The quest for orthologs, the tree of basal animals, and taxonomic profiles of metagenomes

Dissertation
for the award of the degree
"Doctor rerum naturalium" (Dr. rer. nat.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

submitted by
Fabian Schreiber
from Kassel

Göttingen 2010

Prof. Dr. Burkhard Morgenstern (1st Referee)
Abteilung für Bioinformatik, Institut für Mikrobiologie und Genetik,
Universität Göttingen

Prof. Dr. Lutz Walter (2nd Referee)
Abteilung Primatengenetik, Deutsches Primatenzentrum,
Göttingen

Prof. Dr. Gert Wörheide
Department für Geo- und Umweltwissenschaften,
Ludwig-Maximilians-Universität München

Date of the oral examination: June 25th, 2010

Affidavit

I hereby insure that I wrote this PhD thesis independently and with no other sources and aids than quoted.

Fabian Schreiber

May, 2010
Göttingen, Germany

List of Publications

Papers in Peer Reviewed Journals

- ◇ Hervé Philippe, Romain Derelle, Philippe Lopez, Kerstin S. Pick, Carole Borchiellini, Nicole Boury-Esnault, Jean Vacelet, Emmanuelle Deniel, Evelyn Houliston, Eric Quéinnec, Corinne Da Silva, Patrick Wincker, Hervé Le Guyader, Sally Leys, Daniel J. Jackson, Bernard M. Degnan, **Fabian Schreiber**, Dirk Erpenbeck, Burkhard Morgenstern, Gert Wörheide and Michael Manuel.
Phylogenomics restores traditional views on deep animal relationships.
Current Biology (2009) 19, 706-712.

- ◇ **Fabian Schreiber**, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A web server for selecting orthologous gene alignments from EST sequences.
Nucleic Acids Research (2009) 37, W185-W188.

- ◇ **Fabian Schreiber**, Kerstin S. Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A protocol for selecting orthologous groups in phylogenomics.
BMC Bioinformatics (2009) 10, 219.

- ◇ Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke.
HIV Classification using Coalescent Theory.
Bioinformatics (2010), doi:10.1093/bioinformatics/btq159.

- ◇ Kerstin S. Pick[1], Hervé Philippe[1], **Fabian Schreiber**, Dirk Erpenbeck, Daniel J. Jackson, Petra Wrede, Mathias Wiens, Alexandre Alié, Burkhard Morgenstern, Michael Manuel and Gert Wörheide.
Broader phylogenomic sampling improves the accuracy of non-bilaterian relationships.
Molecular biology and evolution (2010), doi:10.1093/molbev/msq089.

[1] These authors contributed equally

◇ **Fabian Schreiber**, Peter Gumrich, Rolf Daniel and Peter Meinicke.
Treephyler: fast taxonomic profiling of metagenomes.
Bioinformatics (2010), 26(7):960-961.

Posters at Conferences

◇ Katharina J. Hoff, **Fabian Schreiber**, Maike Tech, Peter Meinicke.
The effect of sequencing errors on metagenomic gene prediction.
Presented at the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB) 2009, Stockholm, Sweden.

◇ **Fabian Schreiber**, Kerstin S. Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A protocol for selecting ortholog groups in phylogenomics.
Presented at the Conference *Celebrating Darwin: From the Origin of Species to Deep Metazoan Phylogeny* (2009), Berlin, Germany.

◇ Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke.
Classification of HIV-1 Using Coalescent Theory.
Presented at the German Conference on Bioinformatics, GCB 2008, Dresden, Germany.

◇ Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke.
Classification of HIV-1 Using Coalescent Theory.
Presented at the European Conference on Computational Biology, ECCB 2008, Cagliari, Italy.

◇ **Fabian Schreiber**, Kerstin S. Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A protocol for selecting ortholog groups in phylogenomics.
Presented at the Göttingen Graduate School for Neurosciences and Molecular Biosciences opening (2008).

Acknowledgments

I much appreciate the support of my supervisors Professor Burkhard Morgenstern and Professor Gert Wörheide throughout the last three years. They have guided me through this thesis and gave me the freedom to work in their departments. Thanks to Professor Lutz Walter for being the second referee, as well as for support and helpful discussions. I like to thank Katharina Hoff for patiently reviewing my manuscripts and correcting my English language and grammar.

A special thanks goes to all my colleagues in the department of Professor Burkhard Morgenstern as well as in the department of Professor Gert Wörheide and in the department of Professor Edgar Wingender. It was a pleasure working with you and getting to know all of you.

The coordination office of the Göttingen Graduate School for Neurosciences and Molecular Biosciences was extremely helpful during all phases of my PhD studies and I want to thank them for their commitment.

Many thanks go to my parents, as well as my brother and grandmas for supporting me through my studies. Thanks go to my roommate Göck for sharing my passion for making pasta.

Last but not least, I would like to thank my beloved girlfriend Pascale for making my life so wonderful.

Contents

Abstract

This PhD study covers the development and application of methods in basal animal phylogenomics as well as the development of methods in metagenomics.

Studies of the tree of the basal animals Cnidaria, Ctenophora (comb-jellies), Porifera, and Placozoa are - despite the use large sets of DNA sequences - equivocal and are in conflict with traditional phylogenies based on morphological data. A resolved tree allows implications about the early evolution of the animal bauplan. New methods as well as enriched taxon sampling are needed to test existing hypotheses and to come to a consensus regarding the tree of basal animals as well as whether sponges are monophyletic or not. Using existing methods for similarity search, EST translation, sequence alignment and filtering of noisy characters in alignments, a new pipeline will automatically construct large-scale datasets for phylogenetic studies. In this work, we developed the new method *OrthoSelect* that - for the first time - automatically constructs datasets suitable for phylogenetic studies on a large scale. We assembled and analysed two large-scale datasets with enriched taxon sampling for basal animals and more sophisticated outgroup selection. *OrthoSelect* is generally applicable to all taxonomic groups and therefore a valuable tool for all phylogenetic large-scale studies. Our studies could further support the hypothesis that sponges are a monophyletic phylum. However, the studies were unequivocal concerning the relationships of basal animals due to cases of undetected hemiplasy (gene tree/species tree conflict).

In the field of metagenomics, the study of unculturable microorganisms, new methods are needed for constructing taxonomic profiles that scales with the increased size of datasets from next generation sequencing technologies and large-scale studies. A new method based on PFAM assignments allows the computation of taxonomic profiles from large metagenomic datasets. We developed the new tool *Treephyler* for taxonomic profiling of metagenomes. It is as accurate as existing methods, but ~ 10 times faster. This makes *Treephyler* the first tool that is ready to handle large datasets as e.g. in the study to explore the human microbiome.

Chapter 1

Introduction: Phylogenomics and issues of basal animal evolution

> **"The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of nature."**
>
> Charles Darwin, 1857

1.1 From Darwin to the *Tree of Life*

A prerequisite for almost any evolutionary study is the understanding of the phylogenetic relationships between organisms. All evolutionary history of genes and contemporary species is related by a phylogenetic tree (Page *et al.*, 1988). This directly follows from the famous evolutionary theory of natural selection presented by *Charles Darwin* in *The Origin of Species* (Darwin, 1859). In that book, the evolutionary relationships between organisms were, for the first time, illustrated as a phylogenetic tree. The publication of *Ernst Haeckel's* famous tree in 1866 corroborates the enthusiasm of biologists in phylogenies (Haeckel, 1866).

Today, phylogenetic trees are drawn from features of recent species using methods that rely on mathematical models. The basis for reconstructing the evolutionary history of species is the identification of homologous characters[1] that the different organisms share. These characters are then compared and reconstruction methods are used to construct a phylogenetic tree. The accuracy of the tree strongly depends on how well the evolutionary history is described by the mathematical model. Because the underlying biological mechanisms are not yet well

[1] There is no generally agreed-upon definition of a character in phylogenetics. However, a character can be thought of as an attribute, which can be used to distinguish taxa (e.g. the shape of teeth or an amino acid).

understood, these models do not have a sufficient fit. This makes the reconstruction of the *Tree of Life* (Maddison *et al.*, 2007) - the ultimate goal of phylogenetics - a difficult task.

Phylogenetics - trees from morphological and molecular data

The 1970s brought the development of tools to sequence DNA and proteins. Until that time, phylogenetics was based on the analysis of morphological or ultrastructural data only. Using the comparative anatomy of fossils and recent species, the main groups of plants and animals could be separated. However, the limited number of available reliable morphological characters restricts the ability to get highly resolved trees for parts of the *Tree of Life*. Morphological characters that can be used to distinguish species are almost not present in microorganisms and are limited in complex organisms, e.g. animals (van Niel, 1955).

