zoomed in view on the section marked by the dashed frame in **(A)**. Furthermore, the search window is shown which was used to search and select cluster 25. All members of this cluster are drawn in green with an increased size. expTBDTs and pTBDTs are distinguishable from yet uncharacterized TBDTs by their light green color. By selecting a cluster, a download button appears on the bottom left, which is linked to a sequence file containing all the sequences of that cluster. Due to the enlarged view it is now possible to place the mouse pointer on exactly one dot. Again, this dot is drawn in yellow and additional information about this specific sequence appears in the info area on the left.


A
Azorhizobium_caulinodans_ORS_571|158423537
cluster_25
Azorhizobium_caulinodans_ORS_571|158424326
cluster_25
Burkholderia_cenocepacia_HI2424|116691556
cluster_25
Burkholderia_thailandensis_E264|83720594
cluster_25
Delftia_acidovorans_SPH-1|160899228
cluster_25
Gramella_forsetii_KT0803|120436537
cluster_92
Methylobacterium_extorquens_PA1|163852051
cluster_25
Nitrobacter_winogradskyi_Nb-255|75676503
cluster_25
Pseudomonas_fluorescens_Pf-5|70734148
cluster_25
Pseudomonas_syringae_pv._tomato_str._DC3000|28872665
cluster_25
Rhodopseudomonas_palustris_BisB5|91975180
cluster_25
Rhodopseudomonas_palustris_CGA009|39933484
Substrate: "Vitamin_B12" [predicted]
cluster_25
B
Rhodopseudomonas_palustris_CGA009|39933484
Substrate: "Vitamin_B12" [predicted]
cluster_25
Search 390x450
25
TYPE IN SPECIES NAME OR A DIGIT TO FIND SPECIFIC CLUSTERS
CLUSTER_25
selected:
cluster_25
Download

for the cyanobacterial sequences (Fig. 10.7). The 97 cyanobacterial TBDT sequences were aligned with MAFFT (Katoh *et al.* (2005)) and a maximum likelihood tree was constructed with IQPNNI v3.3.b4 (Minh *et al.* (2005)). As a substitution model we selected VT (Müller and Vingron (2000)) with gamma-distributed substitution rates. Support values were calculated from 1,000 bootstrap replicates. The consensus tree was reconstructed with TREE-PUZZLE v5.2 (Schmidt *et al.* (2002)) applying the majority consensus rule. Seven 'subtrees' (a-f) were identified and mapped to regions I-X.

The six sequences in subtree 'a' belong to region I (Fig. 10.6, Fig. 10.7) and show a relation to heme transporters such as HutA (Figures 10.3, 10.4, sequence 13). The sequences are found in *Synechococcus* sp., *Acaryochloris marina* and *Anabaena* sp. PCC 7120 (see new assignment in Table 10.1, column 3). Subtrees 'b' and 'c' contain only sequences from *Gloeobacter violaceus*. The subtree 'b' resides within region I and is equidistant to enterobactin and heme transporters. Thereby, a clear assignment to a characterized TBDT family is currently impossible. Subtree 'd' is close to the BtuB transporter cluster (region II) (Figure 10.3). In this region we find sequences from most of the analyzed cyanobacteria (8 of 12), suggesting that transporters with similarity to BtuB are common. Subtree 'e' (Figure 10.6, 10.7) represents transporters, which can clearly be assigned as specific for aerobactin/rhizobactin (IutA-/RhtA-type). Subtree 'f' represents sequences of transporters with the closest relation to FhuA-type transporters of cluster 0. The sequences of subtree 'g' (cluster 1), closely related to ViuA, are probably transporters for catecholates. The sequences of subtree 'g' are also close to cluster 118, which contains putative thiamin transporters. Nevertheless, since the two putative thiamin transporters have not yet been experimentally confirmed, we consider these cyanobacterial TBDTs to be iron transporters of the ViuA-type.

To summarize our findings, the assignment of the cyanobacterial TBDTs to regions with functional characterization was successful with the exception of some TBDTs from *Gloeobacter violaceus* (subtrees 'b' and 'c'). Although BtuB-like transporters and hydroxamate-type metallophore transporters were found in cyanobacteria, we did not find FecA-type (diferric dicitrate) TBDTs, even though they occur in α-, β-, γ-, δ-, and ϵ-proteobacteria, bacteroidetes and spirochaetes.

10.4.6 Identification of TBDTs in *Anabaena* sp. PCC 7120

In order to explore the cyanobacterial TBDTs in more detail, we analyzed the full genome of *Anabaena* sp. PCC 7120. We identified 21 TBDT genes carrying the plug domain and β-barrel domain characteristic for TBDTs. In addition, we identified four genes (*all2620*, *alr2179*, *all2578*, *alr4028*) containing the plug domain of the TBDT, but an incomplete β-barrel domain. Downstream of *all2620* (Fig. 10.8A) and also *alr4028* (Fig. 10.8B) a gene coding for the 'missing part' of the β-barrel domain is present (*all2619* and alr4029),

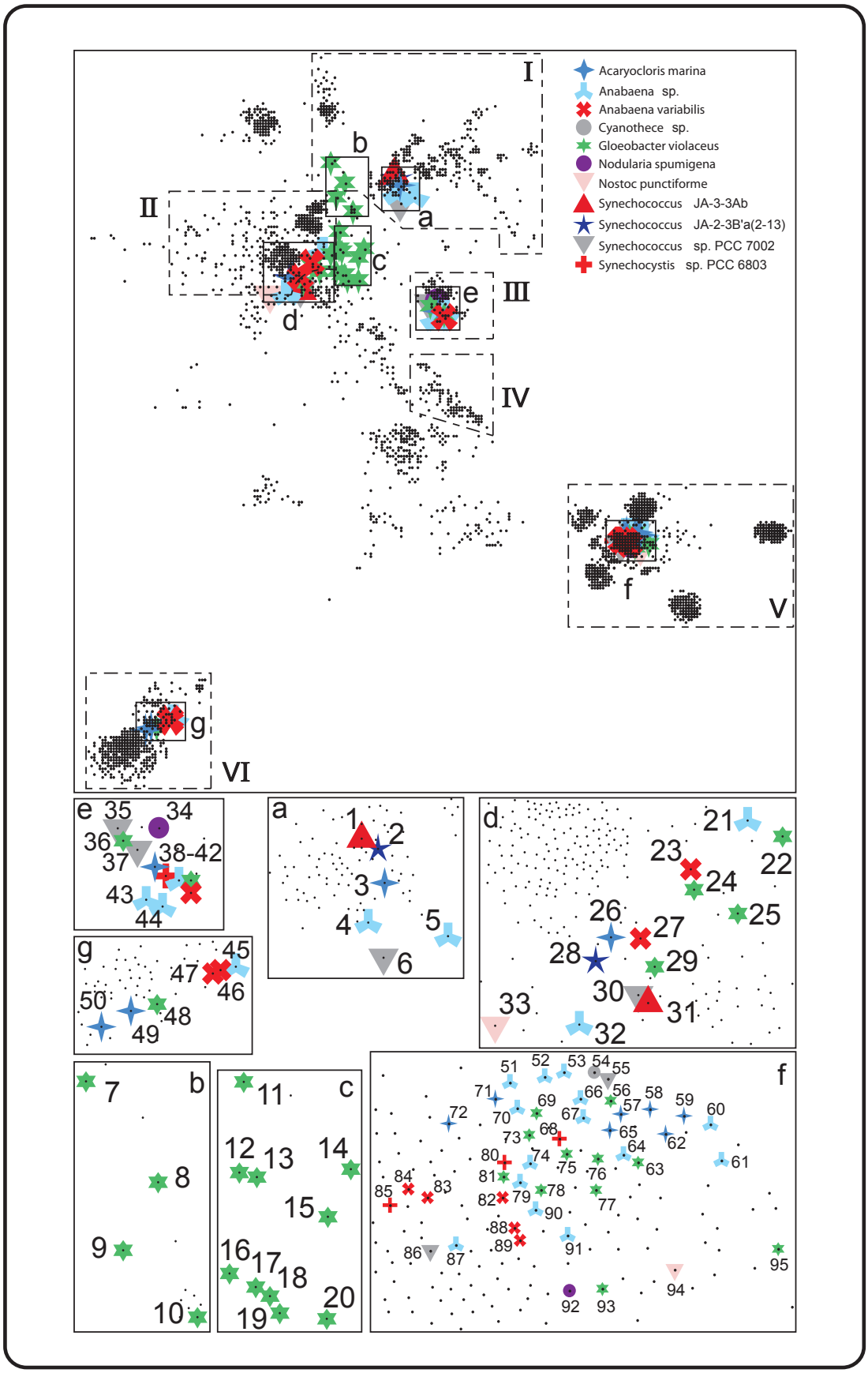


Figure 10.6: Distribution of TBDTs found in genomes of cyanobacteria Cyanobacterial sequences of TBDTs are highlighted and a blow-up of the corresponding frames is shown on the bottom. Dashed boxes correspond to the regions as shown in Fig. 10.3. The color code shows the different species as indicated in the right corner. The numbers are assigned according to Tab. 10.1

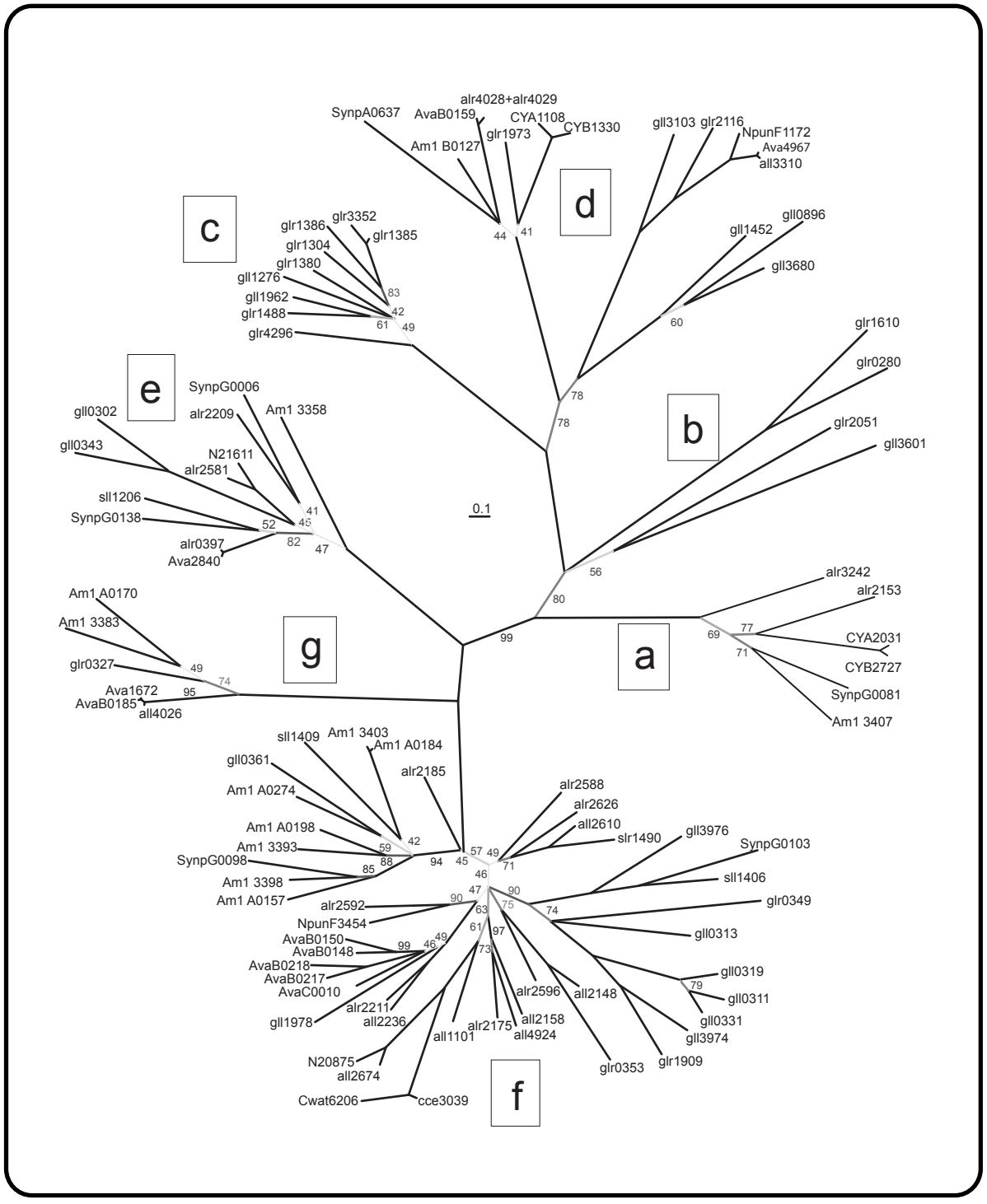


Figure 10.7: Maximum Likelihood Analysis of TBDTs found in Cyanobacteria An alignment of sequences of TonB-dependent transporters listed in Table 10.1 was used to reconstruct a maximum likelihood phylogeny. Support values were calculated from 1,000 bootstrap replicates. To indicate the probability of occurrence of an edge in these trees the edges are shown in shades of grey, with darker lines corresponding to higher support values.

respectively. Consequently, we checked the stop codon separating the two gene pairs. For that, we isolated genomic DNA of *Anabaena* sp. as described in Cai and Wolk (1990). The intergenic sequences between *all2619* and *all2620* and between *alr4028* and alr4029, respectively, and additional ∼250 bp inside each flanking gene were amplified with 5' Prime PCR Extender Polymerase (5' Prime, Hamburg, Germany) according to the manufacturer's protocol. The PCR product was cloned into pCR2.1 (Invitrogen, Karlsruhe, Germany), transformed into DH5α (GibcoBRL, Eggenstein, Deutschland) and the resulting plasmids purified for sequencing.
We confirmed the stop codon between *all2620* and *all2619* (Fig. 10.8A) and could not identify a frame shift in the sequence of the region 500 bp upstream or downstream of the stop codon. If *all2620* is, indeed, part of a TBDT it has to form a heterodimer. A putative interaction partner would be *all2619*. It would, therefore, be interesting to investigate the existence of such complex and to understand whether it is just a remnant of a genetic accident which led to a split of the TBDT gene in *all2620* and *all2619*. In contrast to *all2620* and *all2619*, for *alr4028* and alr4029 we found a T to C exchange in the sequence when comparing our results with that of the deposited sequence. Hence, we conclude that the stop codon does not exist and that the two genes *alr4028* and alr4029 encode one protein. Therefore, 22 TBDTs exist in *Anabaena* sp. PCC 7120.

For 19 TBDTs the genomic organization suggests the integration of the genes in an operon (Figure 10.8C). Twelve TBDTs are directly positioned behind a gene coding for a (putative) transcriptional regulator (Figure 10.8C, dark blue), and most of the (putative) operon structures contain genes coding for proteins involved in iron transport. The gene coding for a ViuA-type transporter is in a putative operon with subunits of a cytochrome D ubiquinol oxidase, which is rather unexpected, because, to date, a relation between this oxidase and iron transport has not been reported (Figure 10.8C). Of the BtuB transporters, one is a single gene (*all3310*), whereas the other (the gene which we confirmed and which is still annotated as *alr4028*/alr4029) is in a rather typical genomic environment, namely in front of three genes encoding the periplasmic and the plasma membrane localized iron transport machinery. The same holds true for the *hutA*-like gene *alr3242*. The other *hutA*-like gene (*alr2153*) is in a putative operon with a gene encoding a tetracenomycin C synthesis protein and a gene of unknown function. An interaction between this TBDT and the downstream genes has not been shown so far.

Three genes are classified as *iutA*-like. *alr0397* (*schT*) is single standing in the genome. Downstream of *alr2581* we found two genes coding for an unknown protein and a dicitrate binding protein, respectively. *Alr2209* is a component of a large genomic region (∼14 kbp alr2208-alr2215) containing upstream a transcription regulator and downstream a cluster with three genes coding for two periplasmic dicitrate binding proteins and one *fhuA*-like gene (*alr2211*). Fourteen of the 22 TBDTs encoding genes in *Anabaena* sp. PCC 7120 show a close relation to FhuA-type transporters (subtree 'f' in Fig. 10.7, see

Tab. 10.1). Thirteen of these genes are upstream of a gene coding for a protein annotated as dicitrate-binding. However, most of the genes found in the putative operons defined by the 14 *fhuA*-like genes encode for proteins of unknown function. Three of the *fhuA*-like genes (*alr2588*, *alr2592*, *alr2596*) are in the same chromosomal region. Upstream of these, a gene coding a transcription regulator and downstream a gene encoding a dicitrate binding protein are found. However, the phylogenetic analysis (Figure 10.7) argues against a recent gene duplication that could have give rise to the multiple *fhuA*-like genes.

10.4.7 Variations of the Number of Genes encoding TBDTs in Cyanobacteria

The number of TBDTs varies among cyanobacteria. We found 22 TBDTs in *Anabaena* sp. PCC 7120, 10 in *Anabaena variabilis*, six in *Synechococcus* sp. PCC 7002, four in *Synechocystis* sp. PCC 6803, 33 in *Gloeobacter violaceus*, but no TBDTs in the genomes of, for example, *Prochlorococcus* (Figures 10.6, 10.7, 10.8 and Table 10.1). This variation of the number of genes, however, does not reflect an elevated amount of outer membrane protein coding genes in *Anabaena* sp. PCC 7120 in general, because other outer membrane proteins (Omp85, TolC, OstA and others) are not found with higher counts in *Anabaena* sp. PCC 7120 than in other cyanobacteria (not shown). Furthermore, in a previous report, a correlation of the number of TBDTs to the number of open reading frames as, for example, for transporters in the cytoplasmic membrane (Ren and Paulsen (2005)) was not be observed (Schauer *et al.* (2008)), which is supported by our analysis (not shown).

TBDTs are regulated by TonB proteins. Hence, the large number of TBDTs leads to the question of whether each TBDT is regulated by its own TonB protein or whether (at least a sub-population of the TBDTs) is regulated in concert by the same TonB protein. We, therefore, screened cyanobacterial genomes for the presence of *tonB* (Table 10.2). One to three *tonB* genes were detected. Hence, the number of TBDTs largely exceeds the number of TonB proteins. Please note that we identified a TonB-like protein (Slr1484) in *Synechocystis* sp. PCC 6803, which corrects a previous statement excluding the presence of a TonB-like protein in this species (Huang *et al.* (2002)).

10.5 Conclusion

A clustering of ∼4,600 TBDTs revealed that these proteins group by their substrate rather than according to their taxonomy. Exceptions of this observation are the regions IX, XI, XIII and C. The latter region comprises sequences from bacteroidetes and γ-proteobacteria, only. Hence, the transported molecule dominates the sequence variation

[6] `http://bacteria.kazusa.or.jp/cyano/`

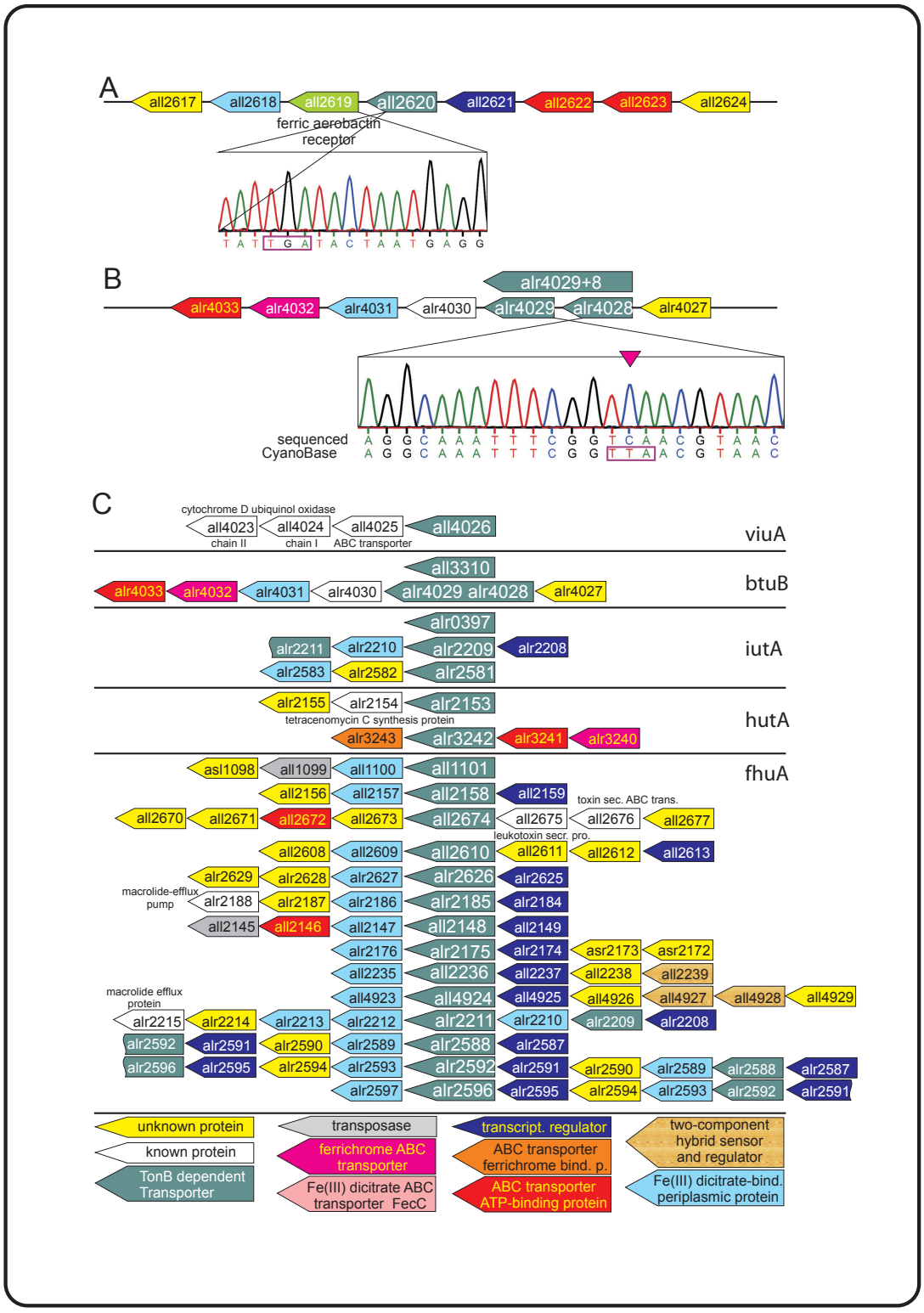


Figure 10.8: The genomic organization of the loci coding for TonB-dependent transporters (TBDTs) in *Anabaena sp.* PCC 7120 For A-C, the genomic structure was excised from Cyanobase[6](Kaneko *et al.* (2001)) and the nomenclature of the color code is given in **(C)**. White boxes indicate genes for which the name is given in the figure. **(A)** shows the genomic orientation surrounding *all2620* (top) and the sequencing profile of the region coding for the stop codon (in reverse orientation, bottom). **(B)** shows the genomic orientation surrounding *alr4028* (top) and the sequencing profile of the region coding for the stop codon (in reverse orientation, bottom). **(C)** The genomic organization surrounding the 22 genes coding for TBDTs.

among TBDTs. Using the occurrence of expTBDTs within clusters, we were able to assign a tentative substrate for almost two-thirds of the analyzed sequences. The exact details of the clustering can be inspected at our website[7] (see section 10.4.4). Here, the individual clusters or sequences can be highlighted based on the presentation in Figure 10.4. However, the current assignment has to be considered preliminary as Schauer and colleagues pointed out that further substrates might be discovered in future (Schauer *et al.* (2008)). We will updated the web interface as soon as data about new substrate is available. We identified several clusters of TBDTs with putatively so far unknown substrates. Further research on a few representative proteins of each from these clusters would be of great interest, as it would significantly advance the knowledge on substrate uptake by bacteria on the protein level. Moreover, it might also reveal new potential drug targets. Many of Gram-negative bacteria are pathogens. A blocking the specific nutrition uptake via TBDTs could potentially reduce the viability of these agents.

We did not observe large differences to previously suggested classifications of iron-transporting TBDTs. Overall, our approach reproduced previous classifications of TBDTs according to their substrates based on a smaller number of sequences and a phylogenetic tree reconstruction (Koebnik *et al.* (1993), LeVier and Guerinot (1996), Rakin *et al.* (1994), Bäumler and Hantke (1992), Stojiljkovic and Hantke (1992)). Only the positioning of the IutA and of the ViuA sequences differs with respect to distances previously proposed (LeVier and Guerinot (1996), Rakin *et al.* (1994)). In contrast to the report by LeVier and Guerinot, who placed ViuA between the lactoferrin and transferrin recognizing transporters (LeVier and Guerinot (1996)), we found that ViuA (sequence 74, Region VI) clearly clusters with FyuA (sequence 72). This discrepancy might reflect the fact that: (i) more sequences of TBDTs are available nowadays; and (ii) the methodology to analyze sequence relationships has improved.

A deviation from this general picture was found for the predicted BtuBs, which are spread over a long stripe from regions II to V. Hence, BtuBs might show a similar diffuse distribution pattern like the heme and hydroxamate transporters (regions I and V, respectively). Candidate TBDTs that clustered with BtuB and we therefore predicted to transport cobalt-complexing vitamin B12 might transport different substrates that are only structurally related.

10.5.1 TBDTs in *Anabaena* sp.

Based on database searches, we have identified 25 proteins in *Anabaena* sp. PCC 7120 that show features characteristic for TBDTs (Lundrigan and Kadner (1986)). Our careful analysis of these 25 sequences suggests that 22 of these 25 are functional TBDTs (see

[7] `http://www.cibiv.at/TBDT`

section 10.4.6, Figs. 10.7 and 10.8). Strikingly, at least five different types of transporters are identified (FhuA, ViuA, IutA, BtuB and HutA type) and this number greatly exceeds the number of genes coding for TBDTs of almost all other (sequenced) cyanobacteria. The only exception is *Gloeobacter violaceus* which, in turn, is the only species for which we could not predict substrates for all TBDTs. As already discussed, it has been hypothesized that iron limitation might have been one of the selective forces in evolution of cyanobacteria. *Gloeobacter violaceus* (33 TBDTs) was isolated in 1972 from a calcareous rock near the Vierwaldstättersee in Switzerland, whereas *Anabaena* variablis (previously *Anabaena flos-aquae* strain A-37; 10 TBDTs) was isolated in 1964 from fresh water of the Mississippi, USA. Both strains are considered non-symbiotic. *G. violaceus* is rather unique with respect to the absence of a thylakoid structure and does not form filaments (Guglielmi *et al.* (1981)), just like *Synechocystis* sp. PCC 6803 (4 TBDTs). The latter was isolated from fresh water in California and deposited in the Pasteur collection[8] in 1968. Hence, the number of TBDTs does not correlate with filament or heterocyst formation. It might, however, correlate with the habitat from which the species were isolated. Relevant parameters might be either the species competition for iron or the iron limitations in the environment *per se*. Therefore, symbiotic cyanobacteria such as *Nostoc punctiforme* may possibly contain a rather low number of TBDTs because iron is provided by the host. Unfortunately, to the best of our knowledge, the source of *Anabaena* sp. strain PCC 7120 (formerly named *Nostoc muscorum* ISU (Adolph and Haselkorn (1971)); further synonyms are *Anabaena* sp. ATCC 27893, *Nostoc* sp. strain PCC 7120) is unknown, and it is considered to be a 'free living cyanobacterium'. The observation that this cyanobacterium is susceptible to viruses isolated from the Lake Mendota, Dane County, Wisconsin, USA, (Adolph and Haselkorn (1971)) might suggest that a similar environment was the place of isolation. This would be in line with an original natural habitat of *Anabaena* sp. PCC 7120 that contained rather limited iron sources. It has been reported that the iron concentration in rivers is higher than in lakes (e.g. Moslavac *et al.* (2005)). The variety of TBDT classes found in *Anabaena* sp. rather agrees with iron limited environmental conditions.

The only TBDT type which could not be identified in any of the analyzed cyanobacterial species, is the FecA-type (diferric dicitrate), which, however, can be found in many other bacteria. To date, schizokinen is the only confirmed siderophore which is secreted by *Anabaena* sp. PCC 7120 (Goldman *et al.* (1983)) and, recently, its transporter was identified (Nicolaisen *et al.* (2008)). However, additional siderophores are secreted by Anabaena sp. (Nicolaisen *et al.* (2008), Jeanjean *et al.* (2008)), but they have not yet been characterized. Nevertheless, other interpretations for the variable number of TBDTs might still be possible.

[8] `http://www.crbip.pasteur.fr`

Table 10.1: Sequences used for the phylogenetic analysis of cyanobacteria The gene name is given in column 1; the initial annotation in the database in column 2; the classification according to Figure 10.6 using the name of a representative transporter of the related category is given in column 3; the spot number according to Figure 10.6 is given in column 4; the accession code in column 5 and the source organism in column 6.

Code	Old	New	Spot	GenBank ID	Species
Am1 B0127	BtuB	BtuB	26	gi|158339996	*Acaryochloris marina MBIC11017*
Am1 3407	CirA	HutA	3	gi|158336543	
Am1 3358	CirA	IutA	38	gi|158336494	
Am1 A0170	CirA	ViuA	50	gi|158339820	
Am1 3383	CirA	ViuA	49	gi|158336519	
Am1 A0184	Fiu	FhuA	58	gi|158339834	
Am1 A0274	FhuE	FhuA	57	gi|158339535	
Am1 A0198	-	FhuA	71	gi|158339845	
Am1 A0157	FhuE	FhuA	59	gi|158339807	
Am1 3393	FhuE	FhuA	72	gi|158336529	
Am1 3398	-	FhuA	65	gi|158336534	
Am1 3403	Fiu	FhuA	62	gi|158336539	
Ava_B0148	FhuE	FhuA	89	gi|75812430	*Anabaena variabilis* ATCC 29413
Ava_B0150	CirA	FhuA	88	gi|75812432	
Ava_B0159	BtuB	FhuA	82	gi|75812441	
Ava_B0185	CirA	ViuA	46	gi|75812467	
Ava_B0217	Fiu	BtuB	23	gi|75812499	
Ava_B0218	CirA	IutA	42	gi|75812500	
Ava_C0010	FhuE	BtuB	27	gi|75812671	
Ava_1672	CirA	ViuA	47	gi|75907894	
Ava_2840	CirA	FhuA	83	gi|75909052	
Ava_4967	BtuB	FhuA	84	gi|75911163	
cce3039	CirA	FhuA	54	gi|172037952	*Cyanothece* sp. CCY0110
glr0280	CirA	?[a]	7	gi|37519849	*Gloeobacter violaceus* PCC7421
gll0302	CirA	IutA	41	gi|37519871	
gll0311	-	FhuA	81	gi|37519880	
gll0313	-	FhuA	56	gi|37519882	
gll0319	-	FhuA	63	gi|37519888	
glr0327	CirA	ViuA	48	gi|37519896	
gll0331	FecA	FhuA	69	gi|37519900	
gll0343	CirA	IutA	36	gi|37519912	
glr0349	FhuE	FhuA	95	gi|37519918	
glr0353	-	FhuA	76	gi|37519922	
gll0361	FhuE	FhuA	78	gi|37519930	
gll0896	FepA	BtuB	22	gi|37520465	
gll1276	OMC[b]	BtuB	20	gi|37520845	
glr1304	OMC[b]	BtuB	17	gi|37520873	
glr1380	-	-	-	gi|37520949	
glr1385	-	BtuB	18	gi|37520954	
glr1386	CirA	BtuB	15	gi|37520955	
gll1452	BtuB	BtuB	25	gi|37521021	

[a] no clear assignment possible
[b] OMC: outer membrane channel

Code	Old	New	Spot	GenBank ID	Species
glr1488	-	BtuB	12	gi|37521057	
glr1610	-	?[a]	8	gi|37521179	
glr1909	FhuE	FhuA	73	gi|37521478	
gll1962	-	BtuB	16	gi|37521531	
glr1973	BtuB	BtuB	29	gi|37521542	
gll1978	Fiu	FhuA	93	gi|37521547	
glr2051	FepA	?[a]	10	gi|37521620	
glr2116	BtuB	BtuB	24	gi|37521685	
gll3103	BtuB	BtuB	11	gi|37522672	
glr3352	OMC[b]	BtuB	19	gi|37522921	
gll3601	BtuB	BtuB	14	gi|37523170	
gll3680	BtuB	?[a]	9	gi|37523249	
gll3974	FhuE	FhuA	77	gi|37523543	
gll3976	CirA	FhuA	75	gi|37523545	
glr4296	CirA	BtuB	13	gi|37523865	
NpunF1172	-	BtuB	33	gi|186681644	*Nostoc punctiforme* PCC73102
NpunF3454	-	FhuA	94	gi|186683610	
all1101	-	FhuA	52	gi|17228596	*Anabaena* sp. PCC 7120
all2148	FhuE	FhuA	66	gi|17229640	
all2158	FhuE	FhuA	53	gi|17229650	
all2236	Fiu	FhuA	87	gi|17229728	
all2610	CirA	FhuA	67	gi|17230102	
all2674	Fiu	FhuA	70	gi|17230166	
all3310	BtuB	BtuB	21	gi|17230802	
all4026	CirA	ViuA	45	gi|17231518	
all4924	FhuE	FhuA	61	gi|17232416	
alr0397	CirA	IutA	40	gi|17227893	
alr2153	CirA	HutA	4	gi|17229645	
alr2175	-	FhuA	69	gi|17229667	
alr2185	Fiu	FhuA	51	gi|17229677	
alr2209	CirA	IutA	44	gi|17229701	
alr2211	CirA	FhuA	90	gi|17229703	
alr2581	CirA	IutA	43	gi|17230073	
alr2588	CirA	FhuA	74	gi|17230080	
alr2592	FhuE	FhuA	91	gi|17230084	
alr2596	FhuE	FhuA	64	gi|17230088	
alr2626	-	FhuA	79	gi|17230118	
alr3242	CirA	HutA	5	gi|17230734	
alr4028 +alr4029	-	BtuB	32	gi|17231520 gi|17231521	
sll1206	IutA	IutA	39	gi|16329186	Synechocystis sp. PCC6803
sll1409	FhuA	FhuA	80	gi|16329191	
sll1406	FhuA	FhuA	85	gi|16329194	
slr1490	FhuA	FhuA	68	gi|16329201	
CYA_1108	BtuB	BtuB	31	gi|86605797	*Synechococcus sp.* JA-3-3Ab
CYA_2031	CirA	HutA	1	gi|86606671	
CYB_1330	BtuB	BtuB	28	gi|86608804	*Synechococcus sp.* JA-2-3B'a(2-13)
CYB_2727	CirA	HutA	2	gi|86610153	
SynpA0637	BtuB	BtuB	30	gi|170077260	*Synechococcus sp.* PCC 7002
SynpG0081	CirA	HutA	6	gi|170076551	
SynpG0006	CirA	IutA	35	gi|170076476	
SynpG0138	CirA	IutA	37	gi|170076608	
SynPG0089	FhuE	FhuA	55	gi|170076568	
SynpG0103	FhuA	FhuA	86	gi|170076573	
Nspu20875	CirA	FhuA	92	gi|119508873	*Nodularia spumigena* CCY9414

Code	Old	New	Spot	GenBank ID	Species
Nspu21611	CirA	IutA	43	gi|119509643	
Cwat6206[c]	-	FhuA	1	gi|67920343	*Crocosphaera watsonii* WH8501

[c]incomplete sequence

Table 10.2: TonB-like genes in cyanobacteria Column 1 gives the species name; column 2 gives the number of Tonb-like genes identified in the genome; column 3 gives the locus tag for each detected TonB-like gene.

species	No. TonBs	Locus tag
Acaryochloris marina MBIC11017	2	AM1_A0167, AM1_3413
Anabaena variabilis ATCC 29413	1	Ava_2295
Crocosphaera watsonii WH 8501	1	CwatDRAFT_6356
Cyanothece sp. CCY0110	2	CY0110_08196, CY0110_24616
Gloeobacter violaceus PCC 7421	3	glr1389, glr1815, glr2404
Nodularia spumigena CCY9414	1	N9414_10453
Nostoc punctiforme PCC 73102	1	Npun_F0783
Anabaena sp. PCC 7120	1	all5036
Synechococcus sp. PCC 7002	2	SYNPCC7002_G0090, SYNPCC7002_A2465
Synechococcus sp. JA-3-3Ab	1	CYA_2030
Synechococcus sp. JA-2-3B'a(2-13)	1	CYB_2726
Synechocystis sp. PCC 6803	1	slr1484

Appendices

A Abbreviations

aa – Amino Acids
A – Adenine
AU – Approximately Unbiased Test
ASL – Average Sequence Length
BI – Bayesian Inference
bp – Base Pairs
C – Cytosine
cDNA – complementary DNA
DMP – Deep Metazoan Phylogeny
ddNTP – dideoxynucleotide-tri-phosphate
dNTP – deoxynucleotide-tri-phosphate
EST – Expressed Sequence Tag
G – Guanine
KH – Kishino-Hasegawa Test
LBA – Long Branch Attraction
LM – Likelihood Mapping
MCS – Multiple Cloning Site
ML – Maximum Likelihood
MLd – Maximum Likelihood Distance
NCBI – National Center of Biotechnology Information
pHMM – Profile Hidden Markov Model
RDMS – Relational Database Management Systems
SH – Shimodaira-Hasegawa Test
SQL – Structured Query Language
T – Thymine
TBDT – TonB-dependent transporter
expTBDT – experimentally characterized TBDT
pTBDT – TBDT with predicted substrate
TCs – Tentative Consensus Sequences
WKH – Weighted Kishino-Hasegawa Test

B EST Processing Pipeline

B.1 Overview on the PERL Modules of the Processing Pipeline

In the following, we give a brief description of the modules that are part of the processing pipeline. A graphical overview for the cleaning, the clustering and the annotation section are shown in Figures B.2, B.3 and B.4, respectively.

B.1.1 Cleaning

The pipeline (ESTcc.pl) is started with our internal EST project id. First, the module `get_taxon_info.pm` requests species information for the specific EST project from the table TAXON (see Section B.2.1). The next module, `getFromDB.pm`, downloads the sequences and base quality values of the project and generates a sequence file in fasta format. If vector screening with Lucy has been chosen by the user via program options, the module `lucy.pm` is called next. It will start Lucy with the right parameter settings, parse the results and clip the sequences accordingly. Afterwards, the module `run_crossmatch.pm` will start Crossmatch, that compares all sequences to the vector database UniVec and scans for contaminations witch genomic DNA from *E. coli*. Since Crossmatch only reports contaminations, a module called `vector.pm`, subsequently applies the clipping information generated by Crossmatch to the sequence and the quality value files. The module `seqclean.pm` controls the programs SeqClean and cln2qual. The first program performs an additional scan for vector contaminations using the UniVec database, but also evaluates the general sequence quality. The sequences are clipped and the clipping information is passed to cln2qual, which removes the quality values of clipped nucleotides from the quality files. Masking of repetitive elements and low complexity regions is controlled by the module `mask.pm`. It starts RepeatMasker and collects the results. This module also takes care of sorting out too short sequences. `write2db.pm` uploads all the processed sequences and quality values into the EST table in dbDMP. Finally, a module called `delete.pm` deletes all unnecessary files, generated by the programs called in this section of the pipeline, keeping only the sequence and log files.

B.1.2 Clustering

Processed EST sequences are downloaded for the clustering from the table EST via `getfromDB.pm`. For convenience, the ESTs can also be transfered directly to the clustering routine, if they are already stored in a local sequence file, for example as output of the cleaning procedure.
The module `clustering.pm` starts the TGICL program package and collects the results, which are then passed to `extract.pm` to assemble a sequence file containing the consensus sequences of all contigs. If demanded by the user, the module `trim_quals.pm` starts LUCY to scan for low quality regions within the EST contigs. The resulting clipping information is applied by the module to cut the contig sequences accordingly. Depending on whether a second clustering step is performed or not, the sequences are either directly passed to `write2db` which uploads the contigs into the tables EST_LIST, CONTIG and CONTIG_INFO, or the program flow continues with the module `delete.pm`. The latter will first delete files generated by TGICL, before the module `clustering.pm` is called again, to start the second clustering step with the clipped contig sequences. `extract.pm` afterwards extracts the sequences and passes them to `write2db` which uploads them into dbDMP.

B.1.3 Annotation

`getFromDB.pm` queries the DMP database for contigs associated with the defined clustering project (4.3.2). A Fasta file is compiled which is then transferred to `BlastX.pm`. It will start a BLASTX search with these contigs against a user-defined protein database. The resulting BLAST report is parsed and the information is stored in the database tables BLAST and BLAST_RESULT. For each EST contig, the protein sequences of the up to 25 best BLAST hits are extracted from the protein database and passed to the next module `annotate.pm`. This module starts the program GENEWISE, that calculates a codon alignment between the contig and each of its BLAST hit sequences. Subsequently, the protein sequences are ordered by their GENEWISE score and the description of the three highest scoring sequences are uploaded as tentative annotations into the table ANNOTATION. The translated contig sequence is inserted into the TRANSLATION table.

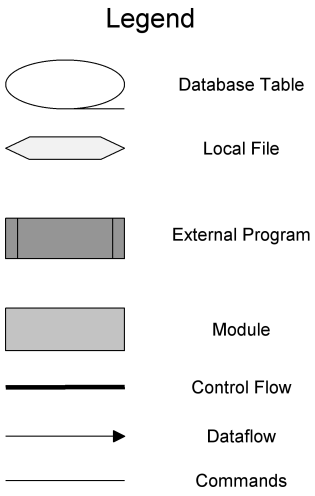


Figure B.1: Legend for following pipeline diagrams These symbols are used in the following diagrams (B.2,B.3 and B.4) showing the individual PERL modules of the EST processing pipeline.

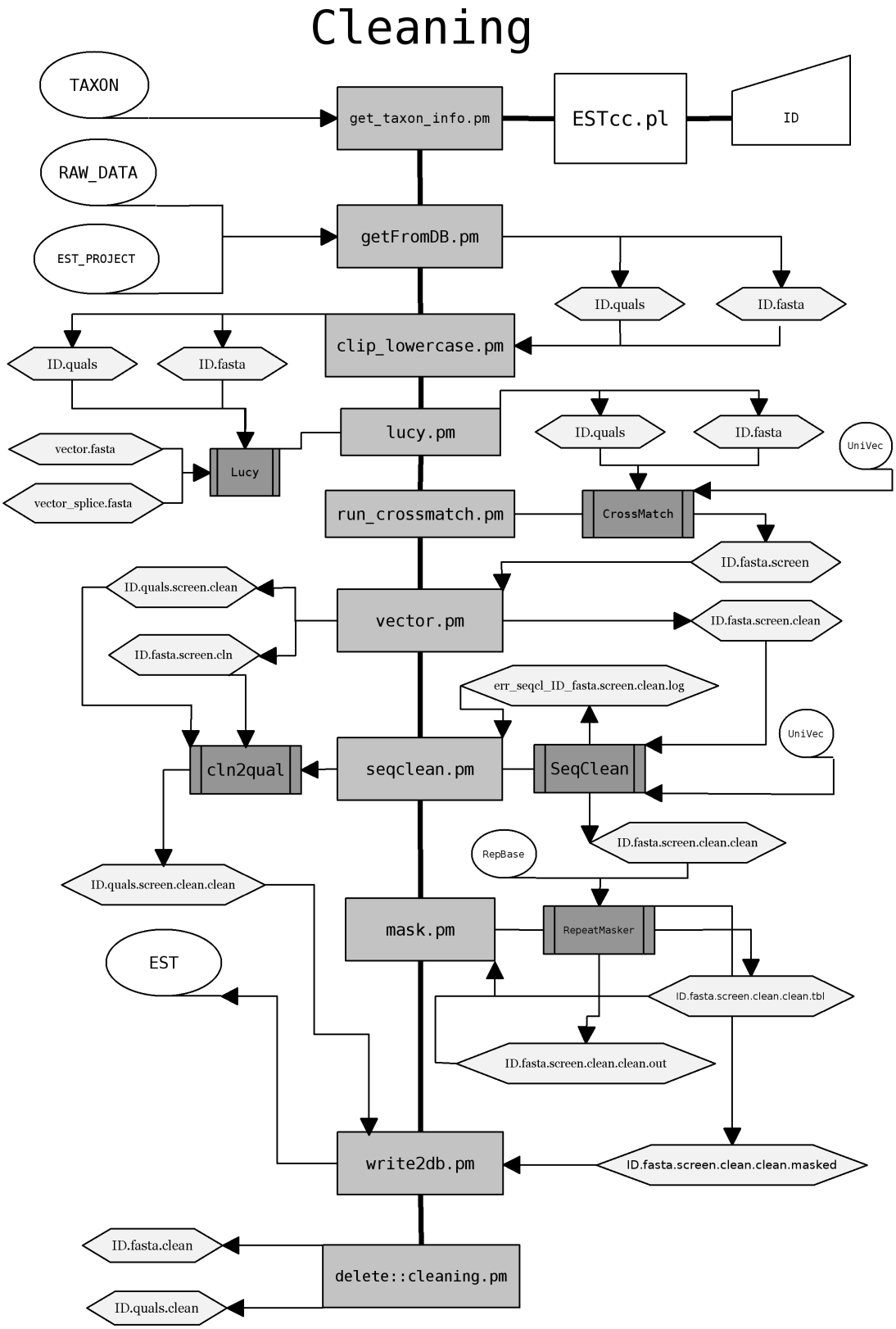


Figure B.2: Cleaning Diagram Program flow of the cleaning section of the processing pipeline. The key to the different box shapes is given in the legend in Fig. B.1. For a description of the modules see Section B.1.1. For a description of the involved database tables see Section B.2.1 and Fig. B.5.

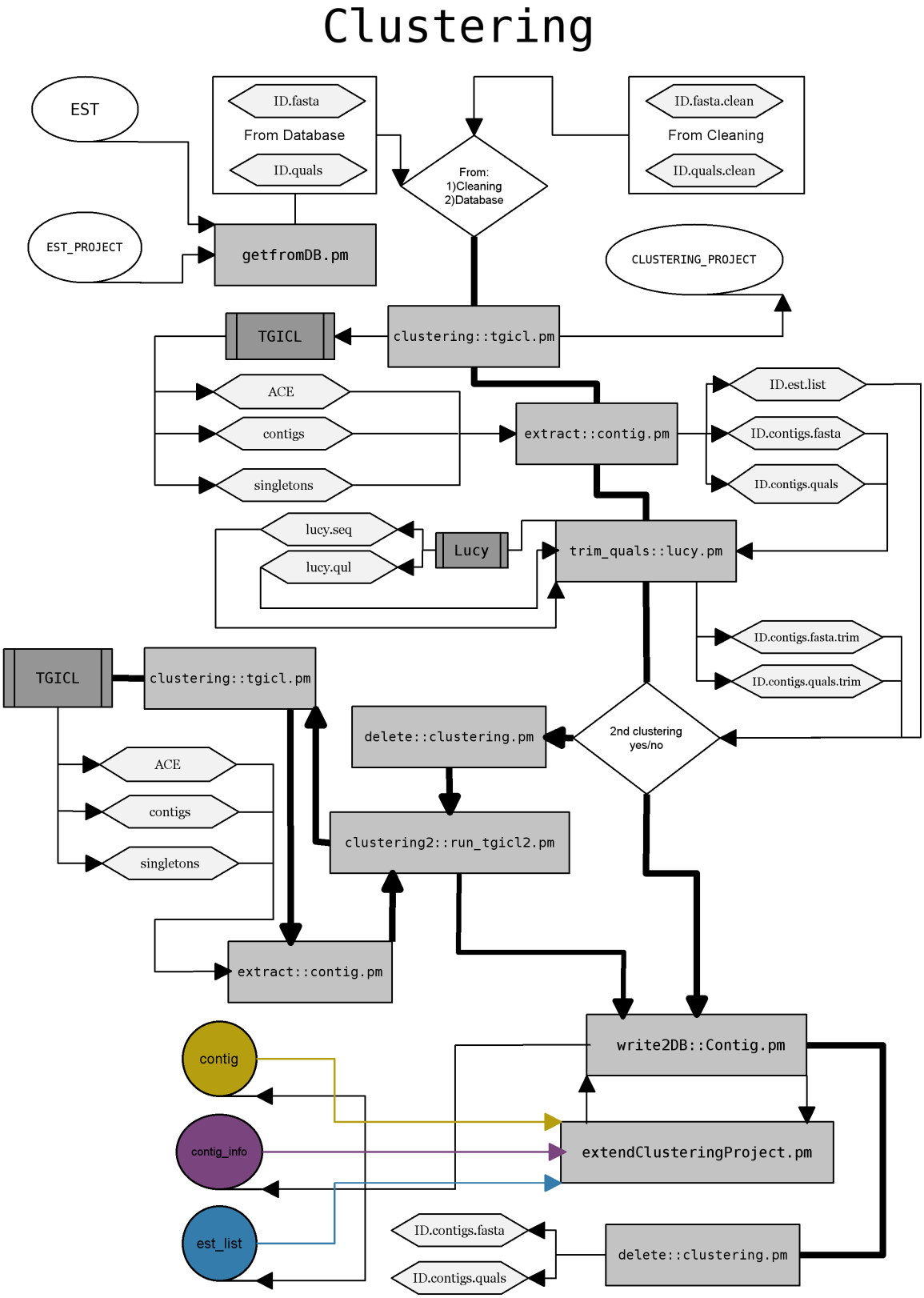


Figure B.3: Clustering Diagram Program flow of the clustering section of the processing pipeline. The key to the different box shapes is given in the legend in Fig. B.1. The colors have been added to make the tracks of the data flow more traceable. For a description of the modules see Section B.1.2. For a description of the involved database tables see Section B.2.1 and Fig. B.5.

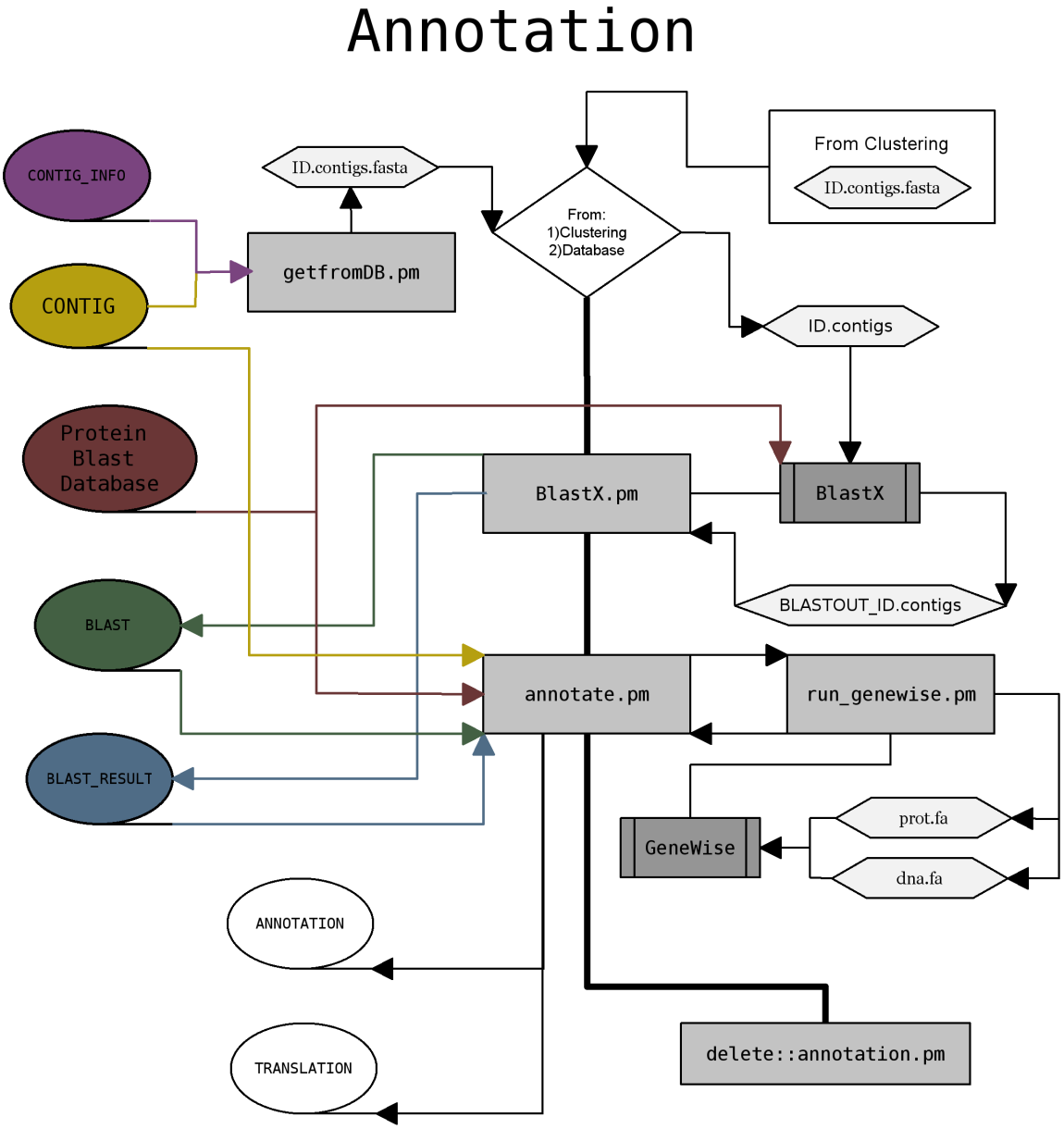


Figure B.4: Annotation Diagram Program flow of the annotation section of the processing pipeline. The key to the different box shapes is given in the legend in Fig. B.1. The colors have been added to make the tracks of the data flow more traceable. For a description of the modules see Section B.1.3. For a description of the involved database tables see Section B.2.1 and Fig. B.5.

B.2 Database

B.2.1 Description of the dbDMP Database Scheme

In the following, we will give a detailed description of all database tables of the DMP project, that are relevant for data processing and managing. A graphical overview is shown in Fig. B.5.
Within the DMP database (dbDMP), ESTs are organized in EST projects. Each project has a unique ID. The table EST_PROJECT contains basic information about each project, such as data source and a link to the table TAXON. This table contains scientific and common name, the systematic description and if present, the NCBI taxonomy ID for each taxon represented by sequence data in dbDMP. Unprocessed ESTs are stored in the table RAW_DATA. Other EST related tables are EST_INFO and EST. EST_INFO contains cross-references between the entries of the dbDMP tables EST, RAW_DATA and EST_PROJECT and NCBI's dbEST. Table EST contains the sequence and base quality values of the cleaned sequences. Finally, CLEANING_PROCEDURE keeps detailed information about the applied cleaning methods, such as program parameters, version of the used vector and repetitive elements databases. Each clustering results in a clustering project, details about which are stored in CLUSTERING_PROJECT. Keeping EST projects and clustering projects separated allows a more flexible handling of the data. For example, it might be desirable to cluster different EST projects of the same species, e.g., to combine public data with private data. But at the same time a clustering without private data has to be held ready for those users who are granted access only to public data. EST_LIST provides the information about which ESTs have been assembled to which contig. COUNT_INFO is used to keep track of the number of raw, processed and assembled sequence for each clustering project. Contig related information is divided into two tables. The sequence itself with base quality values for each contig are stored in CONTIG. CONTIG_INFO contains cross-references for each contig, which assigns each entry in CONTIG to one or more entries in CLUSTERING_PROJECT. This design prevents redundancies in the database. As described above, each clustering initiates a new clustering project. If a contig is not altered due to the addition of new ESTs, a new entry in CONTIG_INFO is created linking to an existing entry in CONTIG, containing the actual sequence.
Attached to CONTIG_INFO are those tables storing information about annotation of the contigs. BLAST contains the parameters of the BLAST search. The up to 25 best BLAST hits per contig are stored in BLAST_RESULTS. TRANSLATION contains the results of the GeneWise alignment of each hit with the contig, including the translation of each contig into an amino acid sequence. The purpose of ANNOTATION_NEU is to archive the protein sequences that resulted in the up to three highest scoring GeneWise alignments per contig, which are used as tentative annotations.