The emergence of molecular data in phylogenetic studies promised to improve the resolving power and, by this, to overcome the limitation of insufficient comparable characters (Zuckerkandl and Pauling, 1965). In the course of phylogenetic research, some genes proved to be more suitable than others to reconstruct trees, making them reference markers. One of these markers is the gene that encodes the small ribosomal subunit RNA (SSU rRNA). Investigations using the SSU rRNA gene shaped the tree of Bacteria and Archaea in the 1980s (Woese, 1987). It also led to the recognition of Archaea as a third distinct domain of life (Woese and Fox, 1977).

Different genes - different answers

To further increase the resolution of the *Tree of Life*, researchers started using several genes rather than only single genes. However, it turned out that the resolving power is still limited and usually allows to obtain firm support for some parts of a phylogeny, only. This is especially true for the deepest inner nodes in a phylogeny. They are very old and therefore supported by less characters and more likely to undergo multiple substitutions. Furthermore, the analysis of different genes revealed rather different and often contradictory tree topologies.

Consequently, large parts of the *Tree of Life* remain unresolved due to the limited amount of data, while other parts - for which enough data are available - are unresolved because of incongruencies between the genes under study. These incongruencies are widespread and have been reported to occur on all taxonomic ranges; between closely related species (Kopp and True, 2002; Mason-Gamer and Kellogg, 1996), major classes (Giribet *et al.*, 2001; Hwang *et al.*, 2001) or phyla (Löytynoja and Milinkovitch, 2001; Rokas *et al.*, 2003a).

1.2 Phylogenomics - the more data the better

"... a classification founded on any single character, however important that may be, has always failed."

Charles Darwin, 1857

Based on the hypothesis that the more data is used the more likely it is to get the correct tree (Miyamoto and Fitch, 1995), newly developed sequencing techniques were used to generate thousands of base pairs in the time it requires to read these lines (Delsuc *et al.*, 2005). This wealth of sequencing information lead to a new branch of molecular phylogenetics, called phylogenomics (Eisen and Fraser, 2003), which tries to apply phylogenetic methods on genomic-scale data. The availability of this immense mass of data could decrease the impact of factors that cause incongruencies (Eisen and Fraser, 2003).

1.2.1 A lot of different genes - still different answers

Reasons for incongruencies are the limited data availability (Cummings *et al.*, 1995), the use of inappropriate taxa (taxon sampling) (Graybeal, 1998), inadequate modeling of sequence evolution (Yang *et al.*, 1994), as well as biological processes such as natural selection or genetic drift (Maddison, 1997; Martin and Burg, 2002; Satta *et al.*, 2000). The year 2000 marked a shift from single gene to multi-gene analyses, with studies using up to 20 genes predating the era of phylogenomics (Baldauf *et al.*, 2000; Madsen *et al.*, 2001; Murphy *et al.*, 2001; Stiller and Hall, 1997; Qiu *et al.*, 1999)

Following this trend to increase the size of datasets, studies using large sets of genes were published focussing on e.g. the phylogeny of deuterostomes (Bourlat *et al.*, 2006), tunicates (Delsuc *et al.*, 2006) or ecdysozoa (Philippe *et al.*, 2005). However, the limitation of taxon sampling to model organisms allowed to resolve some parts of the *Tree of Life*, only. The use of expressed sequence tags (EST) opened up new prospects.

1.2.2 Large datasets of EST sequences

Many recent phylogenomic studies are based on EST sequences (Bapteste *et al.*, 2002; Philippe *et al.*, 2004, 2005). EST sequences are short ($\approx$ 200 - 800 bases), unedited, randomly selected single-pass reads from cDNA libraries that sample the diversity of genes expressed by an organism or tissue at a particular time and under particular conditions.

The relatively low cost and rapid generation of ESTs led to studies using more than 100 genes and a broader taxonomic spectrum (Bapteste *et al.*, 2002; Blair *et al.*, 2002; Lerat *et al.*, 2003; Rokas *et al.*, 2003b; Wolf *et al.*, 2004).

1.2.3 Large datasets, but wrong answers

In general, the use of large datasets results in increased resolution of phylogenetic trees. Additionally, the phylogenetic methods used to construct trees are statistically consistent (Delsuc *et al.*, 2005). That means that analyses converge towards the correct tree as dataset size increase. This is true as long as basic assumptions are met, but failure to do so can lead to inconsistencies (Felsenstein, 2004). Cases when phylogenetic reconstruction methods can produce inconsistent results are:

Compositional bias In cases where species possess a similar sequence composition, phylogenetic methods can group them together, although they are not closely related.

Long branch attraction A common problem to phylogenetic methods is that fast evolving unrelated taxa can be artefactually grouped together and lead to wrong trees (Felsenstein, 1978). An example is the long branch attraction artefact in Philippe *et al.* (2005) where the fast evolving species *C. elegans* is attracted by the distant fungal outgroup *S. cerevisiae*, but correctly groups with *D. melanogaster* when the fungi outgroup is replaced by a more closely related choanoflagellate outgroup.

Heterotachy A character or alignment site is called heterotachous, if its evolutionary rate varies through time. Heterotachy is an essential process of sequence evolution and can lead to wrong trees (Lopez *et al.*, 2002; Kolaczkowski and Thornton, 2004). Heterotachy is difficult to detect as its presence cannot be judged from simply looking at the sequences (Inagaki *et al.*, 2004; Kolaczkowski and Thornton, 2004; Philippe and Germot, 2000).

These factors can lead to highly supported trees that are not guaranteed to be correct. Although the use of large-scale datasets - as with phylogenomics - seems promising, there are cases where different studies yielded different results. One of these cases concerns the basal taxa of the animal *Tree of Life* and the branching order of early-diverging metazoa.

1.3 The quest for the tree of basal animals

Our project deals with the phylogenetic relationships between the basal animal (non-bilaterian) taxa Porifera, Ctenophora, Cnidaria, and Placozoa. Phylogenomics provided a robust picture of bilaterian relationships (Delsuc *et al.*, 2006; Dunn *et al.*, 2008; Philippe *et al.*, 2005). At the beginning of our project, the most complete picture of the animal *Tree of Life* was published in Dunn *et al.* (2008). It used sequences from 21 animal phyla and could confirm previously highly disputed hypotheses, e.g. velvet worms as the sister group of arthropoda and monophyletic molluscs. Due to insufficient taxon sampling the relationships of the basal taxa could not be resolved, leaving their phylogenetic status as well as the phylogenetic origin of sponges unresolved.

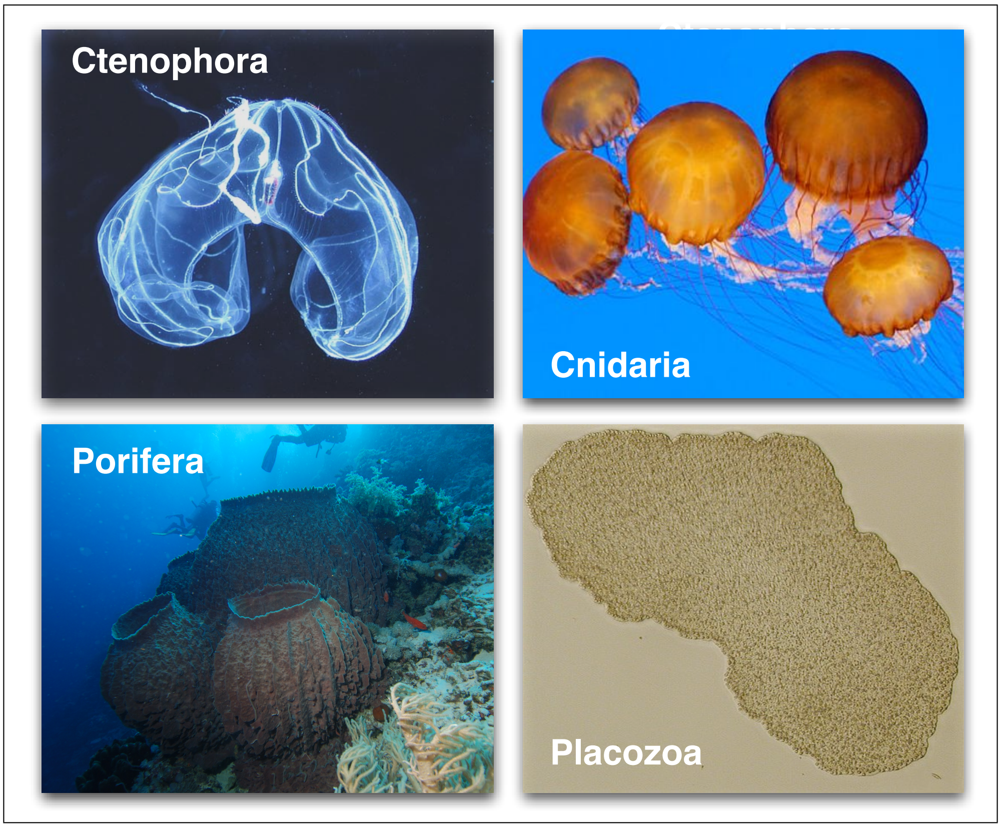


Figure 1.1: The figure shows the ctenophore *Bathocyroe fosteri*, the cnidarian *Chrysaora fuscescens*, the sponge *Xestospongia testudinaria*, and the placozoan *Trichoplax adhaerens*. Pictures taken from (Wikipedia, 2010a,b,c,d).

1.3.1 The basal branches of the metazoan tree

Recent studies that try to resolve the animal *Tree of Life* or parts of it led to contradictory and poorly resolved trees regarding the relationships between basal taxa (Rokas *et al.*, 2005; Schierwater *et al.*, 2009). In the following, we briefly summarize the working hypotheses for the evolutionary relationship of basal metazoan taxa (see Figure 1.1):

Cnidaria and Ctenophora - Coelenterata or not?

Cnidaria is an animal phylum containing over 9,000 species. It includes jellyfish, corals, sea pansies, sea pens, box jellies, and sea wasps and is found exclusively in aquatic and mostly marine environments. The Ctenophora (comb jellies) are an animal phylum that lives in marine waters worldwide (see figure 1.1).

During the long history of animal phylogenetics, mostly two different (of the three possible) scenarios regarding the branching order of Eumetazoa (Cnidaria, Ctenophora, and Bilateria) were found (see figure 1.2). In one of these trees, Ctenophora and Cnidaria form the clade Coelenterata as a sister group to Bilateria. In another hypothesis, Cnidaria are basal to a Ctenophora + Bilateria clade, called Acrosomata. The third tree has Ctenophora branching off first and contains a Cnidaria + Bilateria clade.

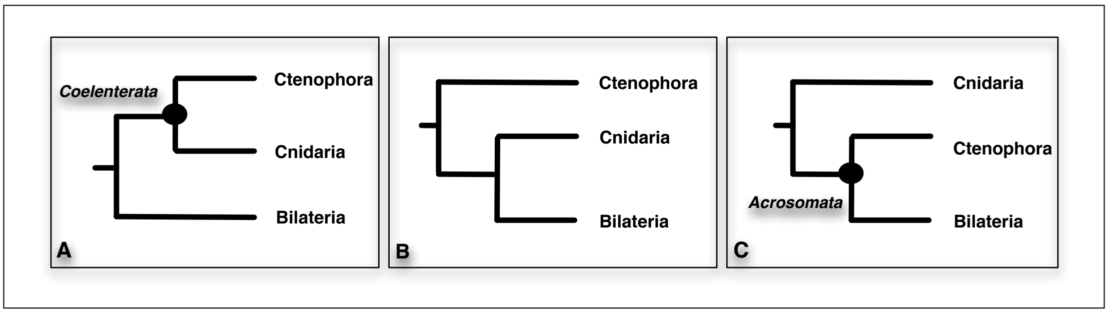


Figure 1.2: The three possible hypotheses about the branching order of Ctenophora, Cnidaria, and Bilateria. Picture redrawn from (Minelli, 2009).

The latter describes a rather uncommon scenario because it implies that slightly more complex taxa (Ctenophora) have a larger evolutionary distance to Bilateria than obviously simpler Cnidaria.
The three trees find support by the following character sets:

- The Coelenterata hypothesis (Ctenophora + Cnidaria) is based on unilateral cleavage (see figure 1.2, A).

- The presence of Hox and Parahox genes as well as collinearity of the Hox gene expression support the Cnidaria + Bilateria clade (Finnerty and Martindale, 1997; Martinez *et al.*, 1998; Martindale *et al.*, 2002) (see figure 1.2, B).

- The clade Acrosomata (Ctenophora+Bilateria) finds support by the presence of true muscle cells, multiciliate cells, complex sensory organs, a through-gut, and a highly stereotyped cleavage (Martindale and Henry, 1997) (see figure 1.2, C).

The recent study by Dunn *et al.* (2008) placed Ctenophora as the most basal animal taxon. This contradicts classical concepts as it implies that morphological more simple animals like sponges are younger than more complex animals like Ctenophora.

Are placozoans reduced cnidarians?

Placozoans are basal, multicellular animals. They are very flat creatures, about 1mm wide, lacking any organs or internal structures (see figure 1.1). The phylogenetic position of Placozoa within the basal tree of animals is still uncertain. The phylum Placozoa was traditionally regarded to be represented by the taxon *Trichoplax adhaerens* only, but is now assumed to be greater in diversity (Signorovitch *et al.*, 2005; Voigt *et al.*, 2004). An analysis using 18S rRNA suggests placozoans to be secondary reduced cnidarians (Cavalier-smith and Chao, 2003). Contradictory, not only does the organization of the mitochondrial genome of *T. adhaerens* and Cnidaria differ, but also the predicted secondary structure of the 16S rRNA is different between the two phyla (Ender and Schierwater, 2003). Syed and Schierwater (2002) proposed that Placozoa would represent a basal metazoan stem line that branched off first to the group Porifera + Eumetazoa. This view finds support when looking at the size and structure of the mitochondrial DNA (Dellaporta *et al.*, 2006).

This unclear picture of the phylogenetic position of Placozoa does not change despite the use of mitochondrial genomes (Haen *et al.*, 2007; Wang and Lavrov, 2007), and 50 nuclear genes (Rokas *et al.*, 2005). An analysis using SSU rRNA sequences of 528 metazoan taxa supported a sister group relationship of Placozoa to a Cnidaria + Bilateria clade (Wallberg *et al.*, 2004). These inconsistencies are mainly due to insufficient and/or inadequate taxon sampling of basal metazoa.

1.3.2 The phylogenetic origin of sponges

Sponges are a diverse group of animals with many body plan features in common and are classified into *Demospongiae*, *Hexactinellida*, *Calcarea*, and *Homoscleromorpha* (see figure 1.1). The presence of a system of internal canals and choanocyte chambers, through which water flows, together with the pinacoderm, a thin epithelial covering, firmly support a monophyletic origin of sponges. A monophyletic origin of sponges supports the idea that these features evolved only once (see figure 1.3).

Whole-genome analysis supports a sister-group relationship of sponges to all other metazoa (Srivastava *et al.*, 2008). This is in accordance with morphology (Ax, 1996).

Molecular data contradicts morphology

This picture gets blurred when looking at studies using molecular evidence. A paraphyletic origin of sponges has been supported by studies based on 18S rRNA (Borchiellini *et al.*, 2001; Cavalier-smith *et al.*, 1996; Collins, 1998; Peterson and Eernisse, 2001), protein kinase C (Kruse *et al.*, 1998), and seven nuclear-encoded genes (Peterson and Butterfield, 2005). In these studies, Calcarea form a sister group together with non-sponge metazoans (Epitheliozoa). This grouping finds morphological support by the presence of striated ciliary rootlets in the larvae of calcareous sponges as well as in Epitheliozoa (Rieger, 1976), but not in other sponges (Woollacott, 1995).

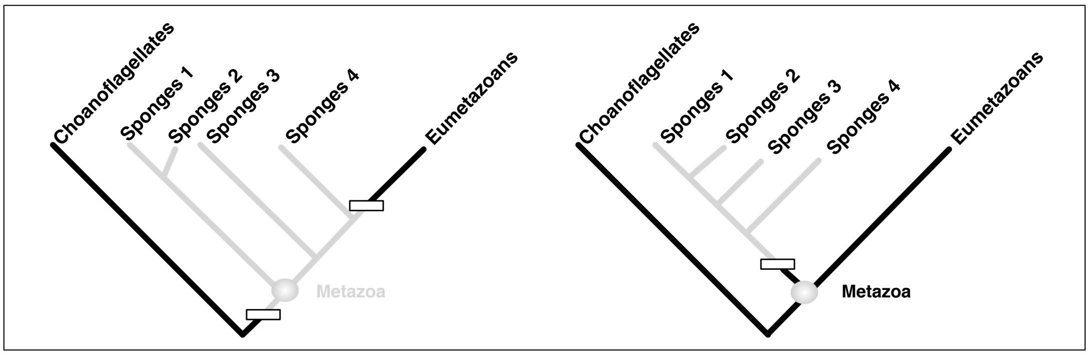


Figure 1.3: (A) Most parsimonious scenario for sponge paraphyly. (B) most parsimonious scenario for sponge monophyly. Picture redrawn from (Philippe *et al.*, 2009).