The access of data is controlled with two tables. In EST_ZUGRIFF is noted, which group of users present in PEOPLE (defined by GROUPS) can access which EST project. Analogously, CLUSTERING_PROJECT_ZUGRIFF manages access of clustering projects for specific groups.

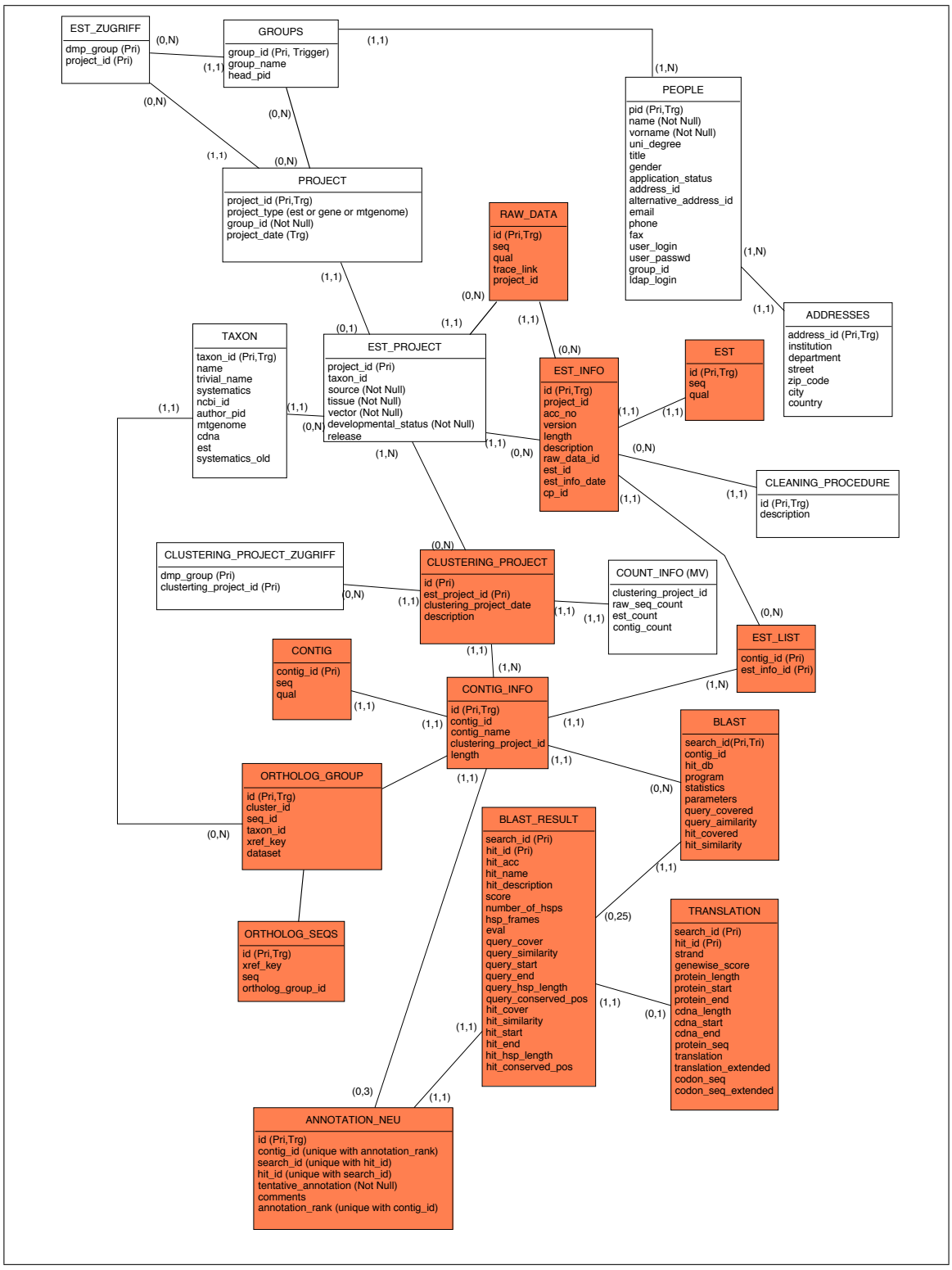


Figure B.5: Database Scheme of dbDMP This diagram includes all database tables of the DMP database. Each box represents a database table. The table name is given above, the column names below the horizontal line. Cross-linkage between tables is illustrated by lines connecting the boxes. The relationship of two related tables, their *cardinality*, is written in brackets at each line. For example, the table EST_INFO is connected to EST_PROJECT. Each entry in EST_PROJECT can be linked to zero or many entries in EST_INFO (0,N). But every entry in EST_INFO is linked to exactly one entry in EST_PROJECT (1,1). Orange boxes mark tables with restricted access.

C Pterygota Phylogeny

Table C.1: Taxa List List of selected taxa for data set creation. Putative orthologs were identified using the HaMStR. An asterisk inidcates represented taxa in the final *maxspe* and *maxgen* data set respectively

Taxon	group	Data source	No. of contigs	No. of orthologs
Agrotis segetum	Lepidoptera	dbEST	812	158
Anopheles albimanus	Diptera	dbEST	3	57
Anopheles anthropophagus	Diptera	dbEST	141	5
*Anopheles gambiae**	Diptera	Ensembl	13	3,096
*Antheraea mylitta**	Lepidoptera	dbEST	1	193
*Aphis gossypii**	Hemiptera	dbEST	4	550
*Apis mellifera**	Hymenoptera	Ensembl	25	3,096
*Baetis sp.**	Ephemeroptera	this study	3	436
Biphyllus lunatus	Coleoptera	dbEST	260	72
Blatella germanica	Dictyoptera	dbEST	2	191
*Bombyx mori**	Lepidoptera	dbEST	40	1,490
*Danaus plexippus**	Lepidoptera	dbEST	10	1,178
Diaprepes abbreviatus	Coleoptera	dbEST	2	143
Euclidia glyphica	Lepidoptera	dbEST	187	28
Folsomia candida	Collembola	dbEST	6	360
*Ischnura elegans**	Odonata	this study	3	527
*Laupala kohalensis**	Ensifera	dbEST	8	700
*Maconellicoccus hirsutus**	Hemiptera	dbEST	4	631
Meladema coriacea	Coleoptera	dbEST	328	60
Melipona quadrifasciata	Hymenoptera	dbEST	321	3
*Nasonia giraulti**	Hymenoptera	dbEST	7	477
*Onychiurus arcticus**	Collembola	dbEST	10	755
*Plodia interpunctella**	Lepidoptera	dbEST	4	431
*Plutella xylostella**	Lepidoptera	dbEST	1	161
Tenebrio molitor	Coleoptera	dbEST	100	12
*Tribolium castaneum**	Coleoptera	dbEST	18	1,164
Tricholepisma aurea	Thysanura	dbEST	344	85
Vespula squamosa	Hymenoptera	dbEST	1	165

Table C.2: ID – the numerical identifier assigned to the gene during the HaMStR process, FlyBaseID/gene name/symbol – the corresponding ID number/gene name/symbol of the Drosophila melanogaster genome database (`http://flybase.org/`). The molecular function and biological process involved description is based on the FlyBase Gene Reports. maxspe/maxgen – genes represented in the alignments. These genes were also selected for the extended ML analyses of individual alignments.

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
6621	FBgn0002031	lethal (2) 37Cc	Dmel\l(2)37Cc	unknown	unknown	+	
6637	FBgn0014028	Succinate dehydrogenase B	Dmel\SdhB	electron carrier	unknown	+	
6639	FBgn0004907	14-3-3*zeta*	Dmel\14-3-3*zeta*	diacylglycerol-activated phospholipid-dependent protein kinase C inhibitor activity	Ras protein signal transduction		+
6671	FBgn0029897	Ribosomal protein L17	Dmel\RpL17	structural constituent of ribosome	translation	+	+
6692	FBgn0032290	CG6443	Dmel\CG6443	unknown	unknown	+	
6715	FBgn0024558	Diphthamide methyltransferas	Dmel\Dph5	enzyme	unknown	+	
6716	FBgn0024733	Qm	Dmel\Qm	structural constituent of ribosome	translation		+
6754	FBgn0000559	Elongation factor 2b	Dmel\Ef2b	GTPbinding	unknown		+
6790	FBgn0001942	Eukaryotic initiation factor 4a	Dmel\eIF-4a	RNA helicase activity	dorsal/ventral axis specification	+	+
6841	FBgn0028336	lethal (1) G0255	Dmel\l(1)G0255	fumarate hydratase activity	tricarboxylic acid cycle	+	
6898	FBgn0033699	Ribosomal protein S11	Dmel\RpS11	structural constituent of ribosome	translation		+
6906	FBgn0034967	eIF-5A	Dmel\eIF-5A	translation initiation factor activity	translational initiation	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
6910	FBgn0014857	Histone H3.3A	Dmel\His3.3A	DNA binding	cell adhesion	+	+
6913	FBgn0039757	Ribosomal protein S7	Dmel\RpS7	structural constituent of ribosome	translation	+	+
6926	FBgn0035753	Ribosomal protein L18	Dmel\RpL18	structural constituent of ribosome	translation		+
6927	FBgn0037351	Ribosomal protein L13A	Dmel\RpL13A	structural constituent of ribosome	translation	+	+
6935	FBgn0010411	Ribosomal protein S18	Dmel\RpS18	structural constituent of ribosome	translation		+
6936	FBgn0038166	CG9588	Dmel\CG9588	ptotein binding	proteolysis	+	
6951	FBgn0010612	lethal (2) 06225	Dmel\l(2)06225	hydrogen-exporting ATPase activity	proton transport		+
6958	FBgn0036580	PDCD-5	Dmel\PDCD-5	DNA binding	apoptosis	+	+
6972	FBgn0021906	Rieske iron-sulfur protein	Dmel\RFeSP	ubiquinol-cytochrome-c reductase activity	mitochondrial electron transport, ubiquinol to cytochrome c	+	+
6978	FBgn0086710	Ribosomal protein L30	Dmel\RpL30	structural constituent of ribosome	translation		+
6983	FBgn0013954	FK506-binding protein 2	Dmel\FK506-bp2	FK506 binding	protein folding		+
6987	FBgn0034242	CG14480	Dmel\CG14480	unknown	unknown	+	
6990	FBgn0037686	Ribosomal protein L34b	Dmel\RpL34b	structural constituent of ribosome	translation		+
6999	FBgn0015393	hoi-polloi	Dmel\hoip	mRNA binding	nuclear mRNA splicing, via spliceosome	+	
7007	FBgn0010265	Ribosomal protein S13	Dmel\RpS13	structural constituent of ribosome	translation	+	+
7013	FBgn0020255	ran	Dmel\ran	GTP binding	actin filament organization	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7015	FBgn0086254	CG6084	Dmel\CG6084	aldehyde reductase activity	unknown	+	+
7021	FBgn0052687	CG32687	Dmel\CG32687	protein binding	unknown	+	
7065	FBgn0086656	shrub	Dmel\shrb	unknown	dendrite morphogenesis; protein	+	
7083	FBgn0040284	SF2	Dmel\SF2	mRNA binding	nuclear mRNA splicing, via spliceosome transport	+	
7089	FBgn0002607	Ribosomal protein L19	Dmel\RpL19	structural constituent of ribosome	translation		+
7098	FBgn0036213	Ribosomal protein L10Ab	Dmel\RpL10Ab	structural constituent of ribosome	translation	+	+
7171	FBgn0020618	Receptor of activated protein kinase	Dmel\Rack1	protein kinase C binding	oviposition	+	+
7181	FBgn0031148	CG1753	Dmel\CG1753	cystathionine beta-synthase activity	cysteine biosynthetic process via cystathionine	+	
7185	FBgn0039537	CG5590	Dmel\CG5590	oxidoreductase activity	metabolic proces	+	
7188	FBgn0001961	Suppressor of profilin 2	Dmel\Sop2	actin binding	anatomical structure development	+	
7189	FBgn0029176	Ef1γ	Dmel\Ef1γ	translation elongation factor activity	translational elongation	+	+
7214	FBgn0000116	Arginine kinase	Dmel\Argk	arginine kinase activity	phosphorylation	+	+
7236	FBgn0014455	Adenosyl-homot-cystein-ase at 13	Dmel\Ahcy13	adenosyl-homo-cysteinase activity	one-carbon compound metabolic process		+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7307	FBgn0023514	CG14805	Dmel\CG14805	unknown	unknown	+	
7310	FBgn0016691	Oligomycin sensitivity-conferring pr	Dmel\Oscp	hydrogen-exporting ATPase activity	proton transport		+
7316	FBgn0034743	Ribosomal protein S16	Dmel\RpS16	structural constituent of ribosome	translation	+	+
7318	FBgn0010078	Ribosomal protein L23	Dmel\RpL23	structural constituent of ribosome	translation		+
7321	FBgn0030968	CG7322	Dmel\CG7322	oxidoreductase activity	metabolic process	+	+
7326	FBgn0076597	-	Dpse\GA16582	structural constituent of ribosome.	translation	+	+
7331	FBgn0047135	CG32276	Dmel\CG32276	unknown	protein modification process	+	+
7339	FBgn0011760	cut up	Dmel\ctp	ATPase activity	microtubule-based movement	+	+
7340	FBgn0052230	CG32230	Dmel\CG32230	NADH dehydrogenase activity	mitochondrial electron transport	+	
7358	FBgn0017579	Ribosomal protein L14	Dmel\RpL14	structural constituent of ribosome	translation		+
7370	FBgn0028342	lethal (1) G0230	Dmel\l(1)G0230	hydrogen-exporting ATPase activity	proton transport		+
7383	FBgn0023211	Elongin C	Dmel\Elongin-C	transcription elongation regulator activity	dendrite morphogenesis	+	
7395	FBgn0000253	Calmodulin	Dmel\Cam	calcium ion binding	kinetochore organization and biogenesis	+	+
7400	FBgn0004432	Cyclophilin 1	Dmel\Cyp1	peptidyl-prolyl cis-trans isomerase activity	protein folding	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7408	FBgn0010408	Ribosomal protein S9	Dmel\RpS9	structural constituent of ribosome	translation		+
7413	FBgn0015790	Rab-protein 11	Dmel\Rab11	GTP binding	anatomical structure development;	+	
7427	FBgn0032597	CG17904	Dmel\CG17904	nucleotide binding	unknown	+	
7430	FBgn0017545	Ribosomal protein S3A	Dmel\RpS3A	structural constituent of ribosome	translation		+
7434	FBgn0039697	CG7834	Dmel\CG7834	electron carrier activity	oxidative phosphorylation	+	+
7437	FBgn0030263	CG2076	Dmel\CG2076	unknown	unknown	+	
7443	FBgn0004117	Tropomyosin 2	Dmel\Tm2	actin binding	heart development		+
7512	FBgn0037346	extra bases	Dmel\exba	protein binding	long-term memory	+	+
7533	FBgn0004867	string of pearls	Dmel\sop	RNA binding	translation	+	+
7538	FBgn0034709	CG3074	Dmel\CG3074	cysteine-type endopeptidase activity	proteolysis	+	+
7542	FBgn0086785	Vps36	Dmel\Vps36	mRNA 3'-UTR binding	unknown	+	
7596	FBgn0019644	ATP synthase, subunit b	Dmel\ATPsyn-b	hydrogen-exporting ATPase activity	proton transport	+	+
7606	FBgn0005593	Ribosomal protein L7	Dmel\RpL7	structural constituent of ribosome	translation	+	+
7609	FBgn0035964	Dihydropteridine reductase	Dmel\Dhpr	6,7-dihydropteridine reductase activity	metabolic process	+	
7631	FBgn0032192	CG5731	Dmel\CG5731	alpha-N-acetylgalactos- aminidase activity	carbohydrate metabolic process	+	+
7640	FBgn0015282	Proteasome 26S subunit subunit 4	Dmel\Pros26.4	ATPase activity	proteolysis	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7660	FBgn0012036	Aldehyde dehydrogenase	Dmel\Aldh	aldehyde dehydrogenase (NAD) activity	pyruvate metabolic process	+	
7720	FBgn0025700	CG5885	Dmel\CG5885	unknown	cotranslational protein targeting	+	+
7731	FBgn0086904	Nascent polypeptide associated co	Dmel\Nac$\alpha$	protein binding	regulation of pole plasm oskar mRNA localization		+
7736	FBgn0035528	CG15012	Dmel\CG15012	beta-N-acetyl-hexosaminidase activity.	unknown	+	
7741	FBgn0016119	ATPase coupling factor 6	Dmel\ATPsyn-Cf6	hydrogen-exporting ATPase activity	proton transport		+
7742	FBgn0038739	CG4686	Dmel\CG4686	unknown	unknown	+	
7771	FBgn0023212	Elongin B	Dmel\Elongin-B	transcription elongation regulator activity	protein modification process	+	
7772	FBgn0002579	Ribosomal protein L36	Dmel\RpL36	structural constituent of ribosome	translation		+
7785	FBgn0005533	Ribosomal protein S17	Dmel\RpS17	structural constituent of ribosome	translation		+
7792	FBgn0035871	CG7188	Dmel\CG7188	unknown	negative regulation of apoptosis	+	+
7795	FBgn0000409	Cytochrome c proximal	Dmel\Cyt-c-p	electron carrier activity	oxidative phosphorylation		+
7799	FBgn0002626	Ribosomal protein L32	Dmel\RpL32	structural constituent of ribosome	translation		+
7805	FBgn0037551	CG7891	Dmel\CG7891	GTP binding	small GTPase mediated signal transduction	+	

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7864	FBgn0036928	Translocase of outer membrane 20	Dmel\Tom20	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	protein targeting to mitochondrion	+	
7867	FBgn0029785	Ribosomal protein L35	Dmel\RpL35	structural constituent of ribosome	translation		+
7868	FBgn0030733	CG3560	Dmel\CG3560	ubiquinol-cytochrome-c reductase activity	mitochondrial electron transport, ubiquinol to cytochrome c		+
7878	FBgn0031459	CG2862	Dmel\CG2862	nucleotidase activity	unknown		+
7883	FBgn0039713	Ribosomal protein S8	Dmel\RpS8	structural constituent of ribosome	translation	+	+
7884	FBgn0015288	Ribosomal protein L22	Dmel\RpL22	structural constituent of ribosome	translation	+	+
7902	FBgn0037231	CG9779	Dmel\CG9779	unknown	phagocytosis	+	
7903	FBgn0029161	slowmo	Dmel\slmo	unknown	larval behavior	+	
7907	FBgn0035853	CG7375	Dmel\CG7375	ubiquitin-protein ligase activity	regulation of protein metabolic process	+	
7914	FBgn0031090	Rab35	Dmel\Rab35	GTP binding	cytokinesis	+	+
7915	FBgn0036460	CG5114	Dmel\CG5114	unknown	unknown	+	
7932	FBgn0062413	Copper transporter 1A	Dmel\Ctr1A	copper ion transmembrane transporter activity	copper ion transport	+	
7935	FBgn0004404	Ribosomal protein S14b	Dmel\RpS14b	structural constituent of ribosome	translation		+
7950	FBgn0034751	Ribosomal protein S24	Dmel\RpS24	structural constituent of ribosome	translation	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
7970	FBgn0001197	Histone H2A variant	Dmel\His2Av	DNA binding	chromatin assembly	+	
7981	FBgn0035588	CG10672	Dmel\CG10672	oxidoreductase activity	metabolic process	+	
8009	FBgn0035471	Sc2	Dmel\Sc2	oxidoreductase activity	protein modification process	+	+
8013	FBgn0002590	Ribosomal protein S5a	Dmel\RpS5a	structural constituent of ribosome	translation	+	+
8016	FBgn0036318	CG11009	Dmel\CG11009	unknown	unknown	+	
8022	FBgn0015756	Ribosomal protein L9	Dmel\RpL9	structural constituent of ribosome	translation		+
8023	FBgn0010409	Ribosomal protein L18A	Dmel\RpL18A	structural constituent of ribosome	translation	+	+
8030	FBgn0041191	Rheb	Dmel\Rheb	GTP binding	imaginal disc growth	+	
8032	FBgn0010226	Glutathione S transferase S1	Dmel\GstS1	glutathione transferase activity	response to oxidative stress	+	+
8051	FBgn0014026	Ribosomal protein L7A	Dmel\RpL7A	structural constituent of ribosome	translation		+
8073	FBgn0023174	Proteasome β2 subunit	Dmel\Pros$\beta$2	endopeptidase activity	ubiquitin-dependent protein catabolic process	+	+
8075	FBgn0003150	Proteasome 29kD subunit	Dmel\Pros29	endopeptidase activity	ATP-dependent proteolysis	+	+
8076	FBgn0033879	CG6543	Dmel\CG6543	enoyl-CoA hydratase activity	fatty acid beta-oxidation	+	+
8090	FBgn0011013	lethal (3) s1921	Dmel\l(3)s1921	deoxyhypusine monooxygenase activity	peptidyl-lysine modification to hypusine	+	

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
8092	FBgn0030724	Nipsnap	Dmel\Nipsnap	unknown	unknown	+	
8185	FBgn0037001	CG6020	Dmel\CG6020	NADH dehydrogenase (ubiquinone) activity	mitochondrial electron transport, NADH to ubiquinone		+
8207	FBgn0000064	Aldolase	Dmel\Ald	fructose-bisphosphate aldolase activity	glycolysis	+	+
8216	FBgn0033902	Transport and Golgi organization 7	Dmel\Tango7	catalytic activity	Golgi organization and biogenesis		+
8220	FBgn0004169	upheld	Dmel\up	tropomyosin binding	mesoderm development		+
8247	FBgn0001145	Glutamine synthetase 2	Dmel\Gs2	glutamate-ammonia ligase activity	glutamate catabolic process	+	+
8307	FBgn0000579	Enolase	Dmel\Eno	phosphopyruvate hydratase activity	glycolysis	+	
8323	FBgn0024833	AP-47	Dmel\AP-47	protein binding	neurotransmitter secretion	+	
8333	FBgn0015808	Sterol carrier protein X-related thiolase	Dmel\ScpX	sterol carrier protein X-related thiolase activity	phospholipid transport	+	
8344	FBgn0031771	CG9140	Dmel\CG9140	NADH dehydrogenase activity	mitochondrial electron transport	+	
8359	FBgn0032444	CG5525	Dmel\CG5525	ATPase activity	mitotic spindle organization and biogenesis	+	
8391	FBgn0011211	bellwether	Dmel\blw	hydrogen-exporting ATPase activity	permatid development		+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
8396	FBgn0001098	Glutamate dehydrogenase	Dmel\Gdh	glutamate dehydrogenase [NAD(P)+] activity	sperm storage	+	
8453	FBgn0037893	CG6719	Dmel\CG6719	chaperone binding	de novo' protein folding	+	
8456	FBgn0034138	Ribosomal protein S15	Dmel\RpS15	structural constituent of ribosome	translation		+
8473	FBgn0037874	Translationally controlled tumor	Dmel\Tctp	guanyl-nucleotide exchange factor activity	positive regulation of multicellular organism growth	+	+
8474	FBgn0032509	CG6523	Dmel\CG6523	disulfide oxidoreductase activity	cell redox homeostasis	+	
8490	FBgn0033544	CG7220	Dmel\CG7220	ubiquitin-protein ligase activity	proteolysis	+	
8517	FBgn0004926	Eukaryotic initiation factor 2β	Dmel\eIF-2β	translation initiation factor activity	translational initiation		+
8547	FBgn0010638	Sec61β	Dmel\Sec61β	protein transporter activity.	SRP-dependent cotranslational protein targeting to membrane		+
8565	FBgn0024939	Ribosomal protein L8	Dmel\RpL8	structural constituent of ribosome	translation		+
8581	unknown	unknown	unknown	unknown	unknown	+	+
8607	FBgn0030082	HP1b	Dmel\HP1b	chromatin binding	chromatin assembly	+	+
8609	FBgn0037328	Ribosomal protein L35A	Dmel\RpL35A	structural constituent of ribosome	translation		+
8613	FBgn0019624	Cytochrome c oxidase subunit Va	Dmel\CoVa	cytochrome-c oxidase activity	mitochondrial electron transport, cytochrome c to oxygen		+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
8622	FBgn0086687	desat1	Dmel\desat1	stearoyl-CoA 9-desaturase activity	cuticle hydrocarbon biosynthetic process	+	+
8624	FBgn0003279	Ribosomal protein L4	Dmel\RpL4	structural constituent of ribosome	translation		+
8627	FBgn0028690	Rpn5	Dmel\Rpn5	endopeptidase activity	proteolysis	+	+
8640	FBgn0027291	lethal (1) G0156	Dmel\l(1)G0156	socitrate dehydrogenase (NAD+) activity	tricarboxylic acid cycle	+	+
8653	FBgn0022774	Ornithine aminotransferase precursor	Dmel\Oat	ornithine-oxo-acid transaminase activity	ornithine metabolic process	+	+
8657	FBgn0028665	VhaAC39	Dmel\VhaAC39	hydrogen-exporting ATPase activity	proton transport	+	
8661	FBgn0001248	Isocitrate dehydrogenase	Dmel\Idh	isocitrate dehydrogenase (NADP+) activity	glyoxylate cycle	+	
8671	FBgn0033663	ERp60	Dmel\ERp60	protein disulfide isomerase activity	protein folding	+	+
8690	FBgn0086133	knockdown	Dmel\kdn	citrate (Si)-synthase activity	tricarboxylic acid cycle	+	
8714	FBgn0030086	CG7033	Dmel\CG7033	ATP-dependent helicase activity	protein folding	+	
8717	FBgn0022097	Vha36	Dmel\Vha36	hydrogen-exporting ATPase activity	proton transport		+
8732	FBgn0064225	Ribosomal protein L5	Dmel\RpL5	structural constituent of ribosome	translation	+	+
8736	FBgn0023477	Tal	Dmel\Tal	ransaldolase activity	pentose-phosphate shunt	+	+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
8740	FBgn0025366	Intronic Protein 259	Dmel\Ip259	unknown	phagocytosis	+	+
8782	FBgn0010602	lesswright	Dmel\lwr	protein binding	regulation of biological process	+	
8784	FBgn0011217	effete	Dmel\eff	protein binding	gamete generation	+	+
8785	FBgn0013325	Ribosomal protein L11	Dmel\RpL11	structural constituent of ribosome	translation		+
8789	FBgn0023175	Proteasome α7 subunit	Dmel\Pros$\alpha$7	endopeptidase activity	ubiquitin-dependent protein catabolic process	+	+
8799	FBgn0004922	Ribosomal protein S6	Dmel\RpS6	structural constituent of ribosome	translation	+	+
8804	FBgn0037063	CG9391	Dmel\CG9391	inositol-1(or 4)-monophosphatase activity	dephosphorylation	+	
8822	FBgn0010348	ADP ribosylation factor 79F	Dmel\Arf79F	GTP binding	protein amino acid ADP-ribosylation	+	+
8884	FBgn0014868	Oligosaccharyltransferase 48kD subunit	Dmel\Ost48	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	protein amino acid N-linked glycosylation	+	
8932	FBgn0032987	Ribosomal protein L21	Dmel\RpL21	structural constituent of ribosome	translation		+
8942	FBgn0024188	separation anxiety	Dmel\san	N-acetyltransferase activity	mitotic sister chromatid cohesion; metabolic process	+	

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
8985	FBgn0025638	Roc1a	Dmel\Roc1a	ubiquitin-protein ligase activity	proteolysis	+	
8986	FBgn0019936	Ribosomal protein S20	Dmel\RpS20	structural constituent of ribosome	translation	+	+
8997	FBgn0039129	Ribosomal protein S19b	Dmel\RpS19b	structural constituent of ribosome	translation	+	+
9007	FBgn0250837	Deoxyuridine triphosphatase	Dmel\dUTPase	dUTP diphosphatase activity	dUTP metabolic process	+	
9017	FBgn0000150	abnormal wing discs	Dmel\awd	microtubule binding	biopolymer modification	+	+
9021	FBgn0039163	CG5515	Dmel\CG5515	unknown	unknown	+	
9093	FBgn0037637	CG9836	Dmel\CG9836	iron-sulfur cluster binding	iron-sulfur cluster assembly	+	
9095	FBgn0028833	Dak1	Dmel\Dak1	cytidylate kinase activity	nucleotide and nucleic acid metabolic process	+	
9097	FBgn0035631	Thioredoxin-like	Dmel\Txl	disulfide oxidoreductase activity	cell redox homeostasis	+	+
9165	FBgn0029133	REG	Dmel\REG	proteasome activator activity	unknown	+	
9169	FBgn0021814	Vps28	Dmel\Vps28	protein binding	actin cytoskeleton organization and biogenesis	+	
9195	FBgn0014020	Rho1	Dmel\Rho1	GTPase activity; protein binding	anatomical structure development;		+
9284	FBgn0035726	CG9953	Dmel\CG9953	serine-type carboxypeptidase activity	proteolysis	+	
9336	FBgn0003941	Ribosomal protein L40	Dmel\RpL40	structural constituent of ribosome	translation		+

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
9344	FBgn0037314	Prosβ4	Dmel\Pros$\beta$4	endopeptidase activity	cell proliferation	+	+
9384	FBgn0011361	mitochondrial acyl carrier protein 1	Dmel\mtacp1	phosphopantetheine binding	mitochondrial electron transport, NADH to ubiquinone	+	
9404	FBgn0031980	Ribosomal protein L36A	Dmel\RpL36A	structural constituent of ribosome	translation		+
9414	FBgn0052672	Autophagy-specific gene 8a	Dmel\Atg8a	unknown	determination of adult life span	+	+
9421	FBgn0032518	Ribosomal protein L24	Dmel\RpL24	structural constituent of ribosome	translation	+	+
9434	FBgn0039132	AP-1σ	Dmel\AP-1σ	protein transporter activity	neurotransmitter secretion	+	
9438	FBgn0039857	Ribosomal protein L6	Dmel\RpL6	structural constituent of ribosome	translation	+	+
9489	FBgn0002174	lethal (2) tumorous imaginal discs	Dmel\l(2)tid	patched binding	smoothened signaling pathway	+	
9502	FBgn0038742	Arc42	Dmel\Arc42	RNA polymerase II transcription mediator activity	transcription initiation from RNA polymerase II promoter	+	
9503	FBgn0005585	Calreticulin	Dmel\Crc	calcium ion binding	central nervous system development		+
9511	FBgn0014189	Helicase at 25E	Dmel\Hel25E	RNA helicase activity	nuclear mRNA splicing, via spliceosome	+	

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
9562	FBgn0028985	Serine protease inhibitor 4	Dmel\Spn4	serine-type endopeptidase inhibitor activity	peptide hormone processing	+	+
9569	FBgn0028662	VhaPPA1-1	Dmel\VhaPPA1-1	hydrogen-exporting ATPase activity	mitotic spindle organization and biogenesis	+	
9590	FBgn0028737	Elongation factor 1 β	Dmel\Ef1β	translation elongation factor activity	translational elongation	+	+
9594	FBgn0250814	-	Dmel\CG4169	ubiquinol-cytochrome-c reductase activity	proteolysis	+	
9603	FBgn0035679	CG10467	Dmel\CG10467	aldose 1-epimerase activity	carbohydrate metabolic process	+	
9616	FBgn0037756	CG8507	Dmel\CG8507	low-density lipoprotein receptor binding	unknown	+	
9666	FBgn0020369	Pros45	Dmel\Pros45	endopeptidase activity	proteolysis	+	
9667	FBgn0036762	CG7430	Dmel\CG7430	dihydrolipoyl dehydrogenase activity	glycine catabolic process	+	+
9684	FBgn0024832	AP-50	Dmel\AP-50	protein binding	neurotransmitter secretion	+	
9751	FBgn0004436	Ubiquitin conjugating enzyme	Dmel\UbcD6	ubiquitin-protein ligase activity	centrosome organization and biogenesis	+	
9753	FBgn0011272	Ribosomal protein L13	Dmel\RpL13	structural constituent of ribosome	translation		+
9813	FBgn0031436	CG3214	Dmel\CG3214	NADH dehydrogenase (ubiquinone) activity	mitochondrial electron transport, NADH to ubiquinone	+	

ID	FlyBase ID	gene name (FlyBase)	symbol	molecular function	biological process involved	*maxgen*	*maxspe*
9821	FBgn0036825	Ribosomal protein L26	Dmel\RpL26	structural constituent of ribosome	translation		+
9826	FBgn0011726	twinstar	Dmel\tsr	actin binding	anatomical structure development	+	+
9827	FBgn0025637	skpA	Dmel\skpA	protein binding	DNA endoreduplication	+	+

Table C.3: Maximum likelihood support of individual alignments (assigned with the numerical identifier) Left: ML support of individual alignments for maxspe data set. **Right:** ML support of individual alignments for maxgen data set. The support for the three different phylogenetic hypotheses of the individual alignments is expressed as the ΔlogL : S.E. and the P-SH value. For the best tree the -logL value is given. Tree1 —Palaeoptera hypothesis, Tree2 —Metapterygota hypothesis, Tree3 —Chiastomyaria hypothesis

maxspe

6671	Tree	ΔlogL:S.E.	P-SH
	1	-1942.68	1
	2	0.73	0.24
	3	0.61	0.25

6790	Tree	ΔlogL:S.E.	P-SH
	1	1.2	0.14
	2	0.52	0.32
	3	-3621.21	1

6906	Tree	ΔlogL:S.E.	P-SH
	1	0.4	0.49
	2	-1306.3	1
	3	0.16	0.53

6927	Tree	ΔlogL:S.E.	P-SH
	1	0.47	0.4
	2	-2902.9	1
	3	0.32	0.43

6958	Tree	ΔlogL:S.E.	P-SH
	1	0.17	0.54
	2	0.83	0.33
	3	-1719.68	1

7007	Tree	ΔlogL:S.E.	P-SH
	1	0.61	0.44
	2	-1287.81	1
	3	0.02	0.64

7015	Tree	ΔlogL:S.E.	P-SH
	1	-3778.23	1
	2	0.26	0.48
	3	0.78	0.29

maxgen

6637	Tree	ΔlogL:S.E.	P-SH
	1	0.08	0.59
	2	0.71	0.4
	3	-2194.7	1

6671	Tree	ΔlogL:S.E.	P-SH
	1	-1278.93	1
	2	0.73	0.22
	3	0.73	0.22

6715	Tree	ΔlogL:S.E.	P-SH
	1	1.16	0.13
	2	-1619.02	1
	3	1.27	0.11

6790	Tree	ΔlogL:S.E.	P-SH
	1	0.97	0.24
	2	0.25	0.5
	3	-2572.28	1

6906	Tree	ΔlogL:S.E.	P-SH
	1	0.53	0.29
	2	0.53	0.29
	3	-982.28	1

6927	Tree	ΔlogL:S.E.	P-SH
	1	0.5	0.44
	2	-1911.92	1
	3	0.21	0.5

6936	Tree	ΔlogL:S.E.	P-SH
	1	-2854.79	1
	2	0.35	0.34
	3	0.35	0.34

maxspe ***maxgen***

7098	Tree	ΔlogL:S.E.	P-SH
	1	-1866.4	1
	2	0.94	0.18
	3	0.94	0.18

6958	Tree	ΔlogL:S.E.	P-SH
	1	5350	0.25
	2	1.21	0.14
	3	-1284.5	1

7214	Tree	ΔlogL:S.E.	P-SH
	1	-2653.18	1
	2	1.26	0.11
	3	1.26	0.11

7007	Tree	ΔlogL:S.E.	P-SH
	1	0.9	0.26
	2	0.41	0.42
	3	-894.09	1

7316	Tree	ΔlogL:S.E.	P-SH
	1	-1252.16	1
	2	0	0.51
	3	0	0.6

7015	Tree	ΔlogL:S.E.	P-SH
	1	-2919.77	1
	2	0.39	0.4
	3	0.75	0.26

7339	Tree	ΔlogL:S.E.	P-SH
	1	-428.72	1
	2	0.67	0.22
	3	0.67	0.22

7098	Tree	ΔlogL:S.E.	P-SH
	1	-1370.06	1
	2	0.53	0.25
	3	0.53	0.25

7434	Tree	ΔlogL:S.E.	P-SH
	1	0.8	0.22
	2	-2309.35	1
	3	0.64	0.24

7214	Tree	ΔlogL:S.E.	P-SH
	1	0.08	0.49
	2	0.08	0.49
	3	-2321.86	1

7512	Tree	ΔlogL:S.E.	P-SH
	1	-4812.26	1
	2	1.02	0.15
	3	0.04	0.14

7316	Tree	ΔlogL:S.E.	P-SH
	1	-865.36	1
	2	0.55	0.28
	3	0.55	0.28

7538	Tree	ΔlogL:S.E.	P-SH
	1	0.68	0.26
	2	0.68	0.26
	3	-6151.21	1

7339	Tree	ΔlogL:S.E.	P-SH
	1	-411.01	1
	2	0	0.14
	3	0	0.15

7606	Tree	ΔlogL:S.E.	P-SH
	1	0.98	0.3
	2	0.08	0.6
	3	-3203.78	1

7383	Tree	ΔlogL:S.E.	P-SH
	1	0	0.58
	2	0	0.26
	3	-422.69	1

7631	Tree	ΔlogL:S.E.	P-SH
	1	-4762.97	1

7434	Tree	ΔlogL:S.E.	P-SH
	1	0.69	0.23

maxspe				*maxgen*			
	2	0.88	0.22		2	-1622.66	1
	3	0.45	0.35		3	0.64	0.24

7640	Tree	ΔlogL:S.E.	P-SH
	1	0	0.11
	2	-1897.11	1
	3	0	0.25

7512	Tree	ΔlogL:S.E.	P-SH
	1	-3521.72	1
	2	1.02	0.15
	3	1.03	0.14

7720	Tree	ΔlogL:S.E.	P-SH
	1	0.75	0.26
	2	0.5	0.33
	3	-2070.27	1

7538	Tree	ΔlogL:S.E.	P-SH
	1	3.19	<0.000
	2	3.19	<0.000
	3	-4515.29	1

7883	Tree	ΔlogL:S.E.	P-SH
	1	0.36	0.32
	2	0	0.32
	3	-2093.84	1

7606	Tree	ΔlogL:S.E.	P-SH
	1	1.19	0.16
	2	0.54	0.32
	3	-2060.65	1

7950	Tree	ΔlogL:S.E.	P-SH
	1	0.64	0.26
	2	0.01	0.28
	3	-1203.47	1

7631	Tree	ΔlogL:S.E.	P-SH
	1	-3929.51	1
	2	0.88	0.18
	3	0.77	0.2

7970	Tree	ΔlogL:S.E.	P-SH
	1	-489.5	1
	2	0	0.38
	3	0	0.4

7640	Tree	ΔlogL:S.E.	P-SH
	1	0	0.47
	2	-3387.26	1
	3	0	0

8013	Tree	ΔlogL:S.E.	P-SH
	1	-1793.43	1
	2	0.45	0.3
	3	0.45	0.3

7720	Tree	ΔlogL:S.E.	P-SH
	1	0.71	0.24
	2	0.71	0.24
	3	-1405.59	1

8023	Tree	ΔlogL:S.E.	P-SH
	1	0.63	0.27
	2	-1978.66	1
	3	1.09	0.18

7736	Tree	ΔlogL:S.E.	P-SH
	1	0.05	0.59
	2	-1684.04	1
	3	0.67	0.43

8032	Tree	ΔlogL:S.E.	P-SH
	1	0.91	0.2
	2	1.31	0.1
	3	-4775.66	1

7742	Tree	ΔlogL:S.E.	P-SH
	1	1.21	0.11
	2	-1732.85	1
	3	1.21	0.11

maxspe				***maxgen***			
8073	Tree	ΔlogL:S.E.	P-SH	**7771**	Tree	ΔlogL:S.E.	P-SH
	1	0.07	0.48		1	1.23	0.11
	2	-3536.16	1		2	-1083.55	1
	3	0.07	0.48		3	1.23	0.11
8075	Tree	ΔlogL:S.E.	P-SH	**7864**	Tree	ΔlogL:S.E.	P-SH
	1	0.58	0.27		1	0.23	0.42
	2	-2078.01	1		2	-1576.5	1
	3	0.67	0.27		3	0.22	0.4
8076	Tree	ΔlogL:S.E.	P-SH	**7883**	Tree	ΔlogL:S.E.	P-SH
	1	0	0.58		1	0.27	0.38
	2	0	0.69		2	0.27	0.38
	3	-2968.41	1		3	-1437.09	1
8456	Tree	ΔlogL:S.E.	P-SH	**7902**	Tree	ΔlogL:S.E.	P-SH
	1	0	1		1	0.04	0.61
	2	0	0.51		2	1.23	0.23
	3	-1180.76	0.32		3	-2242.52	1
8547	Tree	ΔlogL:S.E.	P-SH	**7950**	Tree	ΔlogL:S.E.	P-SH
	1	0.21	0.38		1	0.44	0.48
	2	-1012.64	1		2	0.33	0.5
	3	0.21	0.38		3	-816.95	1
8671	Tree	ΔlogL:S.E.	P-SH	**7970**	Tree	ΔlogL:S.E.	P-SH
	1	0.54	0.3		1	-444.18	1
	2	0.54	0.3		2	0.56	0.27
	3	-6150.07	1		3	0.56	0.27
8732	Tree	ΔlogL:S.E.	P-SH	**8013**	Tree	ΔlogL:S.E.	P-SH
	1	0	0.6		1	-1285.73	1
	2	0	0.47		2	0.72	0.21
	3	-2790.14	1		3	0.72	0.21
8784	Tree	ΔlogL:S.E.	P-SH	**8023**	Tree	ΔlogL:S.E.	P-SH
	1	0	0.02		1	0	0.63
	2	-592.62	1		2	-1306.04	1
	3	0	0.02		3	1.07	0.26
8997	Tree	ΔlogL:S.E.	P-SH	**8032**	Tree	ΔlogL:S.E.	P-SH
	1	0.09	0.57		1	0.65	0.25
	2	-1917.3	1		2	0.6	0.26

maxspe				*maxgen*			
	3	0.4	0.53		3	-3139.07	1

9404	Tree	ΔlogL:S.E.	P-SH
	1	0.12	0.44
	2	-817.29	1
	3	0.12	0.44

9414	Tree	ΔlogL:S.E.	P-SH
	1	0.33	0.34
	2	0.33	0.34
	3	-708.87	1

9562	Tree	ΔlogL:S.E.	P-SH
	1	1.15	0.12
	2	-7451.58	1
	3	1.19	0.12

9590	Tree	ΔlogL:S.E.	P-SH
	1	1.02	0.14
	2	-3019.65	1
	3	0.79	0.2

9821	Tree	ΔlogL:S.E.	P-SH
	1	1.04	0.15
	2	-1579.59	1
	3	1.01	0.17

9827	Tree	ΔlogL:S.E.	P-SH
	1	-1136.47	1
	2	0.75	0.23
	3	0.75	0.23

8073	Tree	ΔlogL:S.E.	P-SH
	1	0	0.78
	2	-2410.92	1
	3	0	0.57

8075	Tree	ΔlogL:S.E.	P-SH
	1	0.36	0.33
	2	-1319.4	1
	3	0.61	0.27

8076	Tree	ΔlogL:S.E.	P-SH
	1	0	0.6
	2	-2301.82	1
	3	0	0.67

8092	Tree	ΔlogL:S.E.	P-SH
	1	0.51	0.29
	2	-2621.95	1
	3	0.61	0.28

8323	Tree	ΔlogL:S.E.	P-SH
	1	-3753.77	1
	2	1.62	0.08
	3	1.63	0.08

8671	Tree	ΔlogL:S.E.	P-SH
	1	-4482.29	1
	2	0.73	0.23
	3	0.73	0.23

8732	Tree	ΔlogL:S.E.	P-SH
	1	0.1	0.45
	2	0.1	0.45
	3	-1933.57	1

8782	Tree	ΔlogL:S.E.	P-SH
	1	0.51	0.27
	2	0.51	0.27
	3	-830.38	1

8784	Tree	ΔlogL:S.E.	P-SH

maxspe	*maxgen*			
		1	0	0.12
		2	-528.5	1
		3	0	0.09
	8942	Tree	ΔlogL:S.E.	P-SH
		1	0.71	0.22
		2	-941.32	1
		3	0.71	0.22
	8997	Tree	ΔlogL:S.E.	P-SH
		1	-1278.6	1
		2	0.42	0.37
		3	0.57	0.35
	9007	Tree	ΔlogL:S.E.	P-SH
		1	1.39	0.09
		2	-1304.15	1
		3	1.31	0.1
	9095	Tree	ΔlogL:S.E.	P-SH
		1	0.86	0.2
		2	1.5	0.08
		3	-2197.76	1
	9169	Tree	ΔlogL:S.E.	P-SH
		1	0.3	0.5
		2	-1596.79	1
		3	0.27	0.53
	9384	Tree	ΔlogL:S.E.	P-SH
		1	1.38	0.09
		2	1.34	0.09
		3	-855.32	1
	9414	Tree	ΔlogL:S.E.	P-SH
		1	0	0.15
		2	0	0.18
		3	-544.6	1
	9489	Tree	ΔlogL:S.E.	P-SH
		1	7.36	<0.000
		2	-2754.7	1

maxspe	*maxgen*			
		3	7.36	<0.000
	9511	Tree	ΔlogL:S.E.	P-SH
		1	11.6	<0.000
		2	-3583.33	1
		3	11.6	<0.000
	9562	Tree	ΔlogL:S.E.	P-SH
		1	0.72	0.23
		2	-4877.63	1
		3	0.89	0.19
	9569	Tree	ΔlogL:S.E.	P-SH
		1	0.74	0.24
		2	-1300.01	1
		3	0.79	0.22
	9590	Tree	ΔlogL:S.E.	P-SH
		1	1.25	0.11
		2	-1982.62	1
		3	1.11	0.14
	9594	Tree	ΔlogL:S.E.	P-SH
		1	1.36	0.1
		2	1.52	0.08
		3	-4281.75	1
	9616	Tree	ΔlogL:S.E.	P-SH
		1	0.36	0.36
		2	-2560.89	1
		3	0.36	0.36
	9813	Tree	ΔlogL:S.E.	P-SH
		1	-765.87	1
		2	0.83	0.18
		3	0.79	0.19
	9827	Tree	ΔlogL:S.E.	P-SH
		1	-846.49	1
		2	0.58	0.24
		3	0.58	0.24

Table C.4: Maximum likelihood bootstrap support of individual alignments (assigned with the numerical identifier). For each gene alignment of the *maxspe* set (Table C.4(a)) and of the *maxgen* set (Table C.4(b)) a maximum likelihood tree with 100 bootstrap replicates was calculated using RAxML. The first column refers to the gene ID of HaMStR, the second column indicates the tree topology. If the topology coincides with a concurrent hypothesis (Palaeoptera, Metapterygota, Chiastomyaria) the bootstrap value of the branch that separates the respective out-group from the respective in-group is written in the third column.

(a)

Core-Ortholog ID	Topology	Bootstrap
6671	Other Topology	
6790	Other Topology	
6906	Other Topology	
6927	Other Topology	
6958	Other Topology	
7007	Other Topology	
7015	Other Topology	
7098	Other Topology	
7214	Other Topology	
7316	Other Topology	
7339	Other Topology	
7434	Other Topology	
7512	Other Topology	
7538	Other Topology	
7606	Other Topology	
7631	Other Topology	
7640	Other Topology	
7720	Other Topology	
7883	Other Topology	
7950	Chiastomyaria	46
7970	Other Topology	
8013	Other Topology	
8023	Metapterygota	34
8032	Other Topology	
8073	Other Topology	
8075	Metapterygota	52
8076	Other Topology	
8456	Other Topology	
8547	Other Topology	
8671	Other Topology	
8732	Palaeoptera	18
8784	Other Topology	
8997	Other Topology	
9404	Other Topology	
9414	Other Topology	
9562	Other Topology	
9590	Other Topology	
9821	Other Topology	
9827	Other Topology	

(b)

Core-Ortholog ID	Topology	Bootstrap
6637	Other Topology	
6671	Other Topology	
6715	Other Topology	
6790	Other Topology	
6906	Other Topology	
6927	Other Topology	
6936	Other Topology	
6958	Other Topology	
7007	Other Topology	
7015	Palaeoptera	45
7098	Other Topology	
7214	Other Topology	
7316	Other Topology	
7339	Other Topology	
7383	Palaeoptera	5
7434	Other Topology	
7512	Other Topology	
7538	Other Topology	
7606	Other Topology	
7631	Other Topology	
7640	Other Topology	
7720	Other Topology	
7736	Other Topology	
7742	Metapterygota	50
7771	Other Topology	
7864	Other Topology	
7883	Chiastomyaria	35
7902	Palaeoptera	24
7950	Chiastomyaria	56
7970	Other Topology	
8013	Other Topology	
8023	Other Topology	
8032	Other Topology	
8073	Other Topology	
8075	Metapterygota	63
8076	Other Topology	
8323	Other Topology	
8671	Palaeoptera	59
8732	Other Topology	
8782	Other Topology	
8784	Other Topology	
8942	Metapterygota	41
8997	Other Topology	
9007	Other Topology	
9095	Other Topology	
9169	Other Topology	
9384	Other Topology	
9414	Other Topology	
9489	Other Topology	
9511	Other Topology	
9562	Other Topology	
9569	Other Topology	
9590	Metapterygota	95
9594	Other Topology	
9616	Other Topology	
9813	Other Topology	
9827	Other Topology	

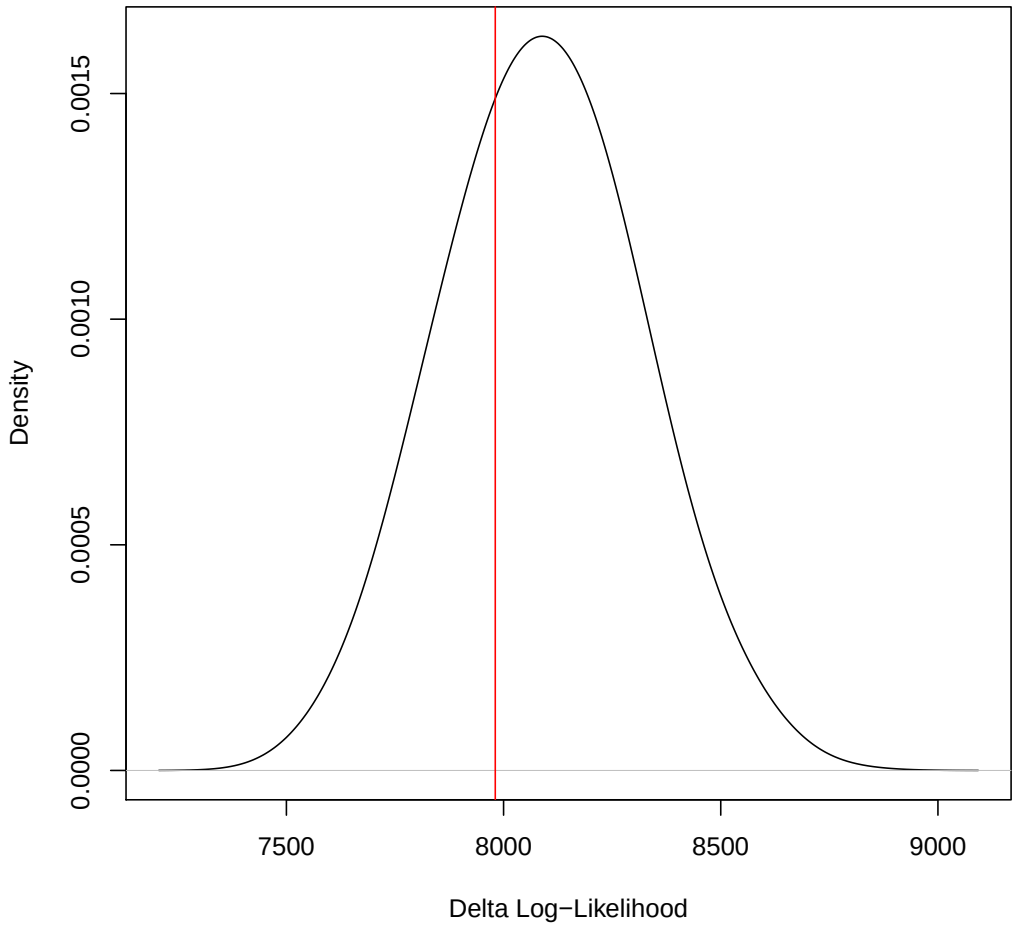


Figure C.1: Distribution of delta values for simulated maxgen alignments without gaps. We removed all columns containing gaps from the concatenated maxgen alignment processed with ALISCORE and calculated a maximum likelihood tree with RAxML under the WAG model. Afterwards we simulated 1,000 alignments of equal length using Seq-Gen and the parameters obtained by the maximum likelihood tree reconstruction. Following the test introduced in Goldman (1993) we reconstructed the maximum likelihood tree and calculated the difference of the unconstrained log-likelihood and the maximum log-likelihood (delta value) for each simulated alignment. Shown is the distribution of delta values for the simulated alignments. The red vertical line marks the delta value for the real alignment.