Another recent contentious issue regards the position of the Homoscleromorpha, a taxon formerly placed within the Demospongiae. The Homoscleromorpha share many morphological and developmental features with non-sponge metazoans (Boury-Esnault *et al.*, 2003).

Implications of sponge paraphyly are interesting for understanding the evolution of early-branching metazoans: characters shared by all sponge lineages are ancestral to Metazoa and eumetazoans are derived from sponge-like organisms (Borchiellini *et al.*, 2001; Peterson, 2001; Nielsen, 2008).

SUMMARY

Our leading questions are:

- ◇ Relationships between basal *Metazoa*?
- ◇ What is the position of *Placozoa* in the basal metazoan tree?
- ◇ Are *sponges* monophyletic or not?
- ◇ Relationship within *sponge* classes?

1.4 How phylogenomics can help

Recent studies focussing on resolving the early branches of the animal tree were equivocal and had either no support or did not include enough taxa to draw conclusions. With an increase in the amount of available sequences, the phylogenomic approach will be suitable to get the true tree of basal animals.

Orthologs and paralogs - the apples and oranges of phylogenetics

Phylogenetic trees are based on orthologous sequences. Following the original definition by Fitch (1970) sequences are called orthologous if they diverged through a speciation event; sequences are called paralogous if they diverged through a duplication event within the same species (see figure 1.4).

Orthology assignment is a crucial prerequisite in phylogenetic studies as falsely predicted orthologs can lead to incorrect tree hypotheses (Zmasek and Eddy, 2002). The selection of orthologous sequences in phylogenomics is even more critical as dataset size increases making manual orthology search impossible.

Orthology Search - a crucial task

A common approach to orthology search is to use similarity search tools like *BLAST* (Altschul *et al.*, 1997) to search query sequences against a sequence database. As a result of that search, the best hit or the best reciprocal hit (Mushegian *et al.*, 1998), two sequences from different datasets that find each other as the best scoring hit, is commonly regarded as an orthologue to the query sequence. However, this is not a sufficient condition to determine orthologous relationship between sequences (Johnson, 2007). The method fails in the case of e.g. gene loss. Several methods for prediction orthologs have been developed and extensively compared (Alexeyenko *et al.*, 2006; Altenhoff and Dessimoz, 2009; Chen *et al.*, 2007; Dutilh *et al.*, 2007). These methods are based on either a phylogenetic analysis (e.g. *Rio* (Zmasek and Eddy, 2002), *PhyOP* (Goodstadt and Ponting, 2006), *Ensembl Compara* (Hubbard *et al.*, 2007)) or all-against-all *BLAST* searches (Dolinski and

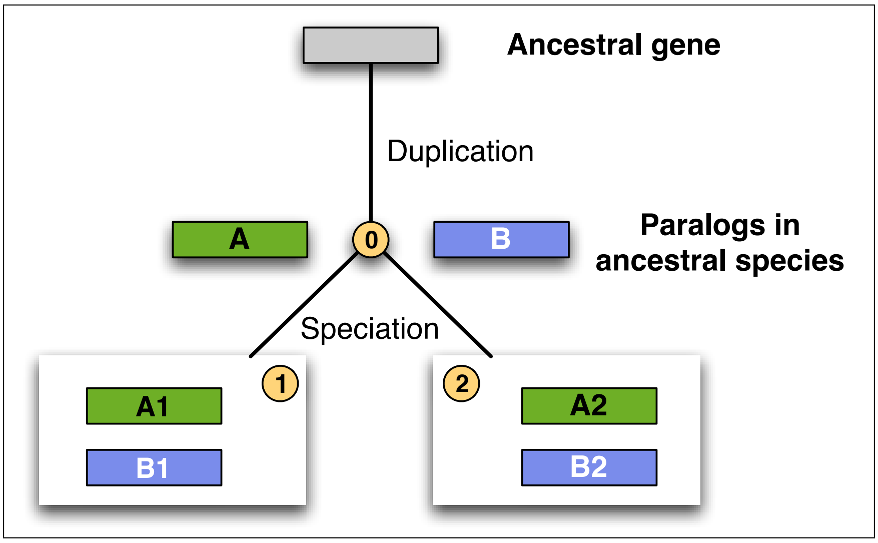


Figure 1.4: The picture describes the relationship between orthologous and paralogous genes. An ancestral gene is duplicated within the genome of species 0 leading to the two copies A and B. After the speciation event, there are two copies in each of the species 1 and 2. The genes A1 and A2 as well as the genes B1 and B2 have an orthologous relationship, because they stem from a speciation event. The genes A1 and B1 and A1 and B2 have a paralogous relationship, since they are the result of a duplication event in the common ancestor 0. Picture redrawn from (Koonin, 2001).

Botstein, 2007). Among the all-against-all methods, several use the reciprocal condition (Waterston *et al.*, 2002; Remm *et al.*, 2001; Tatusov *et al.*, 1997), while others start with reciprocal best-hitting sequence pairs and further cluster those pairs using evolutionary distances (DeLuca *et al.*, 2006), Markov clustering (Li *et al.*, 2003), third-party species (Mclysaght and Huson, 2005; Schneider *et al.*, 2007) or include additional information, e.g. guide trees and gene neighborhood conservation (Sayers *et al.*, 2010).

Problems of existing methods: All existing methods are designed to cluster protein sequences. However, they are not designed to explicitly deal with EST sequences and their correct translation. Furthermore, most existing methods are not capable of dealing with the high redundancy of gene copies. That is, they can not select the sequence most likely to be orthologous from a set of gene copies. Besides that, existing tools that rely on phylogenetic trees require manual curation and are therefore inappropriate for large-scale analysis. Summarized, existing tools are not suitable in EST-based phylogenomics analyses.

EST handling

ESTs are commonly used in large-scale studies because they provide a wealth of phylogenetic information and are relatively cheap to generate. However, ESTs often contain sequencing errors and may cover genes partially, only (James and Mark, 2004). These errors can lead to shifts in the reading frame and make translation non-trivial. Several tools (Iseli *et al.*, 1999; James and Mark, 2004; Shafer *et al.*, 2006; Xu *et al.*, 2007) and web servers (Lee *et al.*, 2007; Schmid and Blaxter, 2008; Smith *et al.*, 2008; Strahm *et al.*, 2006) have been developed to correct sequencing errors and try to avoid frame shift errors.

Phylogenomic workflow - how large datasets are analysed

Phylogenomic studies are based on large sets of sequences. In general, there are two different types of analyses in phylogenomics (see figure 1.5):

- ◇ sequence-based methods,
- ◇ whole-genome methods.

In this study, we focus on sequence-based methods only, because whole-genome data is limited for basal metazoan taxa.

The dataset size of recent phylogenomic studies dramatically increased in the last years. These datasets (e.g. Bapteste *et al.* (2002); Blair *et al.* (2002); Lerat *et al.* (2003); Rokas *et al.* (2003b); Wolf *et al.* (2004)) include many characters, but a considerably lower number of taxa. An interesting issue in phylogenomics is whether the number of taxa or the number of characters should be increased to improve the accuracy of the resulting tree (Graybeal, 1998; Hillis *et al.*, 2003; Lecointre *et al.*, 1993; Poe and Swofford, 1999; Rosenberg and Kumar, 2003). While computer simulations are equivocal (Hillis *et al.*, 2003; Rosenberg and Kumar, 2003), empirical studies support an increased sampling of species (Lecointre *et al.*, 1993; Lin *et al.*, 2002; Philippe, 1997). Datasets with complete genome sequences available would be asymmetrical having either many species and few genes or vice versa. Phylogenomic studies aim at maximizing both, the number of species and the number of genes (Driskell *et al.*, 2004; Sanderson *et al.*, 2003), in order to be able to construct more accurate trees (Lin *et al.*, 2002; Philippe, 1997).

The presence or absence of genes and/or species in such datasets leads to another issue, the impact of missing data on the resulting phylogeny.

In general, there are two different methods - the supermatrix and the supertree approach - to combine the information from single gene alignments that were assembled from local and/or public databases[2]. In this study, we focus on the

[2] Phylogenomics also offers methods that use whole-genome features such as gene content or gene order to build phylogenetic trees, but this is not covered here.

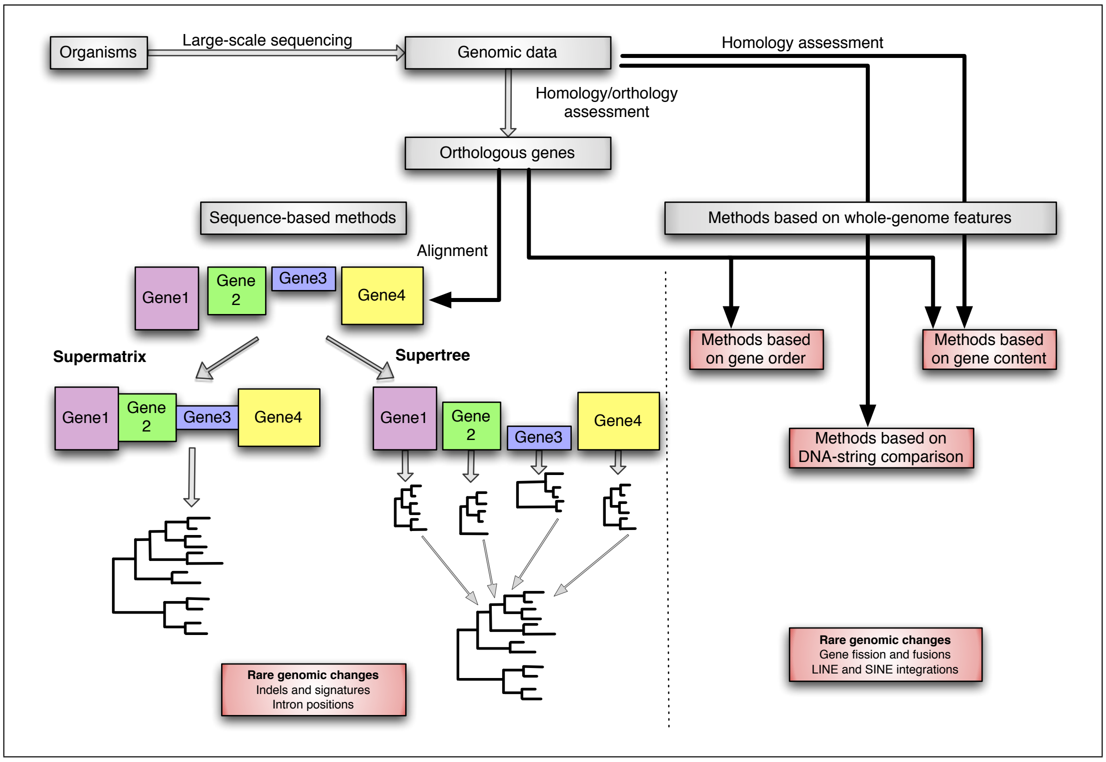


Figure 1.5: The picture shows both currently applied methods of tree inference from genomic data. Obtained from large-scale sequencing projects, sequences are assembled into orthologous genes. Subsequent analysis is based on either sequence-based methods, that construct phylogenetic trees using the supermatrix or supertree approach, or based on whole-genome features. Picture redrawn from (Delsuc *et al.*, 2005).

supermatrix approach only, because it has been shown to be more accurate in simulation studies than the supertree approach (Gadagkar *et al.*, 2005).

Supermatrix - Concatenating single genes

The supermatrix approach is based on the principle of total evidence and tries to use all available data (see figure 1.5). For this, all genes under study are concatenated and missing data - the absence of genes in some species - is marked as a question mark. Recent studies used different levels of missing data (12,5% in Murphy *et al.* (2001), 20% in Qiu *et al.* (1999), 25% in Bapteste *et al.* (2002)) to investigate the impact of missing data. Empirical studies (Driskell *et al.*, 2004; Gatesy *et al.*, 2002; Philippe *et al.*, 2004) as well as simulations (Philippe *et al.*, 2004; Wiens, 2003) found that even species with a large proportion of missing data can be correctly placed in a tree, given the available data are informative enough (Philippe *et al.*, 2004). These findings show that the supermatrix approach is relatively robust against missing data. That makes it applicable to datasets with EST sequences, that are cheap to generate but are an incomplete resource of sequence information.

1.5 Aims - phylogenomics

Despite the use of large datasets, the current situation in phylogenetics of basal metazoan is highly debated.

The aim of this work is to contribute to finding an answer regarding the emergence order of the basal metazoan taxa Ctenophora, Cnidaria, Placozoa, and Porifera as well as whether the phylum Porifera is monophyletic or paraphyletic.

Our hypothesis is that the massive use of newly generated sequences and data from previously neglected taxa combined with new methods for automated dataset construction and the application of complex models of sequence evolution will lead to more accurate trees and answer our questions (see section 1.3.2). The results should give further hints on the early evolution of the animal body plan as well as the phylogenetic origin of sponges.
Our project is divided into two parts (see figure 1.6):

First Part We design, implement and test a new method for automatically constructing datasets in EST-based phylogenomics.

Second Part We apply phylogenomic methods on newly generated EST data and data from public databases to test existing hypotheses.

1.5.1 First part: OrthoSelect

Although cheap and rapid to generate, the use of ESTs in evolutionary studies is hindered by the lack of available tools for automated orthology search in ESTs. Existing methods either require a known species tree or cannot cope with the redundancy in ESTs. A drawback of all existing methods is that they cannot handle sequence redundancies - multiple copies of the same gene. By the development of a new tool - called *OrthoSelect* - for orthology prediction in ESTs, we aim for filling this gap. *OrthoSelect* is able to search large databases for the presence of orthologous genes in ESTs and produce gene alignments ready to use for downstream analysis (e.g. construction of phylogenetic trees). The development of

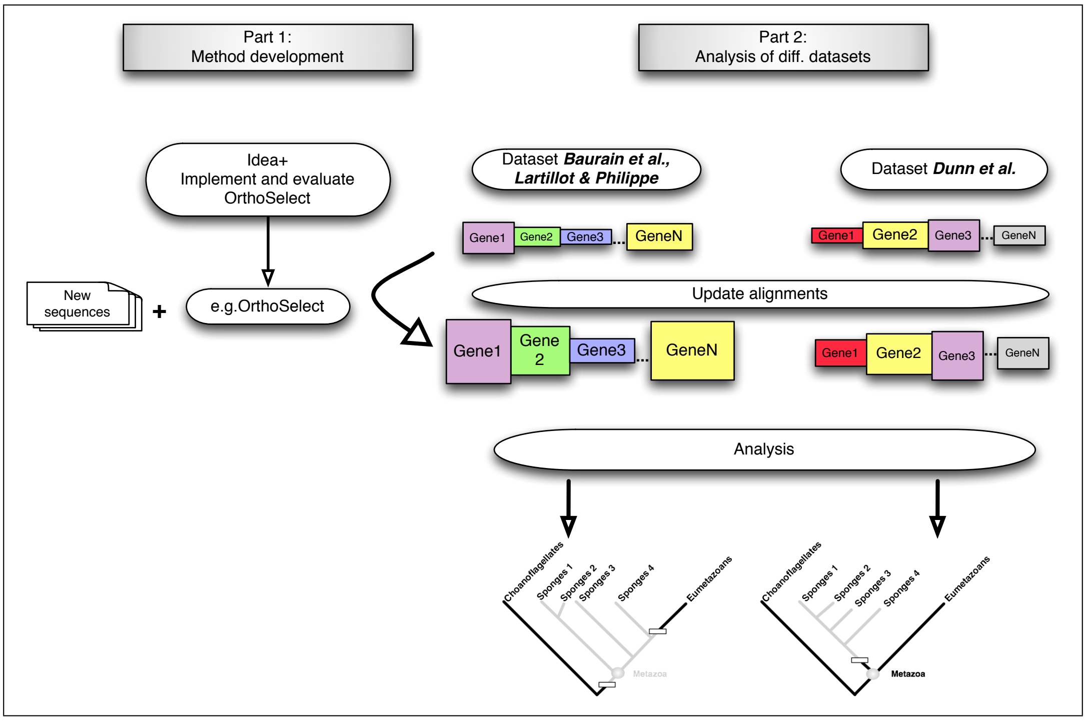


Figure 1.6: This picture shows the two main parts of our project. The first part - on the left side - is the development of an automated tool for constructing datasets for phylogenomics. The second part - on the right side - describes the extension of existing datasets by tools like *OrthoSelect* and its subsequent analyses. The analyses will help to either corroborate or disprove existing hypotheses about the evolution of basal metazoan taxa.