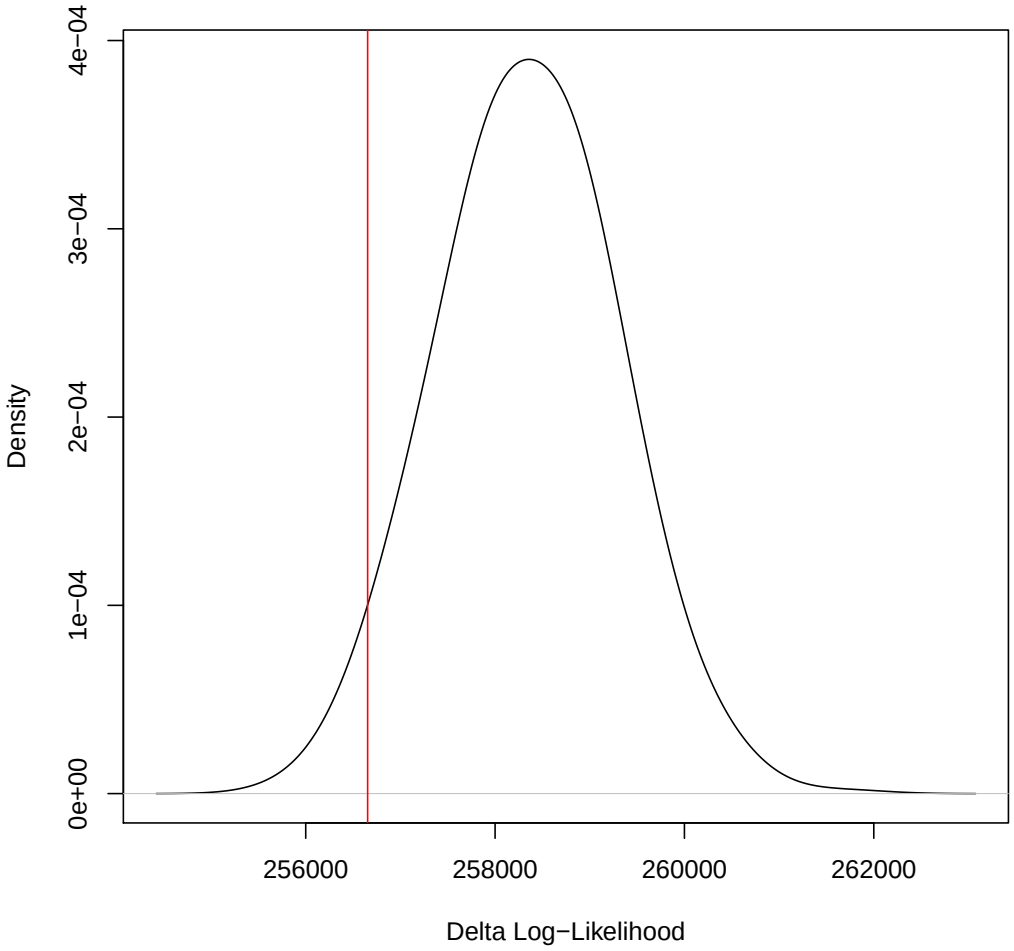


Figure C.2: Distribution of delta values for simulated maxgen alignments with gaps. We simulated 1,000 alignments with Seq-Gen using the parameter obtained by the maximum likelihood tree reconstruction of the concatenated maxgen alignment including all positions with gaps or missing data. We then replaced amino acids of the simulated data with gaps or missing data where there are gaps or missing data in the real alignment. Afterwards we proceeded as described in Supplement Figure C.1. Shown is the distribution of delta values of the simulated alignments. The red vertical line marks the delta value for the real alignment.

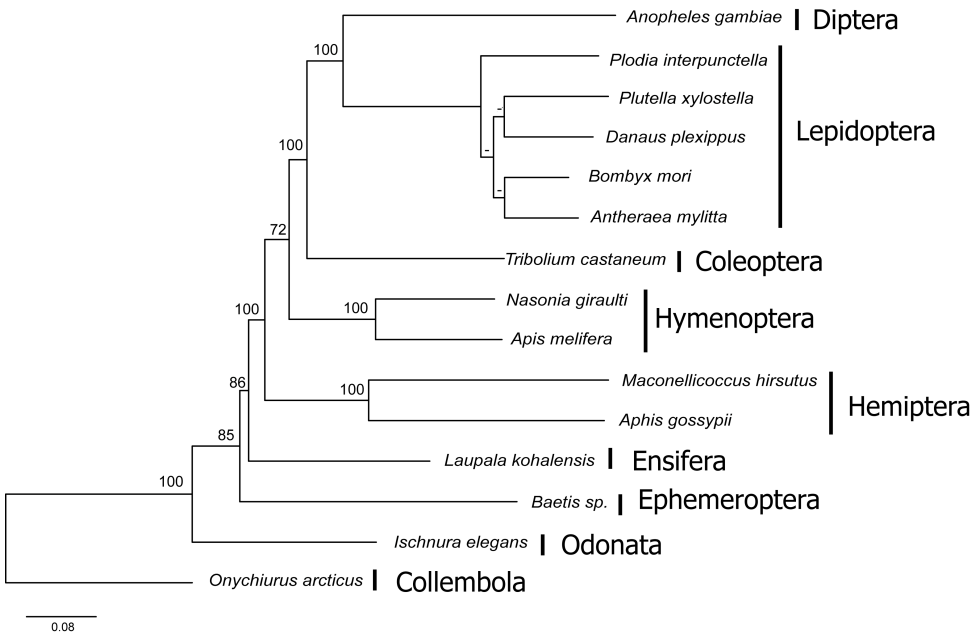


Figure C.3: Maximum likelihood topology of *maxspe* Each gene alignment of the *maxspe* set was processed with ALISCORE. Afterwards the best suited model of evolution was determined for each processed alignment with ProtTest and the alignments were concatenated. The maximum likelihood tree was calculated using RAxML's '-q' option, that allows a partitioning of the alignment with an individual model of evolution for each partition. Support values were assessed by 100 bootstrap replicates. The tree is congruent to the tree shown in 8.2. Thus, even if sequence evolution is modeled individually for each gene, the Chiastomyaria hypothesis is supported.

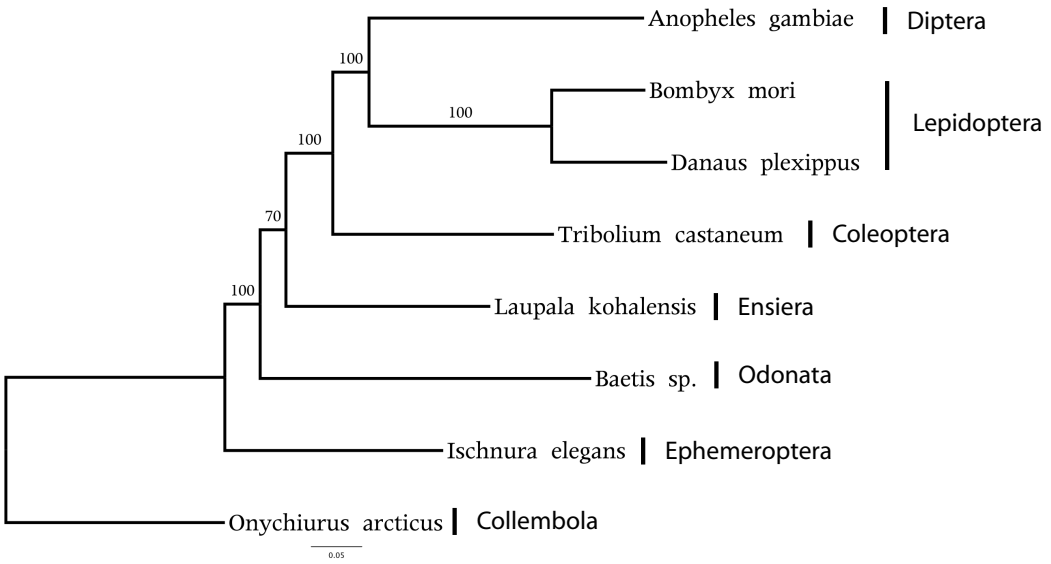


Figure C.4: Maximum likelihood topology of *maxgen* Each gene alignment of the *maxgen* set was processed with ALISCORE. Afterwards the best suited model of evolution was determined for each with ProtTest and the alignments were concatenated. The maximum likelihood tree was calculated using RAxML's '-q' option, that allows a partitioning of the alignment with an individual model of evolution for each partition. Support values were assessed by 100 bootstrap replicates. The tree is congruent to the tree shown in Figure 8.3. Thus, even if sequence evolution is modeled individually for each gene, the Chiastomyaria hypothesis is supported.

D Aspects of EST-based Phylogenetics

Table D.1: Data source Chordata network This table lists the data that were scanned for orthologs to compile a gene set for the network in Fig. 9.1. An 'X' in the last column marks those taxa that were considered for the network after we applied our gene selection strategy.

Species Name	Data Type	Source	Included
Acipenser sinensis	EST	dbEST (NCBI)	
Acipenser transmontanus	EST	dbEST (NCBI)	x
Agkistrodon piscivorus leucostoma	EST	dbEST (NCBI)	
Alligator mississippiensis	EST	dbEST (NCBI)	
Ambystoma mexicanum	EST	dbEST (NCBI)	x
Ambystoma tigrinum tigrinum	EST	dbEST (NCBI)	x
Anas platyrhynchos	EST	dbEST (NCBI)	
Andrias davidianus	EST	dbEST (NCBI)	
Anguilla japonica	EST	dbEST (NCBI)	
Anolis carolinensis	EST	dbEST (NCBI)	x
Anolis sagrei	EST	dbEST (NCBI)	
Apis mellifera	Proteome	HaMStR (Chordata set)	x
Apostichopus japonicus	EST	dbEST (NCBI)	
Astatotilapia burtoni	EST	Gene Index (DFGI)	
Asterina pectinifera	EST	dbEST (NCBI)	x
Bos indicus	EST	dbEST (NCBI)	
Bos sp.	EST	dbEST (NCBI)	
Bos taurus	Proteome	Ensembl	x
Bothrops insularis	EST	dbEST (NCBI)	
Bothrops jararaca	EST	dbEST (NCBI)	
Botryllus schlosseri	EST	dbEST (NCBI)	
Branchiostoma floridae	Proteome	JGI	x
Bubalus bubalis	EST	dbEST (NCBI)	
Caenorhabditis briggsae	Proteome	HaMStR (Chordata set)	x
Caenorhabditis remanei	Proteome	HaMStR (Chordata set)	
Callithrix jacchus	EST	dbEST (NCBI)	
Canis familiaris	Proteome	HaMStR (Chordata set)	x
Capra hircus	EST	dbEST (NCBI)	x
Carassius auratus	EST	dbEST (NCBI)	
Cavia porcellus	Proteome	Ensembl	x
Cervus canadensis nelsoni	EST	dbEST (NCBI)	
Chinchilla lanigera	EST	dbEST (NCBI)	
Ciona intestinalis	Proteome	HaMStR (Chordata set)	x
Ciona savignyi	Proteome	Broad Institute	x
Columba livia	EST	dbEST (NCBI)	
Ctenopharyngodon idella	EST	dbEST (NCBI)	
Cyprinus carpio	EST	dbEST (NCBI)	x

Species Name	Data Type	Source	Included
Danio rerio	Proteome	HaMStR (Chordata set)	x
Dasypus novemcinctus	Proteome	Broad Institute	x
Deinagkistrodon acutus	EST	dbEST (NCBI)	
Dicentrarchus labrax	EST	dbEST (NCBI)	
Diplosoma listerianum	EST	dbEST (NCBI)	
Drosophila melanogaster	Proteome	HaMStR (Chordata set)	x
Echinops telfairi	Proteome	Ensembl	x
Echis ocellatus	EST	dbEST (NCBI)	
Elaphe quadrivirgata	EST	dbEST (NCBI)	
Epinephelus coioides	EST	dbEST (NCBI)	
Eptatretus burgeri	EST	dbEST (NCBI)	x
Equus caballus	Proteome	Ensembl	x
Erinaceus europaeus	Proteome	Ensembl	x
Esox lucius	EST	dbEST (NCBI)	
Eubalaena glacialis	EST	dbEST (NCBI)	
Felis catus	Proteome	Ensembl	x
Fundulus heteroclitus	EST	Gene Index (DFGI)	x
Gadus morhua	EST	dbEST (NCBI)	x
Gallus gallus	Proteome	HaMStR (Chordata set)	x
Gasterosteus aculeatus	Proteome	Ensembl	x
Gekko japonicus	EST	dbEST (NCBI)	
Gillichthys mirabilis	EST	dbEST (NCBI)	
Gobiocypris rarus	EST	dbEST (NCBI)	
Halocynthia roretzi	EST	dbEST (NCBI)	x
Haplochromis chilotes	EST	Gene Index (DFGI)	x
Heliocidaris erythrogramma	EST	dbEST (NCBI)	
Hemicentrotus pulcherrimus	EST	dbEST (NCBI)	
Herdmania momus	EST	dbEST (NCBI)	
Hippocampus comes	EST	dbEST (NCBI)	
Hippoglossus hippoglossus	EST	dbEST (NCBI)	x
Holothuria glaberrima	EST	dbEST (NCBI)	
Homo sapiens	Proteome	HaMStR (Chordata set)	x
Ictalurus furcatus	EST	dbEST (NCBI)	x
Ictalurus punctatus	EST	Gene Index (DFGI)	x
Isoodon macrourus	EST	dbEST (NCBI)	
Lachesis muta	EST	dbEST (NCBI)	
Lama pacos	EST	dbEST (NCBI)	
Lates calcarifer	EST	dbEST (NCBI)	
Leucoraja erinacea	EST	dbEST (NCBI)	x
Lipochromis sp matumbi hunter	EST	dbEST (NCBI)	x
Lithognathus mormyrus	EST	dbEST (NCBI)	
Loxodonta africana	Proteome	Ensembl	x
Macaca fascicularis	EST	Gene Index (DFGI)	x
Macaca nemestrina	EST	dbEST (NCBI)	x
Macropus eugenii	EST	dbEST (NCBI)	x
Marmota monax	EST	dbEST (NCBI)	
Meleagris gallopavo	EST	dbEST (NCBI)	x
Mesocricetus auratus	EST	dbEST (NCBI)	
Microcebus murinus	Proteome	Ensembl	x
Misgurnus anguillicaudatus	EST	dbEST (NCBI)	x
Molgula tectiformis	EST	dbEST (NCBI)	x
Monodelphis domestica	Proteome	HaMStR (Chordata set)	x
Mus musculus	Proteome	HaMStR (Chordata set)	x
Myotis lucifugus	Proteome	Ensembl	x

Species Name	Data Type	Source	Included
Neovison vison	EST	dbEST (NCBI)	
Ochotona princeps	Proteome	Ensembl	x
Odocoileus virginianus	EST	dbEST (NCBI)	
Oikopleura dioica	EST	dbEST (NCBI)	
Oncorhynchus mykiss	EST	Gene Index (DFGI)	x
Oncorhynchus nerka	EST	dbEST (NCBI)	x
Oncorhynchus tshawytscha	EST	dbEST (NCBI)	x
Opsanus beta	EST	dbEST (NCBI)	
Oreochromis mossambicus	EST	dbEST (NCBI)	
Oreochromis niloticus	EST	dbEST (NCBI)	
Ornithorhynchus anatinus	Proteome	Ensembl	x
Oryctolagus cuniculus	Proteome	Ensembl	x
Oryzias latipes	Proteome	Ensembl	x
Osmerus mordax	EST	dbEST (NCBI)	x
Otolemur garnettii	Proteome	Ensembl	x
Ovis aries	EST	dbEST (NCBI)	x
Pan troglodytes	Proteome	Ensembl	x
Pan troglodytes verus	EST	dbEST (NCBI)	
Papio anubis	EST	dbEST (NCBI)	x
Paracentrotus lividus	EST	dbEST (NCBI)	
Paralabidochromis chilotes	EST	dbEST (NCBI)	x
Paralichthys lethostigma	EST	dbEST (NCBI)	
Paralichthys olivaceus	EST	dbEST (NCBI)	x
Patiria miniata	EST	dbEST (NCBI)	
Perca fluviatilis	EST	dbEST (NCBI)	
Peromyscus maniculatus bairdii	EST	dbEST (NCBI)	
Petromyzon marinus	EST	dbEST (NCBI)	x
Philodryas olfersii	EST	dbEST (NCBI)	
Pimephales promelas	EST	dbEST (NCBI)	x
Platichthys flesus	EST	dbEST (NCBI)	x
Plecoglossus altivelis altivelis	EST	dbEST (NCBI)	
Poecilia reticulata	EST	dbEST (NCBI)	x
Polyandrocarpa misakiensis	EST	dbEST (NCBI)	
Pomatomus saltatrix	EST	dbEST (NCBI)	
Pongo pygmaeus	Proteome	Ensembl	x
Psetta maxima	EST	dbEST (NCBI)	
Pseudopleuronectes americanus	EST	dbEST (NCBI)	
Pseudosciaena crocea	EST	dbEST (NCBI)	
Ptychodera flava	EST	Trace Archive (NCBI)	x
Ptyochromis sp.	EST	dbEST (NCBI)	x
Rattus norvegicus	Proteome	Ensembl	x
Rutilus rutilus	EST	dbEST (NCBI)	x
Saccoglossus kowalevskii	EST	dbEST (NCBI)	x
Salmo salar	EST	Gene Index (DFGI)	x
Salvelinus fontinalis	EST	dbEST (NCBI)	
Sarotherodon melanotheron	EST	dbEST (NCBI)	
Sebastes rastrelliger	EST	dbEST (NCBI)	
Seriola quinqueradiata	EST	dbEST (NCBI)	
Sistrurus catenatus edwardsi	EST	dbEST (NCBI)	
Sminthopsis crassicaudata	EST	dbEST (NCBI)	
Solaster stimpsonii	EST	dbEST (NCBI)	
Sparus aurata	EST	dbEST (NCBI)	
Spermophilus lateralis	EST	dbEST (NCBI)	
Spermophilus tridecemlineatus	Proteome	Ensembl	x

Species Name	Data Type	Source	Included
Squalus acanthias	EST	dbEST (NCBI)	x
Strongylocentrotus purpuratus	Proteome	HaMStR (Chordata set)	x
Sus scrofa	EST	Gene Index (DFGI)	x
Taeniopygia guttata	EST	dbEST (NCBI)	x
Takifugu rubripes	Proteome	Ensembl	x
Tetraodon nigroviridis	Proteome	Uniprot	x
Thalassophryne nattereri	EST	dbEST (NCBI)	
Thunnus thynnus	EST	dbEST (NCBI)	
Torpedo californica	EST	dbEST (NCBI)	
Tupaia belangeri	Proteome	Ensembl	x
Tursiops truncatus	EST	dbEST (NCBI)	
Ursus americanus	EST	Gene Index (DFGI)	x
Xenopus tropicalis	Proteome	HaMStR (Chordata set)	x
Xenoturbella bocki	EST	dbEST (NCBI)	x
Xiphophorus maculatus	EST	dbEST (NCBI)	
Zosterisessor ophiocephalus	EST	dbEST (NCBI)	

E TonB-dependent Transporter

Table E.1: List of known TBDTs All TBDTs used for the analysis of the CLANs results are given. Italic indicates TBDTs with predicted substrate. The left columns give the numbers used in Figure 10.3, the second column the Cluster (see Figure 10.3A), the third column gives the assigned name, the fourth column the GenBank ID, the fifth column the source species, the sixth column the identified siderophores recognized by the according protein, the seventh column the siderophore classification and the eighth column a representative reference for the substrate and for substrate classification (o.a.: only annotated). ['metal'](column 6) indicates that the type of transported metal ion is known, but the according metallophore has not yet been identified.

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
1	11	OprC	1498191	Pseudomonas aeruginosa	Copper chelate	Unknown	(Yoneyama and Nakae (1996))
2	12	BfeA	538279	Bordetella pertussis	enterobactin	Catecholate	(Beall and Sanden (1995), Pollack and Neilands (1970))
3	12	PirA	2981053	Pseudomonas aeruginosa	enterobactin	Catecholate	(Pollack and Neilands (1970), Ochsner *et al.* (2000))
4	12	PfeA	548479	Pseudomonas aeruginosa	enterobactin	Catecholate	(Pollack and Neilands (1970), Dean and Poole (1993))
5	12	FepA	2507463	Escherichia coli K12	enterobactin	Catecholate	(Lundrigan and Kadner (1986), Pollack and Neilands (1970))
6	12	IroN	2738252	Salmonella enterica	salmochelin	glycosylated Catecholate	(Hantke *et al.* (2003), Bister *et al.* (2004))
7	12	CfrA	112360090	Campylobacter jejuni	enterobactin	Catecholate	(Pollack and Neilands (1970), Carswell *et al.* (2008))
8	12	CirA	2507462	Escherichia coli K12	2,3-dihydroxybenzoylserine (DHBS)	Catecholate-Carboxylate	(Nau and Konisky (1989), Hantke (1990))
9	12	IrgA	12644182	Vibrio cholerae	Enterobactin	Catecholate	(Pollack and Neilands (1970), Goldberg *et al.* (1992))
10	12	BfrA	1314835	Bordetella bronchiseptica	2,3-dihydroxybenzoylserine (DHBS)	Catecholate-Carboxylate	(Hantke (1990), Beall and Hoenes (1997))
11	18	HutR	147671724	Vibrio cholerae	haem	Porphyrine	(Mey and Payne (2001))
12	18	HuvA	12697532	Listonella anguillarum	haem	Porphyrine	(Henderson and Payne (1994))
13	18	HutA	529727	Vibrio cholerae	haem	Porphyrine	(Mazoy *et al.* (2003))
14	18	PhuR	3044098	Pseudomonas aeruginosa	haem	Porphyrine	(Ochsner *et al.* (2000))

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
15	18	PfhR	4838477	Pseudomonas fluorescens	haem	Porphyrine	(Ochsner *et al.* (2000))
16	132	HpuB	11386826	Neisseria meningitidis	haem	Porphyrine	(Lewis *et al.* (1997))
17	108	HmbR	687640	Neisseria meningitidis	haem	Porphyrine	(Stojijkovic *et al.* (1995))
18	17	HgbA	28194090	Actinobacillus pleuropneumoniae	haem	Porphyrine	(Srikumar *et al.* (2004))
19	86	HemR	6016198	Yersinia enterocolitica	haem	Porphyrine	(Stojiljkovic and Hantke (1992))
20	86	HmuR	2501236	Yersinia pestis	haem	Porphyrine	(Hornung *et al.* (1996))
21	86	ChuA	1763009	Escherichia coli O157:H7	haem	Porphyrine	(Torres and Payne (1997))
22	86	ShuA	1655877	Shigella dysenteriae	haem	Porphyrine	(Mills and Payne (1997))
23	86	HxuC	1170441	Haemophilus influenzae	haem	Porphyrine	(Cope *et al.* (1995))
24	48	TdhA	33151615	Haemophilus ducreyi 35000HP	haem	Porphyrine	(Thomas *et al.* (1998))
25	152	HasR [T]	34787214	Serratia marcescens	haem	Porphyrine	(Letoffe *et al.* (1994))
26	107	MhuA	50403825	Moraxella catarrhalis	haem	Porphyrine	(Furano *et al.* (2005))
27	15	FetA_FrpB	4768684	Neisseria gonorrhoeae	enterobactin	Catecholate	(Pollack and Neilands (1970), Carson *et al.* (1999))
28	59	VctA	18476494	Vibrio cholerae	enterobactin	Catecholate	(Pollack and Neilands (1970), Mey *et al.* (2002))
29	16	LbpA	915278	Neisseria gonorrhoeae	lactoferrin	Fe(III)-binding protein	(Perkins-Balding *et al.* (2004), Biswas and Sparling (1995))
30	16	TbpA	150361	Neisseria gonorrhoeae	transferrin	Fe(III)-binding protein	(Perkins-Balding *et al.* (2004), Cornelissen *et al.* (1992))
31	112	FrpB4	15646121	Helicobacter pylori 26695	[Nickel]	unknown	(Schauer *et al.* (2007))
32	4	BtuB	416728	Escherichia coli K12	Vitamin B12	Porphyrine	(Heller and Kadner (1985))
33	*4*	*XCC3067*	*21232497*	*Xanthomonas campestris pv. campestris str. ATCC 33913*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
34	*4*	*PA1271*	*15596468*	*Pseudomonas aeruginosa PAO1*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
35	*4*	*BPSL0976*	*53718618*	*Burkholderia pseudomallei K96243*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
36	*4*	*RS02718*	*17547119*	*Ralstonia solanacearum GMI1000*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
37	*165*	*CC1750*	*109897435*	*Pseudoalteromonas atlantica T6c*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
38	*4*	*VC0156*	*15640186*	*Vibrio cholera*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
39	*160*	*RSP_2402*	*77462960*	*Rhodobacter sphaeroides 2.4.1*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
40	40	MxcH	162452159	*Sorangium cellulosum* 'So ce 56'	Myxochelin	Catecholate	(Silakowski *et al.* (2000), Kunze *et al.* (1989))
41	10	IutA	1170593	Escherichia coli	aerobactin	Citrate-Hydroxamate	(Krone *et al.* (1985), Gibson and Magrath (1969))
42	10	RhtA	6685883	Sinorhizobium meliloti	Rhizobactin 1021	Citrate-Hydroxamate	(Lynch *et al.* (2001), Persmark *et al.* (1993))
43	*166*	*SO_0815*	*24372404*	*Shewanella oneidensis MR-1*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
44	82	RumA	1247762	Morganella morganii	ferric rhizoferrin	Carboxylate	(Kühn *et al.* (1996), Drechsel *et al.* (1991))
45	82	FecA [T]	729471	Escherichia coli K12	diferric dicitrate	Citrate	(Pressler *et al.* (1988), Ferguson *et al.* (2002))
46	*25*	*PA2911*	*15598107*	*Pseudomonas aeruginosa PAO1*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
47	*25*	*RPA0407*	*39933484*	*Rhodopseudomonas palustris CGA009*	*Vitamin B12*	Porphyrine	(Rodionov *et al.* (2003))
48	0	VciA	147673813	Vibrio cholerae O395	unknown	unknown	(Mey *et al.* (2008))
49	0	PiuA_Fiu	115587765	Pseudomonas aeruginosa	pyochelin	Phenolate	(Ochsner and Vasil (1996), Cox *et al.* (1981))
50	0	FoxA	1169726	Yersinia enterocolitica	desferrioxamine	Hydroxamate	(Bäumler and Hantke (1992), Keller-Schierlein and Prelog (1961))
51	0	FegA	1518696	Bradyrhizobium japonicum	desferrioxamine	Hydroxamate	(LeVier and Guerinot (1996), Keller-Schierlein and Prelog (1961))
52	0	FctA	871032	Erwinia chrysanthemi	chrysobactin	Catecholate	(Sauvage *et al.* (1996), Persmark *et al.* (1989))
53	0	FmtA	53719389	Burkholderia pseudomallei K96243	ferric malleobactin	Hydroxamate	(Alice *et al.* (2006), Yang *et al.* (1991))

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
54	0	OrbA	76810798	Burkholderia pseudomallei	ferric ornibactin	Citrate-Hydroxamate	(Sokol *et al.* (2000), Stephan *et al.* (1993))[a]
55	0	FhuA	2507464	Escherichia coli K12	ferrichrome	Hydroxamate	(Coulton *et al.* (1986), Zalkin *et al.* (1964))
56	0	OptS [T]	116050410	Pseudomonas aeruginosa	desferrioxamine	Hydroxamate	(Keller-Schierlein and Prelog (1961), Llamas *et al.* (2006))
57	0	BfrI	33592999	Bordetella pertussis Tohama I	unknown	unknown	
58	-	BfrZ [T]	6850914	Bordetella bronchiseptica	unknown	unknown	(Pradel and Locht (2001))
59	-	PrhA [T]	17549099	Ralstonia solanacea rum	transducer without transport function	unknown	(Brito *et al.* (2002))
60	6	PbuA [T]	1172035	Pseudomonas sp. M114	pseudobactin M114	Citrate-Catecholate-Hydroxamate	(Morris *et al.* (1994), Teintze and Leong (1981))
61	6	FpvA [T]	12230910	Pseudomonas aeruginosa	pyoverdine	Catecholate-Hydroxamate	(Poole *et al.* (1993), Morris *et al.* (1994))
62	6	PupA [T]	45723	Pseudomonas putida WCS358	pseudobactin A	Citrate-Catecholate-Hydroxamate	(Teintze and Leong (1981), Bitter *et al.* (1991))
63	6	PupB [T]	585759	Pseudomonas putida WCS358	pseudobactin A	Citrate-Catecholate-Hydroxamate	(Teintze and Leong (1981), Koster *et al.* (1993))
64	6	FauA	4589285	Bordetella pertussis	alcaligin	Hydroxamate	(Brickman and Armstrong (1999), Nishio *et al.* (1988))
65	6	FhuE	2507465	Escherichia coli K12	Coprogen, ferrioxamine B, rhodoturolic acid	Hydroxamates	Sauer *et al.* (1990), Dhungana *et al.* (2001)
66	6	FptA	1169730	Pseudomonas aeruginosa	pyochelin	Phenolate	(Cox *et al.* (1981), Ankenbauer and Quan (1994))
67	*9*	*Bcep18194_B2436*	*78063283*	*Burkholderia sp. 383*	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
68	*9*	*XCC0674*	*21230149*	X. campestris pv. Campestris str. ATCC 33913	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
69	7	FatA	132510	Listonella anguillarum	anguibactin	Catecholate-Hydroxamate	(Actis *et al.* (1988), Wuest *et al.* (2009))
70	7	BauA	49175779	Acinetobacter baumannii	anguibactin	Catecholate-Hydroxamate	(Wuest *et al.* (2009), Dorsey *et al.* (2004))

[a]... we have marked one sequence (54; annotated as OrbA in GenBank) with 79% similarity and 67% identity to the one described in the reference (Burkholderia cepacia, 11230853, Sokol *et al.* (2000)); sequences with a higher similarity/identity are in the same cluster, too.

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
71	7	FcuA	1169655	Yersinia enterocolitica	anguibactin	Catecholate-Hydroxamate	(Koebnik *et al.* (1993), Wuest *et al.* (2009))
72	79	FyuA	517234	Yersinia enterocolitica	yersiniabactin	Phenolate	(Rakin *et al.* (1994), Drechsel *et al.* (1995))
73	79	IrpC	17380443	Yersinia pestis	yersiniabactin	Phenolate	(Drechsel *et al.* (1995), Fetherston *et al.* (1995))
74	140	ViuA	267356	Vibrio cholerae	vibriobactin	Catecholate	(Butterton *et al.* (1992), Griffiths *et al.* (1984))
75	*118*	*SO_2715*	*24374256*	*Shewanella oneidensis MR-1*	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
76	*118*	*CPS_0067*	*71281279*	*Colwellia psychrerythraea 34H*	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
77	45	SftP	6019468	Pseudomonas putida	hexylsulfate		(Kahnert and Kertesz (2000))
78	64	SuxA	21232787	Xanthomonas campestris pv. campestris str. ATCC 33913	sucrose	disaccharide	(Blanvillain *et al.* (2007))
79	*64*	*Sfri_3988*	*114565138*	Shewanella frigidimarina NCIMB 400	*sucrose*	*disacharide*	(Blanvillain *et al.* (2007))
80	*52*	*bll6948*	*27382059*	Bradyrhizobium japonicum USDA 110	*[Nickel]*	*unknown*	
81	*52*	*Daro_1684*	*71907314*	Dechloromonas aromaticaRCB	*[Cobalt]*	*unknown*	(Rodionov *et al.* (2006))
82	*52*	*Daro_3944*	*71909555*	Dechloromonas aromatic RCB	*[Nickel]*	*unknown*	(Rodionov *et al.* (2006))
83	*9*	*BF0615*	*53711906*	Bacteroides fragilis YCH46	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
84	*9*	*PG1899*	*34541505*	*Porphyromonas gingivalis* W83	*Thiamin*	*Vitamin B1*	(Rodionov *et al.* (2002))
85	26	RagA	110636973	Cytophaga hutchinsonii ATCC 33406	digested proteins	Polypeptides	(Nagano *et al.* (2007))[b]
86	26	RagA	110636966	Cytophaga hutchinsonii ATCC 33406	digested proteins	Polypeptides	(Nagano *et al.* (2007))*b*

[b]. . . we have marked two sequences (85, 86; annotated as RagA in GenBank) with 65% similarity and 31% identity to the one described in the reference (Porphyromonas gingivalis W83, 34540042, (Nagano *et al.* (2007)).

No.	Cluster	Name	GenBank	Species	Substrate	Siderophore/Substrate classification	Ref.
87	26	SusC	29349110	Bacteroides thetaiotaomicron VPI-5482	Malto-oligosaccharides /starch	Oligo-/Polysaccharides	(Reeves *et al.* (1996))
88	26	CsuF	29348741	Bacteroides thetaiotaomicron VPI-5482	Chondroitin sulfate / hyaluronic acid	unbranched polysaccharides (GlcA GalNAc) / unbranched polymer of N-acetylglucosamine + glucuronic acid	(Cheng *et al.* (1995), HOFFMAN *et al.* (1960), Atkins and Sheehan (1971))
89	26	OmpW	29348978	Bacteroides thetaiotaomicron VPI-5482	unknown	unknown	(Wei *et al.* (2001))
90	63	MalA	16126526	Caulobacter crescentus CB15	Maltodextrins	starch hydrolysate	(Neugebauer *et al.* (2005))
91	*63*	*SO_3514*	*24375018*	*Shewanella oneidensis MR-1*	*Chitooligosaccharides*	N-acetylglucosamine oligomer	(Yang *et al.* (2006), Karlsen and Hough (1995))
92	*63*	*Sden_2708*	*91794059*	Shewanella denitrificans OS217	*Chitooligosaccharides*	N-acetylglucosamine oligomer	(Yang *et al.* (2006), Karlsen and Hough (1995))
93	*63*	*CPS_1021*	*71281574*	Colwellia psychrerythraea 34H	*Chitooligosaccharides*	N-acetylglucosamine oligomer	(Yang *et al.* (2006), Karlsen and Hough (1995))
94	*63*	*CC_0446*	*16124701*	Caulobacter crescentus CB15	*Chitooligosaccharides*	N-acetylglucosamine oligomer	(Yang *et al.* (2006), Karlsen and Hough (1995))
95	*63*	*XCC0120*	*21229598*	X. campestris pv. campestris str. ATCC 33913	*Pectin*	*heteropolysacchride*	(Blanvillain *et al.* (2007))
96	*63*	*XCC2944*	*21232375*	X. campestris pv. campestris str. ATCC 33913	*Chitooligosaccharides*	N-acetylglucosamine oligomer	(Yang *et al.* (2006), Karlsen and Hough (1995))
97	*72*	*XCC4120*	*21233542*	X. campestris pv. campestris str. ATCC 33913	*Xylan*	*heteropolysaccheride*	(Blanvillain *et al.* (2007))
98	*41*		*53713281*	Bacteroides fragilis YCH46	Fibronectin	protein	(Pauer *et al.* (2009))

Table E.2: Number of TBDTs detected in analyzed genomes Species names are given in columns 1 and 3, the number of identified TBDTs in the genome of the according species in columns 2 and 4, respectively. All analyzed species without TBDTs are not listed.

Species	TBDTs	Species	TBDTs
Acaryochloris marina MBIC11017	12	*Bordetella petrii DSM 12804*	12
Acidiphilium cryptum JF-5	3	*Bradyrhizobium japonicum*	1
Acidobacteria bacterium Ellin345	4	*Bradyrhizobium japonicum USDA 110*	10
Acidovorax avenae subsp. citrulli AAC00-1	14	*Bradyrhizobium sp. BTAi1*	13
Acidovorax sp. JS42	8	*Bradyrhizobium sp. ORS278*	8
Acinetobacter baumannii	28	*Brucella abortus biovar 1 str. 9-941*	3
Acinetobacter baumannii ATCC 17978	10	*Brucella abortus S19*	3
Acinetobacter sp. ADP1	25	*Brucella canis ATCC 23365*	2
Actinobacillus pleuropneumoniae	1	*Brucella melitensis 16M*	3
Actinobacillus pleuropneumoniae L20	4	*Brucella melitensis biovar Abortus 2308*	3
Actinobacillus pleuropneumoniae serovar 3 str. JL03	5	*Brucella ovis ATCC 25840*	3
Actinobacillus succinogenes 130Z	1	*Brucella suis 1330*	3
Aeromonas hydrophila subsp. hydrophila ATCC 7966	10	*Brucella suis ATCC 23445*	3
Aeromonas salmonicida subsp. salmonicida A449	10	*Burkholderia ambifaria AMMD*	15
Agrobacterium tumefaciens str. C58	8	*Burkholderia ambifaria MC40-6*	14
Alcanivorax borkumensis SK2	10	*Burkholderia cenocepacia AU 1054*	20
Alkalilimnicola ehrlichei MLHE-1	8	*Burkholderia cenocepacia HI2424*	23
Anabaena sp. PCC 7120	22	*Burkholderia cenocepacia MC0-3*	23
Anabaena variabilis ATCC 29413	10	*Burkholderia mallei ATCC 23344*	8
Anaeromyxobacter dehalogenans 2CP-C	2	*Burkholderia mallei NCTC 10229*	8
Anaeromyxobacter sp. Fw109-5	3	*Burkholderia mallei NCTC 10247*	8
Aquifex aeolicus VF5	1	*Burkholderia mallei SAVP1*	7
Arcobacter butzleri RM4018	17	*Burkholderia multivorans ATCC 17616*	14
Azoarcus sp. BH72	20	*Burkholderia phymatum STM815*	3
Azoarcus sp. EbN1	6	*Burkholderia pseudomallei 1106a*	10
Azorhizobium caulinodans ORS 571	13	*Burkholderia pseudomallei 1710b*	10
Bacteroides fragilis NCTC 9343	73	*Burkholderia pseudomallei 668*	10
Bacteroides fragilis YCH46	76	*Burkholderia pseudomallei K96243*	10
Bacteroides thetaiotaomicron VPI-5482	108	*Burkholderia sp. 383*	27
Bacteroides vulgatus ATCC 8482	74	*Burkholderia thailandensis E264*	10
Bartonella bacilliformis KC583	1	*Burkholderia vietnamiensis G4*	10
Bartonella henselae str. Houston-1	1	*Burkholderia xenovorans LB400*	13
Bartonella quintana str. Toulouse	1	*Campylobacter concisus 13826*	2
Bartonella tribocorum CIP 105476	1	*Campylobacter curvus 525.92*	9
Bdellovibrio bacteriovorus HD100	3	*Campylobacter fetus subsp. fetus 82-40*	4
Beijerinckia indica subsp. indica ATCC 9039	12	*Campylobacter jejuni RM1221*	2
Bordetella avium 197N	10	*Campylobacter jejuni subsp. doylei 269.97*	4
Bordetella bronchiseptica	2	*Campylobacter jejuni subsp. jejuni 81116*	2
Bordetella bronchiseptica RB50	18	*Campylobacter jejuni subsp. jejuni 81-176*	2
Bordetella parapertussis 12822	13	*Campylobacter jejuni subsp. jejuni NCTC 11168*	3
Bordetella pertussis	2	*Candidatus Blochmannia floridanus*	1
Bordetella pertussis Tohama I	15	*Candidatus Blochmannia pennsylvanicus str. BPEN*	1

Species	TBDTs
Caulobacter crescentus CB15	59
Caulobacter sp. K31	83
Chlorobium chlorochromatii CaD3	1
Chlorobium phaeobacteroides DSM 266	3
Chlorobium tepidum TLS	5
Chromobacterium violaceum ATCC 12472	12
Chromohalobacter salexigens DSM 3043	9
Citrobacter koseri ATCC BAA-895	22
Colwellia psychrerythraea 34H	35
Cupriavidus taiwanensis	2
Cyanothece sp. ATCC 51142	1
Cytophaga hutchinsonii ATCC 33406	9
Dechloromonas aromatica RCB	12
Delftia acidovorans SPH-1	37
Desulfococcus oleovorans Hxd3	1
Desulfotalea psychrophila LSv54	1
Desulfovibrio desulfuricans G20	1
Desulfovibrio vulgaris subsp. vulgaris DP4	1
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	1
Dinoroseobacter shibae DFL 12	3
Enterobacter sakazakii ATCC BAA-894	8
Enterobacter sp. 638	13
Erwinia carotovora subsp. atroseptica SCRI1043	18
Erwinia chrysanthemi	1
Erythrobacter litoralis HTCC2594	19
Escherichia coli	1
Escherichia coli 536	16
Escherichia coli APEC O1	15
Escherichia coli ATCC 8739	9
Escherichia coli CFT073	19
Escherichia coli E24377A	9
Escherichia coli HS	8
Escherichia coli K12	6
Escherichia coli O157:H7	1
Escherichia coli O157:H7 EDL933	14
Escherichia coli O157:H7 str. Sakai	13
Escherichia coli SECEC SMS-3-5	13
Escherichia coli str. K-12 substr. DH10B	10
Escherichia coli str. K-12 substr. MG1655	9
Escherichia coli UTI89	18
Escherichia coli W3110	9
Flavobacterium johnsoniae UW101	57
Flavobacterium psychrophilum JIP02/86	8
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	5
Geobacter metallireducens GS-15	2
Geobacter sulfurreducens PCA	1
Geobacter uraniireducens Rf4	2
Gloeobacter violaceus PCC 7421	32
Gluconacetobacter diazotrophicus PAl 5	20
Gluconobacter oxydans 621H	13
Gramella forsetii KT0803	28
Granulibacter bethesdensis CGDNIH1	5
Haemophilus ducreyi 35000HP	2
Haemophilus influenzae	1
Haemophilus influenzae 86-028NP	5
Haemophilus influenzae PittEE	4
Haemophilus influenzae PittGG	4
Haemophilus influenzae Rd KW20	8
Haemophilus somnus 129PT	5
Haemophilus somnus 2336	5
Hahella chejuensis KCTC 2396	8
Halorhodospira halophila SL1	8
Helicobacter acinonychis str. Sheeba	6
Helicobacter hepaticus ATCC 51449	4
Helicobacter pylori 26695	5
Helicobacter pylori HPAG1	6
Helicobacter pylori J99	6
Herminiimonas arsenicoxydans	9
Hyphomonas neptunium ATCC 15444	39
Idiomarina loihiensis L2TR	29
Jannaschia sp. CCS1	1
Janthinobacterium sp. Marseille	31
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	13
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Paris)'	5
Leptospira borgpetersenii serovar Hardjobovis JB197	6
Leptospira borgpetersenii serovar Hardjobovis L550	6
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	9
Leptospira interrogans serovar Lai str. 56601	11
Leptothrix cholodnii SP-6	7
Listonella anguillarum	2
Magnetococcus sp. MC-1	2
Magnetospirillum magneticum AMB-1	2
Mannheimia succiniciproducens MBEL55E	3
Maricaulis maris MCS10	21
Marinobacter aquaeolei VT8	4
Marinomonas sp. MWYL1	15
Mesorhizobium loti MAFF303099	1
Mesorhizobium sp. BNC1	3
Methylibium petroleiphilum PM1	15
Methylobacillus flagellatus KT	21
Methylobacterium extorquens PA1	15
Methylobacterium radiotolerans JCM 2831	20
Methylobacterium sp. 4-46	8
Methylococcus capsulatus str. Bath	5

Species	TBDTs
Moraxella catarrhalis	1
Morganella morganii	1
Myxococcus xanthus DK 1622	11
Neisseria gonorrhoeae	3
Neisseria gonorrhoeae FA 1090	6
Neisseria meningitidis	2
Neisseria meningitidis 053442	6
Neisseria meningitidis FAM18	10
Neisseria meningitidis MC58	9
Neisseria meningitidis Z2491	8
Nitratiruptor sp. SB155-2	2
Nitrobacter hamburgensis X14	11
Nitrobacter winogradskyi Nb-255	13
Nitrosococcus oceani ATCC 19707	8
Nitrosomonas europaea ATCC 19718	29
Nitrosomonas eutropha C91	12
Nitrosospira multiformis ATCC 25196	8
Nodularia spumigena CCY9414	2
Nostoc punctiforme PCC 73102	2
Novosphingobium aromaticivorans DSM 12444	66
Ochrobactrum anthropi ATCC 49188	8
Opitutus terrae PB90-1	8
Parabacteroides distasonis ATCC 8503	66
Paracoccus denitrificans PD1222	22
Parvibaculum lavamentivorans DS-1	7
Pasteurella multocida subsp. multocida str. Pm70	12
Pelobacter carbinolicus DSM 2380	6
Pelobacter propionicus DSM 2379	14
Pelodictyon luteolum DSM 273	3
Photobacterium profundum SS9	7
Photorhabdus luminescens subsp. laumondii TTO1	12
Polaromonas naphthalenivorans CJ2	3
Polaromonas sp. JS666	5
Polynucleobacter sp. QLW-P1DMWA-1	3
Porphyromonas gingivalis W83	7
Prosthecochloris vibrioformis DSM 265	2
Pseudoalteromonas atlantica T6c	62
Pseudoalteromonas haloplanktis TAC125	35
Pseudomonas aeruginosa	6
Pseudomonas aeruginosa PA7	30
Pseudomonas aeruginosa PAO1	35
Pseudomonas aeruginosa UCBPP-PA14	37
Pseudomonas entomophila L48	30
Pseudomonas fluorescens	1
Pseudomonas fluorescens Pf-5	43
Pseudomonas fluorescens PfO-1	26
Pseudomonas mendocina ymp	17
Pseudomonas putida	2
Pseudomonas putida F1	30
Pseudomonas putida GB-1	47
Pseudomonas putida KT2440	30
Pseudomonas putida W619	22
Pseudomonas putida WCS358	1
Pseudomonas sp. M114	1
Pseudomonas stutzeri A1501	13
Pseudomonas syringae pv. phaseolicola 1448A	21
Pseudomonas syringae pv. syringae B728a	19
Pseudomonas syringae pv. tomato str. DC3000	25
Psychrobacter arcticus 273-4	1
Psychrobacter cryohalolentis K5	4
Psychrobacter sp. PRwf-1	8
Psychromonas ingrahamii 37	1
Ralstonia eutropha H16	17
Ralstonia eutropha JMP134	10
Ralstonia metallidurans CH34	16
Ralstonia solanacearum GMI1000	15
Rhizobium etli CFN 42	2
Rhizobium leguminosarum bv. viciae 3841	3
Rhodobacter sphaeroides 2.4.1	4
Rhodobacter sphaeroides ATCC 17025	3
Rhodobacter sphaeroides ATCC 17029	7
Rhodoferax ferrireducens T118	3
Rhodopseudomonas palustris BisA53	8
Rhodopseudomonas palustris BisB18	10
Rhodopseudomonas palustris BisB5	7
Rhodopseudomonas palustris CGA009	18
Rhodopseudomonas palustris HaA2	17
Rhodospirillum rubrum ATCC 11170	13
Roseobacter denitrificans OCh 114	1
Saccharophagus degradans 2-40	43
Salinibacter ruber DSM 13855	17
Salmonella enterica	1
Salmonella enterica subsp. arizonae serovar `62:z4␣z23:--`	8
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	8
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	6
Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7	8
Salmonella enterica subsp. enterica serovar Typhi str. CT18	6
Salmonella enterica subsp. enterica serovar Typhi Ty2	6
Salmonella typhimurium LT2	8
Serratia marcescens	1
Serratia proteamaculans 568	16
Shewanella amazonensis SB2B	23
Shewanella baltica OS155	28
Shewanella baltica OS185	34
Shewanella baltica OS195	38

Species	TBDTs
Shewanella denitrificans OS217	21
Shewanella frigidimarina NCIMB 400	26
Shewanella halifaxensis HAW-EB4	18
Shewanella loihica PV-4	20
Shewanella oneidensis MR-1	23
Shewanella pealeana ATCC 700345	24
Shewanella putrefaciens CN-32	29
Shewanella sediminis HAW-EB3	21
Shewanella sp. ANA-3	34
Shewanella sp. MR-4	35
Shewanella sp. MR-7	35
Shewanella sp. W3-18-1	27
Shewanella woodyi ATCC 51908	39
Shigella boydii CDC 3083-94	8
Shigella boydii Sb227	8
Shigella dysenteriae	1
Shigella dysenteriae Sd197	8
Shigella flexneri 2a str. 2457T	4
Shigella flexneri 2a str. 301	4
Shigella flexneri 5 str. 8401	6
Shigella sonnei Ss046	10
Silicibacter sp. TM1040	1
Sinorhizobium medicae WSM419	4
Sinorhizobium meliloti	1
Sinorhizobium meliloti 1021	8
Solibacter usitatus Ellin6076	5
Sorangium cellulosum 'So ce 56'	10
Sphingomonas wittichii RW1	140
Sphingopyxis alaskensis RB2256	32
Sulfurimonas denitrificans DSM 1251	5
Sulfurovum sp. NBC37-1	2
Synechococcus sp. JA-2-3B'a(2-13)	2
Synechococcus sp. JA-3-3Ab	2
Synechococcus sp. PCC 7002	6
Synechocystis sp. PCC 6803	4
Syntrophobacter fumaroxidans MPOB	3
Syntrophus aciditrophicus SB	2
Thiobacillus denitrificans ATCC 25259	7
Thiomicrospira crunogena XCL-2	5
Verminephrobacter eiseniae EF01-2	6
Vibrio cholerae	4
Vibrio cholerae O1 biovar eltor str. N16961	6
Vibrio cholerae O395	9
Vibrio fischeri ES114	8
Vibrio harveyi ATCC BAA-1116	10
Vibrio parahaemolyticus RIMD 2210633	11
Vibrio vulnificus CMCP6	7
Vibrio vulnificus YJ016	7
Wolinella succinogenes DSM 1740	11
Xanthobacter autotrophicus Py2	3
Xanthomonas axonopodis pv. citri str. 306	66
Xanthomonas campestris pv. campestris str. 8004	65
Xanthomonas campestris pv. campestris str. ATCC 33913	64
Xanthomonas campestris pv. vesicatoria str. 85-10	52
Xanthomonas oryzae pv. oryzae KACC10331	31
Xanthomonas oryzae pv. oryzae MAFF 311018	33
Xylella fastidiosa 9a5c	9
Xylella fastidiosa M12	9
Xylella fastidiosa M23	9
Xylella fastidiosa Temecula1	9
Yersinia enterocolitica	4
Yersinia enterocolitica subsp. enterocolitica 8081	11
Yersinia pestis	2
Yersinia pestis Angola	11
Yersinia pestis Antiqua	11
Yersinia pestis biovar Microtus str. 91001	11
Yersinia pestis CO92	11
Yersinia pestis KIM	12
Yersinia pestis Nepal516	11
Yersinia pestis Pestoides F	11
Yersinia pseudotuberculosis IP 31758	11
Yersinia pseudotuberculosis IP 32953	12
Yersinia pseudotuberculosis YPIII	11
Zymomonas mobilis subsp. mobilis ZM4	17

Summary

The generation of biological sequence data has witnessed a massive reduction in time consumption and costs, yielding a true data flood. More than 63 million publicly available Expressed Sequence Tags (ESTs) and over 900 completely sequenced bacterial genomes are impressive examples of this development. This data now allows to re-address open standing questions concerning the evolution of species and of biological systems.

The research field of metazoan (animal) phylogeny for example particularly benefits from the massive sequencing of ESTs. The splitting events between the main animal lineages occurred hundreds of millions of years ago, leaving only a weak phylogenetic signal. To get a robust resolution of these splits, the signal has to be amplified by incorporating lots of data. Moreover, the broad variety of taxa for which ESTs are now available allows the determination of evolutionary relationships within the main lineages on a fine scale.

In this thesis, we provide an introduction to sequence data (Chapter 1) and their application in phylogeny reconstruction (Chapter 2). We further give a detailed description of ESTs and explain why a processing of these is necessary before they can be used in phylogenetic analyses (Chapter 3). We introduce a pipeline to automatically process millions of raw ESTs (Chapter 4), and evaluate its accuracy (Chapter 5). We further present a database, called dbDMP, which is dedicated to host sequence data generated by processing ESTs. Additionally, we provide insights into the results of clustering ESTs in terms of sequence quality improvement (Chapter 6). In Chapter 7, we describe a method to compile customized sets of orthologous sequences for EST-based phylogeny reconstruction, and demonstrate its application with an investigation of the evolutionary relationships of winged insects (Pterygota) (Chapter 8). Furthermore, we present our finding that common gene selection strategies for EST-based phylogeny reconstruction introduce a bias towards slowly evolving genes. We subsequently investigate the consequences of this bias for the inferred phylogenies (Chapter 9).

By contrast, the protein family of TonB dependent transporters (TBDTs) is an ideal framework to explore the evolution of biological systems. Exclusively found in gram-negative bacteria, they provide passage for several nutrients through the cell wall with high substrate specificity. By searching for homologs to previously characterized TBDTs in almost 700 species, we obtained about 4,600 new candidates from ∼350 taxa. A subsequent clustering analysis revealed a complex system of 195 subclasses within this

family. By labeling the subclasses according to known TBDTs, we were able to suggest putative substrates for ∼3,700 of the 4,600 tentative transporters. Interestingly, TBDTs are grouped by their substrates rather than by the taxonomy of species they are found in. Finally, we present an intuitive web interface that grants access of our results to the research community (Chapter 10)

Both studies demonstrate that by mastering obstacles introduced by the sheer amount of data, nowadays available sequence data provide the opportunity to reconstruct complex evolution on different levels.

Acknowledgments

First and foremost, I wish to thank my advisor Arndt von Haeseler for his advise, his collaboration and for giving me the possibility the work in the pleasant and stimulating environment of his research group.

Next, I would like to thank my co-advisor Ingo Ebersberger for helpful discussions, his collaboration, and for always pointing me into the right directions. I am grateful for his constant support and for carefully reading and commenting this manuscript.

I thank my collaborators at the Institute of Ecology & Evolution in Hannover, Sabrina Simon and Heike Hadrys, who introduced me to the fascinating field of the evolution of winged insects.

I also thank my collaborators Oliver Mirus, Kerstin Nicolaisen and Enrico Schleiff at the Institute for Molecular Biosciences in Frankfurt, who reminded me that evolution does not only take place on the species level.

Furthermore, I would like to thank Enrico Schleiff, as well as Thomas Hankeln, for accepting to referee my thesis.

I thank all my kind colleagues at the CIBIV for creating such an enjoyable and inspiring working environment. Two colleagues deserve special mention. Steffen Klaere for providing mathematical support and Ricardo de Matos Simoes for his help with collecting sequence data and fruitful discussions.

Special thanks go to Jayne Ewing for linguistic help.

I appreciate the financial support from the Deutsche Forschungsgemeinschaft and the Wiener Wissenschafts-, Forschungs- und Technologiefonds granted to Arndt von Haeseler.

I want to thank all my family and friends who supported me in any way. Finally, I cannot thank Nina Schädel enough for her support, her patience and simply for being there.

Curriculum Vitae

Sascha Strauß

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Dr. Bohr Gasse 9
A-1030 Wien

Email: sascha.strauss@univie.ac.at
Homepage: www.cibiv.at/∼sascha

Personal

- Born on August 1, 1979.
- German Citizen.

Education

- 2000–2007: Heinrich-Heine-University, Düsseldorf (Germany)
- 2007–2010: Center for Integrative Bioinformatics Vienna (CIBIV), Vienna (Austira)

Degree

- 2006 'Master of Science' in Biology (Diplom Biologe), Düsseldorf (Germany). Thesis: Computational approach to detect regions of varying ancestries in the human genome on a fine scale

Publications

Journal Articles

- C. Weber, A. Pickl-Herk, **S. Strauss**, O. Carugo and D. Blaas (2009) Predictive bioinformatic identification of minor receptor group human rhinoviruses. *FEBS Lett., 583, 2547-2551.* (DOI: 10.1016/j.febslet.2009.07.015, PMID: 19615999)
- I. Ebersberger, **S. Strauss**, and A. von Haeseler (2009) HaMStR: Profile Hidden Markov Model Based Search for Orthologs in ESTs. *BMC Evol. Biol., 9, 157.* (DOI: 10.1186/1471-2148-9-157, PMID: 19586527)
- S. Simons, **S. Strauss**, A. von Haeseler, and H. Hadrys (2009) A phylogenomic approach to resolve the basal pterygote divergence. *Mol. Biol. Evol. 26, 2719-2730.* (DOI: 10.1093/molbev/msp191, PMID: 19713325)
- O. Mirus*,**S. Strauss***, K. Nicolaisen, A. von Haeseler, E. Schleiff (2009) TonB-dependent transporters and their occurrence in cyanobacteria. *BMC Biol., 7, 68.* (DOI: 10.1186/1741-7007-7-68, PMID: 19821963) - * contributed equally

Bibliography

Abascal, F., Zardoya, R. and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Actis, L. A., Tolmasky, M. E., Farrell, D. H. and Crosa, J. H. (1988) Genetic and molecular characterization of essential components of the vibrio anguillarum plasmid-mediated iron-transport system. *J Biol Chem*, **263**, 2853–2860.