OrthoSelect allows - for the first time - the complete and automated construction of phylogenomic datasets. We evaluate *OrthoSelect* by comparing it to the already published and manually curated phylogenomic dataset of Dunn *et al.* (2008). The tool is written in *Perl* and freely distributed as a command line program.

OrthoSelect - webserver

Additionally, we set up a web server to provide easy access to the command line program. Compared to the command line version, the web server additionally visualizes gene alignments and provide numerous additional statistics about e.g. the generated gene clusters, and the presence of taxa.

1.5.2 Second part: Dataset generation

In order to test the hypotheses mentioned in section 1.3.2, we have to assemble suitable datasets. We use two published datasets as a basis and add our newly generated sponge sequences as well as sequences from other basal metazoan taxa. The extension of both datasets is necessary to be able to test the competing hypotheses from section 1.3.2. The datasets are:

- ◇ The datasets from Baurain *et al.* (2007) and Lartillot and Philippe (2008),
- ◇ The dataset from Dunn *et al.* (2008), the most complete and comprehensive view of the animal phylogeny at that time.

Phylogenetic analysis

The two datasets will be analysed using the supermatrix approach. It is widely accepted that the evolutionary process is stochastic and should be modeled in a statistical way (Maddison and Knowles, 2006). Therefore we use likelihood-based methods of phylogenetic inference rather than distance-based methods or maximum parsimony. Additionally, likelihood-based models incorporate more complex models of sequence evolution (Whelan *et al.*, 2001) and proved to be less affected by model violations (Kolaczkowski and Thornton, 2004). One of these models, the categories model *CAT* (Lartillot and Philippe, 2004), relaxes the assumption of homogeneity of alignment columns. By assigning different evolutionary categories to single sites the *CAT* model can account for among-site heterogeneities (Pagel and Meade, 2004).

We are aware of the following possible limitations: The resolution of a phylogenetic tree depends on how dominant the phylogenetic signal in the dataset is. Time intervals between the deepest nodes in the animal tree, that define the branching pattern of non-bilaterians, are short and it is questionable whether these nodes can be significantly supported even by the use of genome-scale datasets (Philippe *et al.*, 1994). In order to validate current evolutionary hypotheses, the results from phylogenetic analyses using different dataset and/or models should be congruent (Miyamoto and Fitch, 1995).

1.6 Metagenomics - investigating the unculturable

After their discovery by Anton van Leeuwenhoek in the 1670, microorganisms have been intensively studies and much has been learned about their importance to human health, agriculture, industry, and the origin and evolution of life. However, most microorganisms are still unknown (Eisen, 2007). Many bacterial phyla are not culturable, so other methods are needed to access the physiology and genetics of these organisms (Handelsman, 2004). One of these methods is metagenomics, where a population of microorganisms is subject to a genomic analysis. In metagenomics, genomes from microbial communities are randomly sampled resulting in usually large databases of environmental sequence tags. The direct sequencing of genomic DNA from the species that live in these communities allows to study their evolution, lifestyle, and diversity (Béjà *et al.*, 2000; Gill *et al.*, 2006; Hansen *et al.*, 2007).

The development of Pyrosequencing (Margulies *et al.*, 2005) produces a lot of data and allows the direct sequencing of metagenomes without cloning (Edwards *et al.*, 2006). Currently, pyrosequencing generates only short sequence reads (≈ 100 - 450 bp) which makes an assembly into contiguous sequences (contigs) a computationally very demanding task.

An important question in metagenomics is to quantify and characterize microbial communities. To do this, their taxonomic composition can be assessed via the generation of taxonomic profiles.

1.6.1 16S rRNA

After the pioneering work of *Carl Woese* (Woese and Fox, 1977; Woese, 1987), 16S rRNA and 18 rRNA have been established as reliable phylogenetic markers. Besides their high accuracy, methods using these markers can taxonomically profile a small proportion of metagenomes, only.

1.6.2 Alternative approaches

To overcome this limitation, the set of marker gene was extended in Wu and Eisen (2008) and Mering *et al.* (2007) using methods (Huson *et al.*, 2007; Meyer *et al.*, 2008) that rely on sequence similarity search against public databases using *BLAST* (Altschul *et al.*, 1997). The known shortcomings of *BLAST* (requirement of sufficient read length and presence of close homologs in the database) led to the development of tools that directly pursue the classification of the DNA signatures (Brady and Salzberg, 2009; Diaz *et al.*, 2009; McHardy *et al.*, 2007). However, previous methods showed a drop in accuracy as sequence length gets shorter than 1,000 bp and are computationally demanding (Krause *et al.*, 2008). This makes them not suitable to handle the rapidly increasing dataset size in metagenomics and metatranscriptomics. New tools are needed to overcome the computational burden as well as to provide high accuracy.

1.7 Aims - metagenomics

We develop a new tool, called *Treephyler*, for assessing community profiles of metagenomes and metatranscriptomes. We will overcome existing limitations in computational complexity using the new method *UFO* (Meinicke, 2009) that makes fast assignments of sequences to *PFAM* (Finn *et al.*, 2008) families.

Chapter 2

List of Publications

The thesis is based on the following original papers (in chronological order):

◇ **Chapter 3**
Hervé Philippe, Romain Derelle, Philippe Lopez, Kerstin S. Pick, Carole Borchiellini, Nicole Boury-Esnault, Jean Vacelet, Emmanuelle Deniel, Evelyn Houliston, Eric Quéinnec, Corinne Da Silva, Patrick Wincker, Hervé Le Guyader, Sally Leys, Daniel J. Jackson, Bernard M. Degnan, **Fabian Schreiber**, Dirk Erpenbeck, Burkhard Morgenstern, Gert Wörheide and Michael Manuel.
Phylogenomics restores traditional views on deep animal relationships.
Current Biology (2009) 19, 706-712.

◇ **Chapter 4**
Fabian Schreiber, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences.
Nucleic Acids Research (2009) 37, W185-W188.

◇ **Chapter 5**
Fabian Schreiber, Kerstin Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A protocol for selecting orthologous groups in phylogenomics.
BMC Bioinformatics (2009) 10, 219.

◇ **Chapter 6**
Fabian Schreiber, Peter Gumrich, Rolf Daniel and Peter Meinicke.
Treephyler: Fast taxonomic profiling of metagenomes.
BBioinformatics (2010), 26(7):960-961.

- **Chapter 7**
  Kerstin S. Pick[1], Hervé Philippe[1], **Fabian Schreiber**, Dirk Erpenbeck, Daniel J. Jackson, Petra Wrede, Mathias Wiens, Alexandre Alié, Burkhard Morgenstern, Michael Manuel and Gert Wörheide.
  Broader phylogenomic sampling improves the accuracy of non-bilaterian relationships.
  Molecular biology and evolution (2010), doi:10.1093/molbev/msq089.

[1] These authors contributed equally

Chapter 3

Phylogenomics restores traditional views on deep animal relationships

Citation

Hervé Philippe, Romain Derelle, Philippe Lopez, Kerstin S. Pick, Carole Borchiellini, Nicole Boury-Esnault, Jean Vacelet, Emmanuelle Deniel, Evelyn Houliston, Eric Quéinnec, Corinne Da Silva, Patrick Wincker, Hervé Le Guyader, Sally Leys, Daniel J. Jackson, Bernard M. Degnan, Fabian Schreiber, Dirk Erpenbeck, Burkhard Morgenstern, Gert Wörheide and Michael Manuel.
Phylogenomics restores traditional views on deep animal relationships.
Current Biology (2009) 19, 706-712.

Original Contribution

FS helped to assemble the dataset by using a preliminary version of *OrthoSelect* to add newly generated sponge sequences.

Current Biology *19*, 706–712, April 28, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.02.052

Report

Phylogenomics Revives Traditional Views on Deep Animal Relationships

Hervé Philippe,[1,11] Romain Derelle,[2,11] Philippe Lopez,[2] Kerstin Pick,[3,5] Carole Borchiellini,[6] Nicole Boury-Esnault,[6] Jean Vacelet,[6] Emmanuelle Renard,[6] Evelyn Houliston,[7] Eric Quéinnec,[2] Corinne Da Silva,[8] Patrick Wincker,[8] Hervé Le Guyader,[2] Sally Leys,[9] Daniel J. Jackson,[3,10] Fabian Schreiber,[4] Dirk Erpenbeck,[5] Burkhard Morgenstern,[3,4] Gert Wörheide,[5,*] and Michaël Manuel[2,*]

[1]Centre Robert-Cedergren
Département de Biochimie
Université de Montréal
Succursale Centre-Ville
Montréal, Québec H3C3J7
Canada
[2]UPMC, Univ Paris 06
UMR 7138 Systématique, Adaptation, Evolution
CNRS UPMC MNHN IRD, Case 05
Université Pierre et Marie Curie
7 quai St Bernard
75005 Paris
France
[3]Courant Research Center Geobiology
Georg-August-Universität Göttingen Goldschmidtstr. 3
[4]Abteilung Bioinformatik
Institut für Mikrobiologie und Genetik
Goldschmidtstr. 1
37077 Göttingen
Germany
[5]Department of Earth- and Environmental Sciences & GeoBioCenter^{LMU}
Ludwig-Maximilians-Universität München
Richard-Wagner-Str. 10
80333 München
Germany
[6]Aix-Marseille Université
CNRS UMR 6540 DIMAR
Centre d'Océanologie de Marseille
Station Marine d'Endoume
rue de la Batterie des Lions
13 007 Marseille
France
[7]UPMC, Univ Paris 06
CNRS UMR 7009 Biologie du Développement
Observatoire Océanologique
06230 Villefranche-sur-Mer
France
[8]Genoscope and CNRS UMR 8030
2 rue Gaston Crémieux
91057 Evry
France
[9]Department of Biological Sciences
CW 405
University of Alberta
Edmonton, AB T6G 2E9
Canada
[10]School of Integrative Biology
The University of Queensland
Brisbane 4072
Australia

Summary

The origin of many of the defining features of animal body plans, such as symmetry, nervous system, and the mesoderm, remains shrouded in mystery because of major uncertainty regarding the emergence order of the early branching taxa: the sponge groups, ctenophores, placozoans, cnidarians, and bilaterians. The "phylogenomic" approach [1] has recently provided a robust picture for intrabilaterian relationships [2, 3] but not yet for more early branching metazoan clades. We have assembled a comprehensive 128 gene data set including newly generated sequence data from ctenophores, cnidarians, and all four main sponge groups. The resulting phylogeny yields two significant conclusions reviving old views that have been challenged in the molecular era: (1) that the sponges (Porifera) are monophyletic and not paraphyletic as repeatedly proposed [4–9], thus undermining the idea that ancestral metazoans had a sponge-like body plan; (2) that the most likely position for the ctenophores is together with the cnidarians in a "coelenterate" clade. The Porifera and the Placozoa branch basally with respect to a moderately supported "eumetazoan" clade containing the three taxa with nervous system and muscle cells (Cnidaria, Ctenophora, and Bilateria). This new phylogeny provides a stimulating framework for exploring the important changes that shaped the body plans of the early diverging phyla.

Results and Discussion

A Comprehensive Phylogenomic Data Set to Address Basal Metazoan Evolution

Previous studies of basal metazoan relationships by molecular phylogeny techniques (e.g., [3–8, 10, 11]) have proposed contradictory and often poorly supported trees, leaving major issues such as the phylogenetic status (monophyly or paraphyly) of sponges and the position of ctenophores and placozoans unsettled. These inconsistencies may reflect insufficient molecular sampling and/or inadequate taxon sampling of the diversity of extant nonbilaterian metazoan lineages [1, 11–13]. We have adopted a phylogenomic approach specifically aimed at clarifying the basal metazoan relationships, involving more comprehensive sampling of all the major early branching animal lineages. By using newly generated cDNA sequences in addition to publicly available sequences, we have assembled a metazoan data set enriched in species representing the early diverging phyla (see Experimental Procedures and Supplemental Data available online). The data set comprises 128 different protein-coding genes (30,257 unambiguously aligned

*Correspondence: woerheide@lmu.de (G.W.), michael.manuel@snv.jussieu.fr (M.M.)
[11]These authors contributed equally to this work

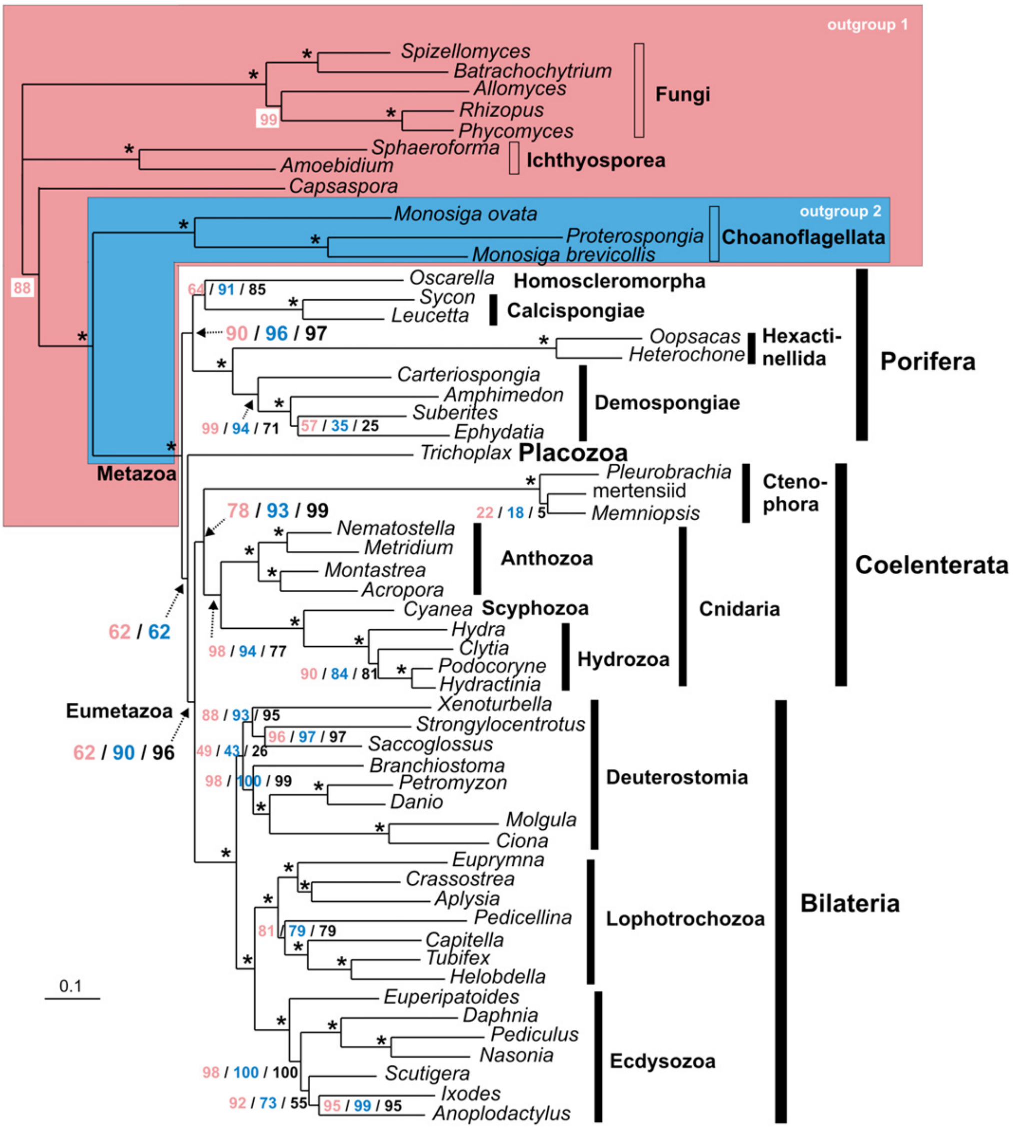


Figure 1. Phylogenetic Analyses of 128 Nuclear-Encoded Proteins

Bayesian tree obtained from the analysis of 30,257 aligned amino acid positions for the 55 terminal taxa with the CAT model. Bootstrap supports (BS) after 100 replicates are indicated for three analyses with different taxon sampling: outgroup 1 (BS values in pink); outgroup 2 (BS values in blue); unrooted analysis (BS values in black). Nodes with maximal support values in all analyses are indicated by an asterisk. The tree obtained with outgroup 1 is shown here (and in Figure S1 with branch posterior probabilities, PP), whereas trees obtained with outgroup 2 and without outgroup are shown in Figures S2 and S3, respectively. Scale bar indicates number of changes per site.

positions) for 11 outgroup species and 44 metazoans, including 9 sponge species, 3 ctenophores, 9 cnidarians, the placozoan *Trichoplax*, and a representative sampling of bilaterian species. Among the 55 terminal taxa, 24 are complete or nearly complete (≤5% of missing data), and only 27% of positions in the final alignment are absent (see Table S2). This is the first phylogenomic data set to include all four main sponge lineages: Demospongiae, by far the most species-rich sponge group, is represented by four species, chosen to maximize morphological and phylogenetic diversity; Hexactinellida and Calcispongia are each represented by two species; and Homoscleromorpha is represented by a chimerical operational taxonomic unit created from two species of the genus *Oscarella*.