Adams, D. G. and Duggan, P. S. (1999) Heterocyst and akinete differentiation in cyanobacteria. *New phytologist*, **144**, 3–33.

Adams, M., Kelley, J., Gocayne, J., Dubnick, M., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R. and al. et (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Adolph, K. W. and Haselkorn, R. (1971) Isolation and characterization of a virus infecting the blue-green alga nostoc muscorum. *Virology*, **46**, 200–208.

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. and Lake, J. A. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, **387**, 489–493.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2007) *Molecular Biology of the Cell*. Garland Science, Fifth edn..

Alice, A. F., López, C. S., Lowe, C. A., Ledesma, M. A. and Crosa, J. H. (2006) Genetic and transcriptional analysis of the siderophore malleobactin biosynthesis and transport genes in the human pathogen burkholderia pseudomallei k96243. *J Bacteriol*, **188**, 1551–1566.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. and Ronquist, F. (2004) Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.

Andrews, S. C., Robinson, A. K. and Rodríguez-Quiñones, F. (2003) Bacterial iron homeostasis. *FEMS Microbiology Reviews*, **27**, 215–237.

Ankenbauer, R. G. and Quan, H. N. (1994) FptA, the Fe(III)-pyochelin receptor of pseudomonas aeruginosa: a phenolate siderophore receptor homologous to hydroxamate siderophore receptors. *J Bacteriol*, **176**, 307–319.

Atkins, E. D. and Sheehan, J. K. (1971) The molecular structure of hyaluronic acid. *Biochemical Journal*, **125**, 92P, PMCID: PMC1178299.

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M. and Philippe, H. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, entamoeba, and mastigamoeba. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1414–1419.

Baurain, D., Brinkmann, H. and Philippe, H. (2007) Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? *Mol Biol Evol*, **24**, 6–9.

Beall, B. and Hoenes, T. (1997) An iron-regulated outer-membrane protein specific to bordetella bronchiseptica and homologous to ferric siderophore receptors. *Microbiology*, **143 (Pt 1)**, 135–145.

Beall, B. and Sanden, G. N. (1995) A bordetella pertussis fepA homologue required for utilization of exogenous ferric enterobactin. *Microbiology*, **141 (Pt 12)**, 3193–3205.

Bechly, G., Brauckmann, C., Zessin, W. and Gröning, E. (2001) New results concerning the morphology of the most ancient dragonflies (Insecta: odonatoptera) from the namurian of Hagen-Vorhalle (Germany). *Journal of Zoological Systematics & Evolutionary Research*, **39**, 209–226.

Belfiore, N. M., Liu, L. and Moritz, C. (2008) Multilocus phylogenetics of a rapid radiation in the genus thomomys (Rodentia: geomyidae). *Syst Biol*, **57**, 294–310.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2009) GenBank. *Nucl. Acids Res.*, **37**, D26–31.

Berglund, A., Sjolund, E., Ostlund, G. and Sonnhammer, E. L. L. (2007) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucl. Acids Res.*, page gkm1020.

Beucher, M. and Sparling, P. (1995) Cloning, sequencing, and characterization of the gene encoding FrpB, a major iron-regulated, outer membrane protein of neisseria gonorrhoeae. *J. Bacteriol.*, **177**, 2041–2049.

Beutel, R. G. . and Pohl, H. . (2006) Endopterygote systematics where do we stand and what is the goal (Hexapoda, arthropoda)?: REVIEW. *Systematic Entomology*, **31**, 202–219.

Beutel, R. G. and Gorb, S. N. (2001) Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. *Journal of Zoological Systematics and Evolutionary Research*, **39**, 177–207.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and genomewise. *Genome Research*, **14**, 988–995.

Bister, B., Bischoff, D., Nicholson, G. J., Valdebenito, M., Schneider, K., Winkelmann, G., Hantke, K. and Süssmuth, R. D. (2004) The structure of salmochelins: C-glucosylated enterobactins of salmonella enterica. *Biometals*, **17**, 471–481.

Biswas, G. D. and Sparling, P. F. (1995) Characterization of lbpA, the structural gene for a lactoferrin receptor in neisseria gonorrhoeae. *Infect Immun*, **63**, 2958–2967.

Bitter, W., Marugg, J. D., de Weger, L. A., Tommassen, J. and Weisbeek, P. J. (1991) The ferric-pseudobactin receptor PupA of pseudomonas putida WCS358: homology to TonB-dependent escherichia coli receptors and specificity of the protein. *Mol Microbiol*, **5**, 647–655.

Blanvillain, S., Meyer, D., Boulanger, A., Lautier, M., Guynet, C., Denancé, N., Vasse, J., Lauber, E. and Arlat, M. (2007) Plant carbohydrate scavenging through tonb-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS One*, **2**, e224.

Bleidorn, C., Podsiadlowski, L., Zhong, M., Eeckhaut, I., Hartmann, S., Halanych, K. and Tiedemann, R. (2009) On the phylogenetic position of myzostomida: can 77 genes get it wrong? *BMC Evolutionary Biology*, **9**, 150.

Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993) dbEST - database for "expressed sequence tags". *Nat Genet*, **4**, 332–333.

Bonaldo, M. F., Lennon, G. and Soares, M. B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, **6**, 791–806.

Boudreaux, H. B. (1979) *Arthropod phylogeny with special reference to insects*. Wiley, New York.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E. S., Thorndyke, M., Nakano, H., Kohn, A. B., Heyland, A., Moroz, L. L., Copley, R. R. and Telford, M. J. (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*, **444**, 85–88.

Braun, V. and Killmann, H. (1999) Bacterial solutions to the iron-supply problem. *Trends Biochem Sci*, **24**, 104–109.

Brickman, T. J. and Armstrong, S. K. (1999) Essential role of the iron-regulated outer membrane receptor FauA in alcaligin siderophore-mediated iron uptake in bordetella species. *J Bacteriol*, **181**, 5958–5966.

Brinkmann, H., van der Giezen, M., Zhou, Y., de Raucourt, G. P. and Philippe, H. (2005) An empirical assessment of Long-Branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*, **54**, 743–757.

Brito, B., Aldon, D., Barberis, P., Boucher, C. and Genin, S. (2002) A signal transfer system through three compartments transduces the plant cell contact-dependent signal controlling ralstonia solanacearum hrp genes. *Mol Plant Microbe Interact*, **15**, 109–119.

Butterton, J. R., Stoebner, J. A., Payne, S. M. and Calderwood, S. B. (1992) Cloning, sequencing, and transcriptional regulation of viuA, the gene encoding the ferric vibriobactin receptor of vibrio cholerae. *J Bacteriol*, **174**, 3729–3738.

Bäumler, A. J. and Hantke, K. (1992) Ferrioxamine uptake in yersinia enterocolitica: characterization of the receptor protein FoxA. *Mol Microbiol*, **6**, 1309–1321.

Cai, Y. P. and Wolk, C. P. (1990) Use of a conditionally lethal gene in anabaena sp. strain PCC 7120 to select for double recombinants and to entrap insertion sequences. *J Bacteriol*, **172**, 3138–3145.

Carson, S. D., Klebba, P. E., Newton, S. M. and Sparling, P. F. (1999) Ferric enterobactin binding and utilization by neisseria gonorrhoeae. *J Bacteriol*, **181**, 2895–2901.

Carswell, C. L., Rigden, M. D. and Baenziger, J. E. (2008) Expression, purification, and structural characterization of CfrA, a putative iron transporter from campylobacter jejuni. *J Bacteriol*, **190**, 5650–5662.

Cartron, M., Maddocks, S., Gillingham, P., Craven, C. and Andrews, S. (2006) Feo – transport of ferrous iron into bacteria. *BioMetals*, **19**, 143–157.

Chen, Y., Lin, C., Wang, C., Wu, H. and Hwang, P. (2007) An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics*, **8**, 416.

Cheng, Q., Yu, M. C., Reeves, A. R. and Salyers, A. A. (1995) Identification and characterization of a bacteroides gene, csuF, which encodes an outer membrane protein that is essential for growth on chondroitin sulfate. *J Bacteriol*, **177**, 3721–3727.

Chou, H. and Holmes, M. H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.

Clarke, S. E., Stuart, J. and Sanders-Loehr, J. (1987) Induction of siderophore activity in anabaena spp. and its moderation of copper toxicity. *Appl Environ Microbiol*, **53**, 917–922.

Clarke, T. E., Tari, L. W. and Vogel, H. J. (2001) Structural biology of bacterial iron uptake systems. *Curr Top Med Chem*, **1**, 7–30.

Comas, I., Moya, A. and González-Candelas, F. (2007) Phylogenetic signal and functional categories in proteobacteria genomes. *BMC Evolutionary Biology*, **7 Suppl 1**, S7, PMID: 17288580.

Cope, L. D., Yogev, R., Muller-Eberhard, U. and Hansen, E. J. (1995) A gene cluster involved in the utilization of both free heme and heme:hemopexin by haemophilus influenzae type b. *J Bacteriol*, **177**, 2644–2653.

Cornelissen, C. N., Biswas, G. D., Tsai, J., Paruchuri, D. K., Thompson, S. A. and Sparling, P. F. (1992) Gonococcal transferrin-binding protein 1 is required for transferrin utilization and is homologous to TonB-dependent outer membrane receptors. *J. Bacteriol.*, **174**, 5788–5797.

Cotton, J. A. and Wilkinson, M. (2009) Supertrees join the mainstream of phylogenetics. *Trends in Ecology & Evolution*, **24**, 1–3.

Coulton, J. W., Mason, P., Cameron, D. R., Carmel, G., Jean, R. and Rode, H. N. (1986) Protein fusions of beta-galactosidase to the ferrichrome-iron receptor of escherichia coli k-12. *J Bacteriol*, **165**, 181–192.

Cox, C. D., Rinehart, K. L., Moore, M. L. and Cook, J. C. (1981) Pyochelin: novel structure of an iron-chelating growth promoter for pseudomonas aeruginosa. *Proc Natl Acad Sci U S A*, **78**, 4256–4260.

Darwin, C. (1859) *On the origin of species by means of natural selection*. John Murray, First edn..

Date, C. J. (1999) *An Introduction to Database Systems*. Addison Wesley Longman, 7th edn..

Dean, C. R. and Poole, K. (1993) Cloning and characterization of the ferric enterobactin receptor gene (pfeA) of pseudomonas aeruginosa. *J Bacteriol*, **175**, 317–324.

Delsuc, F., Brinkmann, H. and Philippe, H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, **6**, 361–375.

Dhungana, S., White, P. S. and Crumbliss, A. L. (2001) Crystal structure of ferrioxamine b: a comparative analysis and implications for molecular recognition. *J Biol Inorg Chem*, **6**, 810–818.

Dorsey, C. W., Tomaras, A. P., Connerly, P. L., Tolmasky, M. E., Crosa, J. H. and Actis, L. A. (2004) The siderophore-mediated iron acquisition systems of acinetobacter baumannii ATCC 19606 and vibrio anguillarum 775 are structurally and functionally related. *Microbiology*, **150**, 3657–3667.

Drechsel, H., Metzger, J., Freund, S., Jung, G., Boelaert, J. R. and Winkelmann, G. (1991) Rhizoferrin — a novel siderophore from the fungusRhizopus microsporus var.rhizopodiformis. *BioMetals*, **4**, 238–243.

Drechsel, H., Stephan, H., Lotz, R., Haag, H., Zähner, H., Hantke, K. and Jung, G. (1995) Structure elucidation of yersiniabactin, a siderophore from highly virulent yersinia strains. *Liebigs Annalen*, **1995**, 1727–1733.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. and Arnold, F. H. (2005) Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14338–14343.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q. and Giribet, G. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Durbin, R. (1998a) *Biological sequence analysis*. Cambridge University Press.

Durbin, R. (1998b) Profile HMMs for sequence families. In *Biological sequence analysis*, pages 100–133, Cambridge University Press.

Dybas, J. M., Madrid-Aliste, C. J., Che, F., Rykunov, D., Angeletti, R. H., Weiss, L. M., Kim, K. and Fiser, A. (2008) Computational analysis and experimental validation of gene predictions in toxoplasma gondii. *PLoS ONE*, **3**, e3899.

Ebersberger, I., Gube, M., Strauss, S., Kupczok, A., Eckart, M., Voigt, K., Kothe, E. and von Haeseler, A. (2009a) A stable backbone for the fungi. http://precedings.nature.com/documents/2901/version/1.

Ebersberger, I., Strauss, S. and von Haeseler, A. (2009b) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, **9**, 157.

Edwards, S. V., Liu, L. and Pearl, D. K. (2007) High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, **104**, 5936–5941.

Engel, M. S. and Grimaldi, D. A. (2004) New light shed on the oldest insect. *Nature*, **427**, 627–630.

Ewing, B. and Green, P. (1998) Base-Calling of automated sequencer traces using phred. - II. error probabilities. *Genome Research*, **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-Calling of automated sequencer traces using phred. - i. accuracy assessment. *Genome Research*, **8**, 175–185.

Feinberg, A. P. and Vogelstein, B. (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochemistry*, **132**, 6–13, PMID: 6312838.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1989) PHYLIP - phylogeny inference package (Version 3.2). *Cladistics*, **5**, 166, 164.

Ferguson, A. D., Chakraborty, R., Smith, B. S., Esser, L., van der Helm, D. and Deisenhofer, J. (2002) Structural basis of gating by the outer membrane transporter FecA. *Science*, **295**, 1715–1719.

Ferguson, A. D. and Deisenhofer, J. (2002) TonB-dependent receptors-structural perspectives. *Biochim Biophys Acta*, **1565**, 318–332.

Ferreira, F. and Straus, N. (1994) Iron deprivation in cyanobacteria. *Journal of Applied Phycology*, **6**, 199–210.

Fetherston, J. D., Lillard, J. W. and Perry, R. D. (1995) Analysis of the pesticin receptor from yersinia pestis: role in iron-deficient growth and possible regulation by its siderophore. *J Bacteriol*, **177**, 1824–1833.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A. (2008) The pfam protein families database. *Nucleic Acids Res*, **36**, D281–D288.

Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst Biol*, **19**, 99–113.

Fitch, W. M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.

Francke, C., Siezen, R. J. and Teusink, B. (2005) Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, **13**, 550–558, PMID: 16169729.

Frickey, T. and Lupas, A. (2004) CLANS: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.

Furano, K., Luke, N. R., Howlett, A. J. and Campagnari, A. A. (2005) Identification of a conserved moraxella catarrhalis haemoglobin-utilization protein, MhuA. *Microbiology*, **151**, 1151–1158.

Gadagkar, S. R., Rosenberg, M. S. and Kumar, S. (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology. Part B. Molecular and Developmental Evolution*, **304**, 64–74, PMID: 15593277.

Gatesy, J., Baker, R. H. and Hayashi, C. (2004) Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of crocodylia. *Systematic Biology*, **53**, 342–355, PMID: 15205058.

Gaunt, M. W. and Miles, M. A. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol*, **19**, 748–761.

Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P. and Rokas, A. (2009) Benchmarking Next-Generation transcriptome sequencing for functional and evolutionary genomics. *Mol Biol Evol*, **26**, 2731–2744.

Gibson, F. and Magrath, D. I. (1969) The isolation and characterization of a hydroxamic acid (aerobactin) formed by aerobacter aerogenes 62-I. *Biochim Biophys Acta*, **192**, 175–184.

Giribet, G. (2008) Assembling the lophotrochozoan (=spiralian) tree of life. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**, 1513–22, PMID: 18192183.

Goldberg, M. B., Boyko, S. A., Butterton, J. R., Stoebner, J. A., Payne, S. M. and Calderwood, S. B. (1992) Characterization of a vibrio cholerae virulence factor homologous to the family of TonB-dependent proteins. *Molecular Microbiology*, **6**, 2407–2418, PMID: 1406279.

Goldman, N. (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Goldman, S. J., Lammers, P. J., Berman, M. S. and Sanders-Loehr, J. (1983) Siderophore-mediated iron uptake in different strains of anabaena sp. *J Bacteriol*, **156**, 1140–1150.

Graur, D. and Li, W. (2000) *Fundamentals of Molecular Evolution*. Sinauer Associates, Second edn..

Graveley, B. R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, **17**, 100–107.

Greca, M. L. (1980) Origin and evolution of wings and flight in insects. *Bulletin of Zoology*, **47**, 65–82.

Greene, J. J. and Rao, V. B. (1998) Recombinant DNA principles and methodologies. pages 354–356, CRC Press.

Griffiths, G. L., Sigel, S. P., Payne, S. M. and Neilands, J. B. (1984) Vibriobactin, a siderophore from vibrio cholerae. *Journal of Biological Chemistry*, **259**, 383–385.

Grimaldi, D. A. and Engel, M. S. (2005) *Evolution of the insects*. Cambridge University Press.

Gu, X., Fu, Y. and Li, W. (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol*, **12**, 546–557.

Guglielmi, G., Cohen-Bazire, G. and Bryant, D. A. (1981) The structure of gloeobacter violaceus and its phycobilisomes. *Archives of Microbiology*, **129**, 181–189.

Guikema and Sherman (1983) Organization and function of chlorophyll in membranes of cyanobacteria during iron starvation. *Plant Physiol*, **73**, 250–256.

Gumbart, J., Wiener, M. C. and Tajkhorshid, E. (2007) Mechanics of force propagation in TonB-Dependent outer membrane transport. *Biophysical Journal*, **93**, 496–504.

Halanych, K. M. (2004) THE NEW VIEW OF ANIMAL PHYLOGENY. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 229–256.

Hantke, K. (1990) Dihydroxybenzolyserine–a siderophore for e. coli. *FEMS Microbiology Letters*, **67**, 5–8.

Hantke, K., Nicholson, G., Rabsch, W. and Winkelmann, G. (2003) Salmochelins, siderophores of salmonella enterica and uropathogenic escherichia coli strains, are recognized by the outer membrane receptor IroN. *Proc Natl Acad Sci U S A*, **100**, 3677–3682.

Hausdorf, B., Helmkampf, M., Meyer, A., Witek, A., Herlyn, H., Bruchhaus, I., Hankeln, T., Struck, T. H. and Lieb, B. (2007) Spiralian phylogenomics supports the resurrection of bryozoa comprising ectoprocta and entoprocta. *Mol Biol Evol*, **24**, 2723–2729.

Hecht, J., Kuhl, H., Haas, S., Bauer, S., Poustka, A., Lienau, J., Schell, H., Stiege, A., Seitz, V., Reinhardt, R., Duda, G., Mundlos, S. and Robinson, P. (2006) Gene identification and analysis of transcripts differentially regulated in fracture healing by EST sequencing in the domestic sheep. *BMC Genomics*, **7**, 172.

Heller, K. and Kadner, R. J. (1985) Nucleotide sequence of the gene for the vitamin b12 receptor protein in the outer membrane of escherichia coli. *J. Bacteriol.*, **161**, 904–908.

Helmkampf, M., Bruchhaus, I. and Hausdorf, B. (2008) Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the lophotrochozoa concept. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 1927–1933.

Henderson, D. P. and Payne, S. M. (1994) Characterization of the vibrio cholerae outer membrane heme transport protein HutA: sequence of the gene, regulation of expression, and homology to the family of TonB-dependent proteins. *Journal of Bacteriology*, **176**, 3269–3277, PMID: 8195082.

Hennig, W. (1950) *Grundzüge Einer Theorie Der Phylogenetischen Systematik*. Deutscher zentralverlag, Berlin.

Hennig, W. (1969) *Die Stammesgeschichte der Insekten.* Kramer, Frankfurt am Main.

Hennig, W. (1981) *Insect phylogeny.* John Wiley & Sons, Bath, UK.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K. and Marra, M. (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807–828.

Hillis, D. M. (1987) Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics*, **18**, 23–42.

Hillis, D. M., Pollock, D. D., McGuire, J. A. and Zwickl, D. J. (2003) Is sparse taxon sampling a problem for phylogenetic inference? *Systematic Biology*, **52**, 124–126, PMID: 12554446.

HOFFMAN, P., LINKER, A., LIPPMAN, V. and MEYER, K. (1960) The structure of chondroitin sulfate b from studies with flavobacterium enzymes. *The Journal of Biological Chemistry*, **235**, 3066–3069, PMID: 13715032.

Hornung, J. M., Jones, H. A. and Perry, R. D. (1996) The hmu locus of yersinia pestis is essential for utilization of free haemin and haem–protein complexes as iron sources. *Molecular Microbiology*, **20**, 725–739, PMID: 9026634.

Hovmöller, R., Pape, T. and Källersjö, M. (2002) The palaeoptera problem: Basal pterygote phylogeny inferred from 18S and 28S rDNA sequences. *Cladistics*, **18**, 313–323.

Huang, F., Parmryd, I., Nilsson, F., Persson, A. L., Pakrasi, H. B., Andersson, B. and Norling, B. (2002) Proteomics of synechocystis sp. strain PCC 6803: identification of plasma membrane proteins. *Molecular & Cellular Proteomics: MCP*, **1**, 956–966, PMID: 12543932.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Huelsenbeck, J. P. and Ronquist, F. (2001) MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Hughes, J., Longhorn, S. J., Papadopoulou, A., Theodorides, K., de Riva, A., Mejia-Chang, M., Foster, P. G. and Vogler, A. P. (2006) Dense taxonomic EST sampling and its applications for molecular systematics of the coleoptera (Beetles). *Mol Biol Evol*, **23**, 268–278.

Huson, D. H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, **23**, 254–267.

Hutber, G., Hutson, K. and Rogers, L. (1977) Effect of iron dificiency on levels of two ferredoxins and flavodoxin in a cyanobacterium. *FEMS Microbiology Letters*, **1**, 193–196.

Irimia, M., Maeso, I., Penny, D., Garcia-Fernandez, J. and Roy, S. W. (2007) Rare coding sequence changes are consistent with ecdysozoa, not coelomata. *Mol Biol Evol*, **24**, 1604–1607.

Jeanjean, R., Talla, E., Latifi, A., Havaux, M., Janicki, A. and Zhang, C. (2008) A large gene cluster encoding peptide synthetases and polyketide synthases is involved in production of siderophores and oxidative stress response in the cyanobacterium anabaena sp. strain PCC 7120. *Environ Microbiol*, **10**, 2574–2585.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006) Phylogenomics: the beginning of incongruence? *Trends in Genetics*, **22**, 225–231.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, **110**, 462–7, PMID: 16093699.

Kaestner, A. (2003) *Lehrbuch der speziellen Zoologie*. G. Fischer.

Kahnert, A. and Kertesz, M. A. (2000) Characterization of a sulfur-regulated oxygenative alkylsulfatase from pseudomonas putida s-313. *J Biol Chem*, **275**, 31661–31667.

Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M. and Tabata, S. (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium anabaena sp. strain PCC 7120. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, **8**, 205–213; 227–253, PMID: 11759840.

Karlsen, S. and Hough, E. (1995) Crystal structures of three complexes between chito-oligosaccharides and lysozyme from the rainbow trout. how distorted is the NAG sugar in site d? *Acta Crystallographica Section D*, **51**, 962–978.

Katoh, H., Hagino, N., Grossman, A. R. and Ogawa, T. (2001) Genes essential to iron transport in the cyanobacterium synechocystis sp. strain PCC 6803. *J Bacteriol*, **183**, 2779–2784.

Katoh, K., ichi Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.*, **33**, 511–518.

Keller-Schierlein, W. and Prelog, V. (1961) Stoffwechselprodukte von actinomyceten. 29. mitteilung. die konstitution des ferrioxamins d1. *Helvetica Chimica Acta*, **44**, 709–713.

Kent, W. J. (2002) BLAT—The BLAST-Like alignment tool. *Genome Research*, **12**, 656–664.

Keren, N., Aurora, R. and Pakrasi, H. B. (2004) Critical roles of bacterioferritins in iron storage and proliferation of cyanobacteria. *Plant Physiol*, **135**, 1666–1673.

Kjer, K. M. (2004) Aligned 18S and insect phylogeny. *Syst Biol*, **53**, 506–514.

Kjer, K. M., Carle, F. L., Litman, J. and Ware, J. (2006) A molecular phylogeny of hexapoda. *Arthropod Systematics and Phylogeny*, **64**, 35–44.

Koebnik, R. (2005) TonB-dependent trans-envelope signalling: the exception or the rule? *Trends in Microbiology*, **13**, 343–347.

Koebnik, R., Hantke, K. and Braun, V. (1993) The TonB-dependent ferrichrome receptor FcuA of yersinia enterocolitica: evidence against a strict co-evolution of receptor structure and substrate specificity. *Mol Microbiol*, **7**, 383–393.

Kolaczkowski, B. and Thornton, J. W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

Koster, M., van de Vossenberg, J., Leong, J. and Weisbeek, P. J. (1993) Identification and characterization of the pupB gene encoding an inducible ferric-pseudobactin receptor of pseudomonas putida WCS358. *Mol Microbiol*, **8**, 591–601.

Kristensen, N. P. *et al.* (1991) Phylogeny of extant hexapods. *The insects of Australia*, **1**, 125–140.

Krizman, D. B., Wagner, L., Lash, A., Strausberg, R. L. and Emmert-Buck, M. R. (1999) The cancer genome anatomy project: EST sequencing and the genetics of cancer progression. *Neoplasia (New York, N.Y.)*, **1**, 101–106, PMC1508126.

Krone, W. J., Stegehuis, F., Koningstein, G., Doorn, C., Roosendaal, B., Graaf, F. K. and Oudega, B. (1985) Characterization of the pColV-K30 encoded cloacin DF13/aerobactin outer membrane receptor protein of escherichia coli; isolation and purification of the protein and analysis of its nucleotide sequence and primary structure. *FEMS Microbiology Letters*, **26**, 153–161.

Kukalova-Peck, J. (1991) Fossil history and the evolution of hexapod structures. *The insects of Australia*, **1**, 141–179.

Kukalová-Peck, J. and Lawrence, J. F. (2004) Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters. *European Journal of Entomology*, **101**, 95–144.

Kunze, B., Bedorf, N., Kohl, W., Höfle, G. and Reichenbach, H. (1989) Myxochelin a, a new iron-chelating compound from angiococcus disciformis (Myxobacterales). production, isolation, physico-chemical and biological properties. *J Antibiot (Tokyo)*, **42**, 14–17.

Kuramae, E., Robert, V., Echavarri-Erasun, C. and Boekhout, T. (2007) Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. *BMC Evolutionary Biology*, **7**, 134.

Kustka, A., Carpenter, E. J. and nudo-Wilhelmy, S. A. S. (2002) Iron and marine nitrogen fixation: progress and future directions. *Res Microbiol*, **153**, 255–262.

Kühn, S., Braun, V. and Köster, W. (1996) Ferric rhizoferrin uptake into morganella morganii: characterization of genes involved in the uptake of a polyhydroxycarboxylate siderophore. *J Bacteriol*, **178**, 496–504.

Latifi, A., Jeanjean, R., Lemeille, S., Havaux, M. and Zhang, C. (2005) Iron starvation leads to oxidative stress in anabaena sp. strain PCC 7120. *J Bacteriol*, **187**, 6596–6598.

Lee, B. C. (1995) Quelling the red menace: haem capture by bacteria. *Molecular Microbiology*, **18**, 383–390.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., Antonescu, V., Chan, A., Cheung, F. and Quackenbush, J. (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucl. Acids Res.*, **33**, D71–74.

Letoffe, S., Ghigo, J. M. and Wandersman, C. (1994) Iron acquisition from heme and hemoglobin by a serratia marcescens extracellular protein. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 9876–9880.

LeVier, K. and Guerinot, M. L. (1996) The bradyrhizobium japonicum fegA gene encodes an iron-regulated outer membrane protein with similarity to hydroxamate-type siderophore receptors. *J Bacteriol*, **178**, 7265–7275.

Lewis, L. A., Gray, E., Wang, Y. P., Roe, B. A. and Dyer, D. W. (1997) Molecular characterization of hpuAB, the haemoglobin-haptoglobin-utilization operon of neisseria meningitidis. *Mol Microbiol*, **23**, 737–749.

Li, L., Stoeckert, C. J. and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.

Linnaeus, C. (1735) *Systema Naturae 1735*.

Liu, L. and Pearl, D. K. (2007) Species trees from gene trees: Reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*, **56**, 504–514.

Llamas, M. A., Sparrius, M., Kloet, R., Jiménez, C. R., Vandenbroucke-Grauls, C. and Bitter, W. (2006) The heterologous siderophores ferrioxamine b and ferrichrome activate signaling pathways in pseudomonas aeruginosa. *J Bacteriol*, **188**, 1882–1891.

Lundrigan, M. D. and Kadner, R. J. (1986) Nucleotide sequence of the gene for the ferrienterochelin receptor FepA in escherichia coli. homology among outer membrane receptors that interact with TonB. *J Biol Chem*, **261**, 10797–10801.

Lynch, D., O'Brien, J., Welch, T., Clarke, P., Cuív, P. O., Crosa, J. H. and O'Connell, M. (2001) Genetic organization of the region encoding regulation, biosynthesis, and transport of rhizobactin 1021, a siderophore produced by sinorhizobium meliloti. *J Bacteriol*, **183**, 2576–2585.

Mallatt, J. and Giribet, G. (2006) Further use of nearly complete 28S and 18S rRNA genes to classify ecdysozoa: 37 more arthropods and a kinorhynch. *Molecular Phylogenetics and Evolution*, **40**, 772–794.

Martynov, A. B. (1925) Über zwei grundtypen der flügel bei den insecten und ihre evolution. *Zoomorphology*, **4**, 465–501.

Massalski, A., Laube, V. M. and Kushner, D. J. (1981) Effects of cadmium and copper on the ultrastructure ofAnkistrodesmus braunii andAnabaena 7120. *Microbial Ecology*, **7**, 183–193.

Matsuda, R. (1970) *Morphology and evolution of the insect thorax.* (Ottawa).

Matsuda, R. (1981) The origin of insect wings (Arthropoda: insecta). *International Journal of Insect Morphology and Embryology*, **10**, 387–398.

May, R. M. (1988) How many species are there on earth? *Science (New York, N.Y.)*, **241**, 1441–1449, PMID: 17790039.

Mazoy, R., Osorio, C. R., Toranzo, A. E. and Lemos, M. L. (2003) Isolation of mutants of vibrio anguillarum defective in haeme utilisation and cloning of huvA, a gene coding for an outer membrane protein involved in the use of haeme as iron source. *Arch Microbiol*, **179**, 329–338.

McCombie, W. R., Adams, M. D., Kelley, J. M., FitzGerald, M. G., Utterback, T. R., Khan, M., Dubnick, M., Kerlavage, A. R., Venter, J. C. and Fields, C. (1992) Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. *Nat Genet*, **1**, 124–131.

Mey, A. R. and Payne, S. M. (2001) Haem utilization in vibrio cholerae involves multiple TonB-dependent haem receptors. *Mol Microbiol*, **42**, 835–849.

Mey, A. R., Wyckoff, E. E., Hoover, L. A., Fisher, C. R. and Payne, S. M. (2008) Vibrio cholerae VciB promotes iron uptake via ferrous iron transporters. *J Bacteriol*, **190**, 5953–5962.

Mey, A. R., Wyckoff, E. E., Oglesby, A. G., Rab, E., Taylor, R. K. and Payne, S. M. (2002) Identification of the vibrio cholerae enterobactin receptors VctA and IrgA: IrgA is not required for virulence. *Infect. Immun.*, **70**, 3419–3426.

Meyer, T. E., Cusanovich, M. A. and Kamen, M. D. (1986) Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 217–220.

Miethke, M. and Marahiel, M. A. (2007) Siderophore-Based iron acquisition and pathogen control. *Microbiol. Mol. Biol. Rev.*, **71**, 413–451.

Mills, M. and Payne, S. M. (1997) Identification of shuA, the gene encoding the heme receptor of shigella dysenteriae, and analysis of invasion and intracellular multiplication of a shuA mutant. *Infect Immun*, **65**, 5358–5363.

Minh, B. Q., Vinh, L. S., von Haeseler, A. and Schmidt, H. A. (2005) pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, **21**, 3794–3796.

Mirus, O.*, Strauss, S.*, Nicolaisen, K., von Haeseler, A. and Schleiff, E. (2009) TonB-dependent transporters and their occurrence in cyanobacteria. *BMC Biology*, **7**, 68. *equal contribution

Misof, B. and Misof, K. (2009) A monte carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst Biol*, page syp006.

Misof, B., Niehuis, O., Bischoff, I., Rickert, A., Erpenbeck, D. and Staniczek, A. (2007) Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology*, **110**, 409–429.

Morris, J., Donnelly, D. F., O'Neill, E., McConnell, F. and O'Gara, F. (1994) Nucleotide sequence analysis and potential environmental distribution of a ferric pseudobactin receptor gene of pseudomonas sp. strain m114. *Mol Gen Genet*, **242**, 9–16.

Moslavac, S., Bredemeier, R., Mirus, O., Granvogl, B., Eichacker, L. A. and Schleiff, E. (2005) Proteomic analysis of the outer membrane of anabaena sp. strain PCC 7120. *J Proteome Res*, **4**, 1330–1338.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, **51 Pt 1**, 263–273, PMID: 3472723.

Müller, T. and Vingron, M. (2000) Modeling amino acid replacement. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **7**, 761–776, PMID: 11382360.

Nagano, K., Murakami, Y., Nishikawa, K., Sakakibara, J., Shimozato, K. and Yoshimura, F. (2007) Characterization of RagA and RagB in porphyromonas gingivalis: study using gene-deletion mutants. *J Med Microbiol*, **56**, 1536–1548.

Nagaraj, S. H., Gasser, R. B. and Ranganathan, S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform*, **8**, 6–21.

Nakamura, T. M., Morin, G. B., Chapman, K. B., Weinrich, S. L., Andrews, W. H., Lingner, J., Harley, C. B. and Cech, T. R. (1997) Telomerase catalytic subunit homologs from fission yeast and human. *Science*, **277**, 955–959.

Nau, C. D. and Konisky, J. (1989) Evolutionary relationship between the TonB-dependent outer membrane transport proteins: nucleotide and amino acid sequences of the escherichia coli colicin i receptor gene. *J Bacteriol*, **171**, 1041–1047.

Neugebauer, H., Herrmann, C., Kammer, W., Schwarz, G., Nordheim, A. and Braun, V. (2005) ExbBD-dependent transport of maltodextrins through the novel MalA protein across the outer membrane of caulobacter crescentus. *J Bacteriol*, **187**, 8300–8311.

Niall, H. D., Hirs, C. H. W. and Timasheff, S. N. (1973) Automated edman degradation: The protein sequenator. In *Part D: Enzyme Structure*, vol. Volume 27, pages 942–1010, Academic Press.

Nicolaisen, K., Moslavac, S., Samborski, A., Valdebenito, M., Hantke, K., Maldener, I., Muro-Pastor, A. M., Flores, E. and Schleiff, E. (2008) Alr0397 is an outer membrane transporter for the siderophore schizokinen in anabaena sp. strain PCC 7120. *J Bacteriol*, **190**, 7500–7505.

Ninomiya, T. . and Yoshizawa, K. (2009) A revised interpretation of the wing base structure in odonata. *Systematic Entomology*, **34**, 334–345.

Nishio, T., Tanaka, N., Hiratake, J., Katsube, Y., Ishida, Y. and Oda, J. (1988) Isolation and structure of the novel dihydroxamate siderophore alcaligin. *Journal of the American Chemical Society*, **110**, 8733–8734.

Ochsner, U. A., Johnson, Z. and Vasil, M. L. (2000) Genetics and regulation of two distinct haem-uptake systems, phu and has, in pseudomonas aeruginosa. *Microbiology*, **146 (Pt 1)**, 185–198.

Ochsner, U. A. and Vasil, M. L. (1996) Gene repression by the ferric uptake regulator in pseudomonas aeruginosa: cycle selection of iron-regulated genes. *Proc Natl Acad Sci U S A*, **93**, 4409–4414.

Ogden, T. and Whiting, M. F. (2003) The problem with "the paleoptera problem:" sense and sensitivity. *Cladistics*, **19**, 432–442.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet*, **2**, 173–179.

Opperdoes, F. R. (2003) Phylogenetic analysis using protein sequences. In *The phylogenetic handbook*, pages 207–235, Cambridge University Press.

Pauer, H., de Oliveira Ferreira, E., dos Santos-Filho, J., Portela, M. B., Zingali, R. B., Soares, R. M. A. and Domingues, R. M. C. P. (2009) A TonB-dependent outer membrane protein as a bacteroides fragilis fibronectin-binding molecule. *FEMS Immunol Med Microbiol*, **55**, 388–395.

Perkins-Balding, D., Ratliff-Griffin, M. and Stojiljkovic, I. (2004) Iron transport systems in neisseria meningitidis. *Microbiol. Mol. Biol. Rev.*, **68**, 154–171.

Persmark, M., Expert, D. and Neilands, J. B. (1989) Isolation, characterization, and synthesis of chrysobactin, a compound with siderophore activity from erwinia chrysanthemi. *J Biol Chem*, **264**, 3187–3193.

Persmark, M., Pittman, P., Buyer, J. S., Schwyn, B., Gill, P. R. and Neilands, J. B. (1993) Isolation and structure of rhizobactin 1021, a siderophore from the alfalfa symbiont rhizobium meliloti 1021. *Journal of the American Chemical Society*, **115**, 3950–3956.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. and Quackenbush, J. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Philippe, H. (2000) Opinion: Long branch attraction and protist phylogeny. *Protist*, **151**, 307–316.

Philippe, H., Delsuc, F., Brinkmann, H. and Lartillot, N. (2005a) Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 541–562.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Silva, C. D., Wincker, P., Guyader, H. L., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. and Manuel, M. (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, **19**, 706–712.

Philippe, H., Lartillot, N. and Brinkmann, H. (2005b) Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol Biol Evol*, **22**, 1246–1253.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W. H. and Casane, D. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular biology and evolution*, **21**, 1740–52, PMID: 15175415.

Phillips, M. J., Delsuc, F. and Penny, D. (2004) Genome-Scale phylogeny and the detection of systematic biases. *Mol Biol Evol*, **21**, 1455–1458.

Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A. and Boyce-Jacino, M. (1999) Mining SNPs from EST databases. *Genome Research*, **9**, 167–174.

Pollack, J. R. and Neilands, J. B. (1970) Enterobactin, an iron transport compound from salmonella typhimurium. *Biochemical and Biophysical Research Communications*, **38**, 989–992, PMID: 4908541.

Poole, K., Neshat, S., Krebes, K. and Heinrichs, D. E. (1993) Cloning and nucleotide sequence analysis of the ferripyoverdine receptor gene fpvA of pseudomonas aeruginosa. *J. Bacteriol.*, **175**, 4597–4604.

Pradel, E. and Locht, C. (2001) Expression of the putative siderophore receptor gene bfrZ is controlled by the extracytoplasmic-function sigma factor BupI in bordetella bronchiseptica. *J Bacteriol*, **183**, 2910–2917.

Pressler, U., Staudenmaier, H., Zimmermann, L. and Braun, V. (1988) Genetics of the iron dicitrate transport system of escherichia coli. *J. Bacteriol.*, **170**, 2716–2724.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., de Jong, P. J., Jurka, J., Kapitonov, V. V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L. A., Salamov, A. A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L. Z., Holland, P. W. H., Satoh, N. and Rokhsar, D. S. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Putney, S. D., Herlihy, W. C. and Schimmel, P. (1983) A new troponin t and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature*, **302**, 718–721.

Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucl. Acids Res.*, **28**, 141–145.

Rakin, A., Saken, E., Harmsen, D. and Heesemann, J. (1994) The pesticin receptor of yersinia enterocolitica: a novel virulence factor with dual function. *Mol Microbiol*, **13**, 253–263.

Rambaut, A. and Drummond, A. J. (2007) Tracer v1. 4. *Available from http://beast. bio. ed. ac. uk/Tracer*.

Reeves, A., D'Elia, J., Frias, J. and Salyers, A. (1996) A bacteroides thetaiotaomicron outer membrane protein that is essential for utilization of maltooligosaccharides and starch. *J. Bacteriol.*, **178**, 823–830.

Regier, J. C. and Shultz, J. W. (1998) Molecular phylogeny of arthropods and the significans of the cambrian explosion for molecular systematics. *American zoologist*, **38**, 918–928.

Remm, M., Storm, C. E. V. and Sonnhammer, E. L. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**, 1041–1052.

Ren, Q. and Paulsen, I. T. (2005) Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comput Biol*, **1**, e27.

von Reumont, B., Meusemann, K., Szucsich, N., Ampio, E. D., Gowri-Shankar, V., Bartel, D., Simon, S., Letsch, H., Stocsits, R., xia Luan, Y., Waegele, J., Pass, G., Hadrys, H. and Misof, B. (2009) Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? a case study on major arthropod relationships. *BMC Evolutionary Biology*, **9**, 119.

Roberts, J., Bebenek, K. and Kunkel, T. (1988) The accuracy of reverse transcriptase from HIV-1. *Science*, **242**, 1171–1173.

Rodionov, D. A., Hebbeln, P., Gelfand, M. S. and Eitinger, T. (2006) Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J Bacteriol*, **188**, 317–327.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2002) Comparative genomics of thiamin biosynthesis in procaryotes. new genes and regulatory mechanisms. *J Biol Chem*, **277**, 48949–48959.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2003) Comparative genomics of the vitamin b12 metabolism and regulation in prokaryotes. *Journal of Biological Chemistry*, **278**, 41148–41159.

Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R. and Burmester, T. (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator). *Molecular Phylogenetics and Evolution*, **53**, 826–834.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., von Haeseler, A., Kube, M., Reinhardt, R. and Burmester, T. (2007) EST sequencing of onychophora and phylogenomic analysis of metazoa. *Molecular Phylogenetics and Evolution*, **45**, 942–951.

Roger, P. A., Tirol, A., Ardales, S. and Watanabe, I. (1986) Chemical composition of cultures and natural samples of n2-fixing blue-green algae from rice fields. *Biology and Fertility of Soils*, **2**, 131–146.

Rokas, A. and Carroll, S. B. (2006) Bushes in the tree of life. *PLoS Biology*, **4**, e352, PMID: 17105342.

Rokas, A., King, N., Finnerty, J. and Carroll, S. B. (2003a) Conflicting phylogenetic signals at the base of the metazoan tree. *Evolution & Development*, **5**, 346–359, PMID: 12823451.

Rokas, A., Kruger, D. and Carroll, S. B. (2005) Animal evolution and the molecular signature of radiations compressed in time. *Science*, **310**, 1933–1938.

Rokas, A., Williams, B. L., King, N. and Carroll, S. B. (2003b) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

Ronquist, F. and Huelsenbeck, J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Roth, A., Gonnet, G. and Dessimoz, C. (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, **9**, 518.

Rudd, S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science*, **8**, 321–329.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Baguñà, J. and Riutort, M. (2002) A phylogenetic analysis of myosin heavy chain type II sequences corroborates that acoela and nemertodermatida are basal bilaterians. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11246–11251.

Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**, 441–446.

Sauer, M., Hantke, K. and Braun, V. (1990) Sequence of the fhuE outer-membrane receptor gene of escherichia coli k12 and properties of mutants. *Mol Microbiol*, **4**, 427–437.

Sauvage, C., Franza, T. and Expert, D. (1996) Analysis of the erwinia chrysanthemi ferrichrysobactin receptor gene: resemblance to the escherichia coli fepA-fes bidirectional promoter region and homology with hydroxamate receptors. *J Bacteriol*, **178**, 1227–1231.

Savard, J., Tautz, D., Richards, S., Weinstock, G. M., Gibbs, R. A., Werren, J. H., Tettelin, H. and Lercher, M. J. (2006) Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. *Genome Research*, **16**, 1334–1338.

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. and Wolfe, K. H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341–345.

Schauer, K., Gouget, B., Carrière, M., Labigne, A. and de Reuse, H. (2007) Novel nickel transport mechanism across the bacterial outer membrane energized by the TonB/ExbB/ExbD machinery. *Mol Microbiol*, **63**, 1054–1068.

Schauer, K., Rodionov, D. A. and de Reuse, H. (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends in Biochemical Sciences*, **33**, 330–338.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Schuster, S. C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18, PMID: 18165802.

Shcolnick, S. and Keren, N. (2006) Metal homeostasis in cyanobacteria and chloroplasts. balancing benefits and risks to the photosynthetic apparatus. *Plant Physiol*, **141**, 805–810.

Shimodaira, H. and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–1247.

Silakowski, B., Kunze, B., Nordsiek, G., Blöcker, H., Höfle, G. and Müller, R. (2000) The myxochelin iron transport regulon of the myxobacterium stigmatella aurantiaca sg a15. *Eur J Biochem*, **267**, 6476–6485.

Simon, S., Strauss, S., von Haeseler, A. and Hadrys, H. (2009) A phylogenomic approach to resolve the basal pterygote divergence. *Mol Biol Evol*, **26**, 2719–2730.

Simpson, F. B. and Neilands, J. B. (1976) Siderochromes in cyanophyceae: isolation and characterization of schizokinen from anabaena sp. *Journal of Phycology*, **12**, 44–48.

Singh, A. K., McIntyre, L. M. and Sherman, L. A. (2003) Microarray analysis of the Genome-Wide response to iron deficiency and iron reconstitution in the cyanobacterium synechocystis sp. PCC 6803. *Plant Physiol.*, **132**, 1825–1839.

Smit, A., Hubley, R. and Green, P. (1996) RepeatMasker open-3.0.

Sokol, P. A., Darling, P., Lewenza, S., Corbett, C. R. and Kooi, C. D. (2000) Identification of a siderophore receptor required for ferric ornibactin uptake in burkholderia cepacia. *Infect Immun*, **68**, 6554–6560.

Spiegelman, S., Watson, K. F. and Kacian, D. L. (1971) Synthesis of DNA complements of natural RNAs: a general approach. *Proceedings of the National Academy of Sciences of the United States of America*, **68**, 2843–2845.

Srikumar, R., Mikael, L. G., Pawelek, P. D., Khamessan, A., Gibbs, B. F., Jacques, M. and Coulton, J. W. (2004) Molecular cloning of haemoglobin-binding protein HgbA in the outer membrane of actinobacillus pleuropneumoniae. *Microbiology*, **150**, 1723–1734.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Staniczek, A. H. (2000) The mandible of silverfish (Insecta: zygentoma) and mayflies (Ephemeroptera): its morphology and phylogenetic significance. *Zoologischer Anzeiger*, **239**, 147–178.

Steel, M. (2005) Should phylogenetic models be trying to 'fit an elephant'? *Trends in Genetics*, **21**, 307–309.

Steel, M. and Rodrigo, A. (2008) Maximum likelihood supertrees. *Syst Biol*, **57**, 243–250.

Stephan, H., Freund, S., Beck, W., Jung, G., Meyer, J. M. and Winkelmann, G. (1993) Ornibactins–a new family of siderophores from pseudomonas. *Biometals*, **6**, 93–100.

Stojijkovic, I., Hwa, V., Martin, L. S., O'Gaora, P., Nassif, X., Heffron, F. and So, M. (1995) The neisseria meningitidis haemoglobin receptor: its role in iron utilization and virulence. *Molecular Microbiology*, **15**, 531–541.

Stojiljkovic, I. and Hantke, K. (1992) Hemin uptake system of yersinia enterocolitica: similarities with other TonB-dependent systems in gram-negative bacteria. *EMBO J*, **11**, 4359–4367.

Strimmer, K. and von Haeseler, A. (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 6815–6819.

Struck, T. H. and Fisse, F. (2008) Phylogenetic position of nemertea derived from phylogenomic data. *Mol Biol Evol*, **25**, 728–736.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 6062–6067, PMID: 15075390.

Swofford, D. L. (2002) PAUP*: phylogenetic analysis using parsimony (* and other methods). version 4.0 b. *Sinauer, Sunderland, Massachusetts, USA*.

Targon, M. L. P. N., Takita, M. A., do Amaral, A. M., de Souza, A. A., Locali-Fabris, E. C., de Oliveira Dorta, S., Borges, K. M., de Souza, J. M., Rodrigues, C. M., Lucheta, A. R., Freitas-Astúa, J. and Machado, M. A. (2007) CitEST libraries. *Genetics and Molecular Biology*, **30**.

Teintze, M. and Leong, J. (1981) Structure of pseudobactin a, a second siderophore from plant growth promoting pseudomonas b10. *Biochemistry*, **20**, 6457–6462.

Temin, H. and Mizutani, S. (1970) Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of rous sarcoma virus. *Nature*, **226**, 1211–1213.

Thomas, C. E., Olsen, B. and Elkins, C. (1998) Cloning and characterization of tdhA, a locus encoding a TonB-dependent heme receptor from haemophilus ducreyi. *Infect Immun*, **66**, 4254–4262.

Ting, C. S., Rocap, G., King, J. and Chisholm, S. W. (2002) Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends Microbiol*, **10**, 134–142.

de la Torre, J., Egan, M., Katari, M., Brenner, E., Stevenson, D., Coruzzi, G. and DeSalle, R. (2006) ESTimating plant phylogeny: lessons from partitioning. *BMC Evolutionary Biology*, **6**, 48.

Torres, A. G. and Payne, S. M. (1997) Haem iron-transport system in enterohaemorrhagic escherichia coli O157:H7. *Mol Microbiol*, **23**, 825–833.

Townsend, J. P. (2007) Profiling phylogenetic informativeness. *Syst Biol*, **56**, 222–231.

van Uitert, M., Meuleman, W. and Wessels, L. (2008) Biclustering sparse binary genomic data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **15**, 1329–1345, PMID: 19040367.

Valentine, J. W. (2004) *On the origin of phyla*. University of Chicago Press.

Walker, J. M. and Rapley, R. (2000) Molecular biology and biotechnology. pages 80–91, Royal Society of Chemistry.

Wandersman, C. and Delepelaire, P. (2004) BACTERIAL IRON SOURCES: from siderophores to hemophores. *Annual Review of Microbiology*, **58**, 611–647.

Wei, B., Dalwadi, H., Gordon, L. K., Landers, C., Bruckner, D., Targan, S. R. and Braun, J. (2001) Molecular cloning of a bacteroides caccae TonB-linked outer membrane protein identified by an inflammatory bowel disease marker antibody. *Infect Immun*, **69**, 6044–6054.

Wheeler, W. C., Whiting, M., Wheeler, Q. D. and Carpenter, J. M. (2001) The phylogeny of the extant hexapod orders. *Cladistics*, **17**, 113–169.

Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a Maximum-Likelihood approach. *Mol Biol Evol*, **18**, 691–699.

Whitfield, J. B. and Kjer, K. M. (2008) Ancient rapid radiations of insects: Challenges for phylogenetic analysis. *Annual Review of Entomology*, **53**, 449–472.

Wiens, J. J., Kuczynski, C. A., Smith, S. A., Mulcahy, D. G., Sites, J. W., Townsend, T. M. and Reeder, T. W. (2008) Branch lengths, support, and congruence: Testing the phylogenomic approach with 20 nuclear loci in snakes. *Syst Biol*, **57**, 420–431.

Wilkinson, M., Cotton, J. A., Lapointe, F. and Pisani, D. (2007) Properties of supertree methods in the consensus setting. *Systematic Biology*, **56**, 330–337.

Willkommen, J. and Hörnschemeyer, T. (2007) The homology of wing base sclerites and flight muscles in ephemeroptera and neoptera and the morphology of the pterothorax of habroleptoides confusa (Insecta: ephemeroptera: Leptophlebiidae). *Arthropod Structure & Development*, **36**, 253–269.

Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J. and Mallatt, J. (2002) Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol Biol Evol*, **19**, 762–776.

Witek, A., Herlyn, H., Meyer, A., Boell, L., Bucher, G. and Hankeln, T. (2008) EST based phylogenomics of syndermata questions monophyly of eurotatoria. *BMC Evolutionary Biology*, **8**, 345.

Wuest, W. M., Sattely, E. S. and Walsh, C. T. (2009) Three siderophores from one bacterial enzymatic assembly line. *J Am Chem Soc*, **131**, 5056–5057.

Yang, C., Rodionov, D. A., Li, X., Laikova, O. N., Gelfand, M. S., Zagnitko, O. P., Romine, M. F., Obraztsova, A. Y., Nealson, K. H. and Osterman, A. L. (2006) Comparative genomics and experimental characterization of n-acetylglucosamine utilization pathway of shewanella oneidensis. *J Biol Chem*, **281**, 29872–29885.

Yang, H. M., Chaowagul, W. and Sokol, P. A. (1991) Siderophore production by pseudomonas pseudomallei. *Infect Immun*, **59**, 776–780.

Yoneyama, H. and Nakae, T. (1996) Protein c (OprC) of the outer membrane of pseudomonas aeruginosa is a copper-regulated channel protein. *Microbiology*, **142 (Pt 8)**, 2137–2144.

Yoshizawa, K. . and Saigusa, T. . (2001) Phylogenetic analysis of paraneopteran orders (Insecta: neoptera) based on forewing base structure, with comments on monophyly of auchenorrhyncha (Hemiptera). *Systematic Entomology*, **26**, 1–13.

Yoshizawa, K. and Johnson, K. P. (2005) Aligned 18S for zoraptera (Insecta): phylogenetic position and molecular evolution. *Molecular Phylogenetics and Evolution*, **37**, 572–580.

Zalkin, A., Forrester, J. D. and Templeton, D. H. (1964) Crystal and molecular structure of ferrichrome a. *Science*, **146**, 261–263.

Zeng, L. and Swalla, B. J. (2005) Molecular phylogeny of the protochordates: chordate evolution. *Canadian Journal of Zoology*, **83**, 24–33.

Zhang, J., Zhou, C., Gai, Y., Song, D. and Zhou, K. (2008) The complete mitochondrial genome of parafronurus youi (Insecta: ephemeroptera) and phylogenetic position of the ephemeroptera. *Gene*, **424**, 18–24.

Zhang, L. and Li, W. (2004) Mammalian housekeeping genes evolve more slowly than Tissue-Specific genes. *Mol Biol Evol*, **21**, 236–239.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **7**, 203–214, PMID: 10890397.

Zmasek, C. M. and Eddy, S. R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14, PMID: 12028595.

Zwickl, D. J. (2006) *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* Ph.D. thesis, The University of Texas at Austin.

Zwickl, D. J. and Hillis, D. M. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol*, **51**, 588–598.