The Sponges Restored as a Monophyletic Group

Our data set was analyzed by Bayesian inference analysis, via the CAT model of sequence evolution [14], conceived to reduce artifacts resulting from mutational saturation and unequal rates of substitution, which are major problems when analyzing ancient events [13, 15]. To explore the effect of outgroup taxa on the metazoan interrelationships obtained, we performed three analyses with different taxon samplings (Figure 1): rooted analysis with a paraphyletic outgroup comprised of the fungi, ichthyosporeans, *Capsaspora*, and choanoflagellates ("outgroup 1;" tree shown in Figure 1 and Figure S1; with bootstrap supports [BS] in pink in Figure 1); analysis rooted with just choanoflagellates, the metazoan sister group [16] ("outgroup 2;" BS in blue in Figure 1, tree in

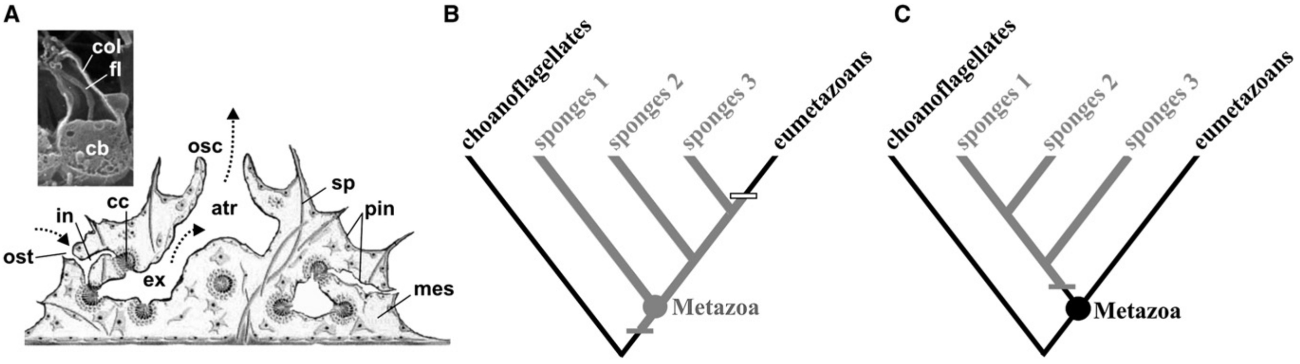


Figure 2. Characters of the Sponge Body Plan and Their Evolution

(A) Schematic section of an adult sponge (bottom) and SEM picture showing a choanocyte, the sponge collar cell (top, choanocyte from *Chelonaplysilla noevus*, Demospongiae). The arrows indicate the direction of circulation of water in the aquiferous system of the sponge. Abbreviations: atr, atrial cavity; cb, cell body; cc, choanocyte chamber; col, collar of microvilli; ex, exhalant canal; fl, flagellum; in, inhalant canal; mes, mesohyl; osc, osculum (or exhalant orifice); ost, ostium (or inhalant orifice); pin, pinacoderm (thin epithelial layer, limiting the sponge body on its external surface and within the canals); sp, spicule.

(B) Most parsimonious scenario for the evolution of sponge body plan characters, imposed on a scheme of sponge paraphyly.

(C) Most parsimonious scenario assuming sponge monophyly.

In (B) and (C), the gray branches indicate the presence of sponge body plan characters (aquiferous system, internalized choanocyte chambers, pinacoderm) and the black branches indicate the absence of these characters. The gray horizontal line indicates character acquisition; the hollow horizontal line indicates character loss. "Sponges 1, 2, and 3" correspond to the major lineages (silicisponges, homoscleromorphs, and calcisponges), of which exact branching order varies among published studies recovering sponge paraphyly.

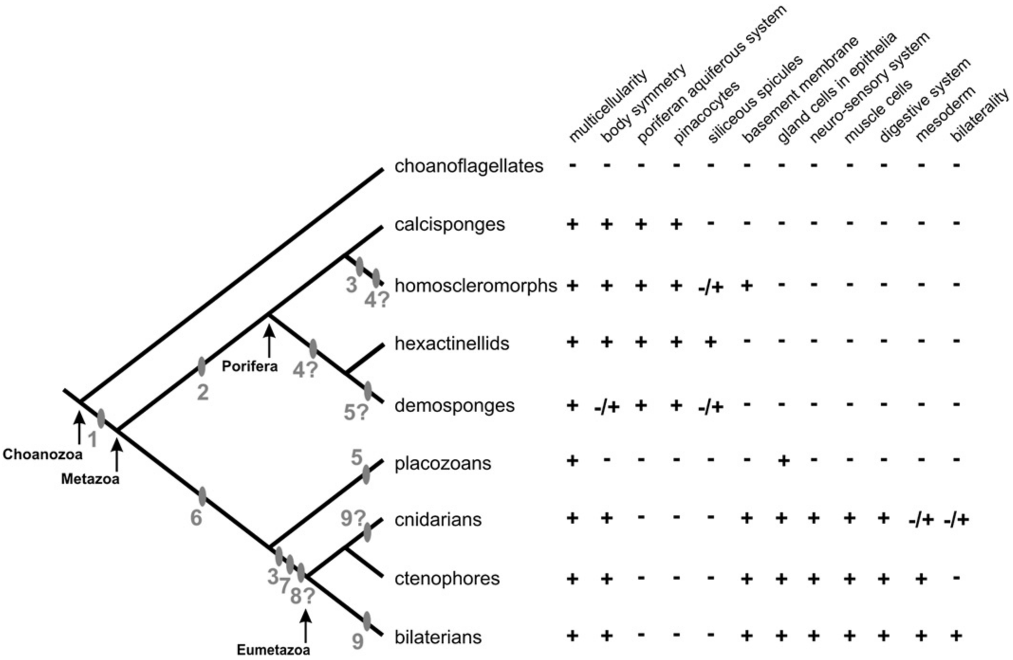
Figure S2); and unrooted analysis (BS in black in Figure 1, tree in Figure S3). The topology resulting from the rooted analyses (trees shown in Figure 1 and Figures S1 and S2) was statistically well supported at most nodes, and its general features were in line with previous studies [2, 3]: choanoflagellates positioned as the sister group to the Metazoa, with Bilateria, Protostomia, Lophotrochozoa, and Ecdysozoa each forming well-supported monophyletic groups. These rooted trees provide strong evidence that the sponge species all belong together in a monophyletic group (Porifera) (bootstrap support = 90% and 96% with outgroup 1 and outgroup 2, respectively). The branch leading to the Porifera is short (Figure 1), accounting for the difficulty in recovering sponge monophyly in previous molecular analyses. This presumably reflects closely spaced splitting events during the Proterozoic era when the sponge lineages emerged.

Extant sponges are a diverse group sharing a number of common body plan features, notably a system of internal canals and choanocyte chambers through which water flows, and a thin epithelial covering called the pinacoderm (Figure 2A). Although morphological character analyses firmly support the hypothesis that the sponges form a monophyletic group [5, 17], rRNA analyses have repeatedly indicated that they are paraphyletic, with the calcisponges and/or the homoscleromorphs positioned closer to eumetazoans than to the other sponges [4–8]. It is worth noting, however, that sponge monophyly could not be ruled out unequivocally in many of these studies because of poor statistical support [6, 7, 10]. The previously proposed hypothesis of sponge paraphyly had significant implications for understanding the origin of multicellular animals, because it would imply that characters shared by all sponge lineages are ancestral for the Metazoa and that eumetazoans are derived from animals with a sponge-like body plan [4, 5, 8, 9] (Figure 2B).

The significant support for sponge monophyly in the present study allows us to return to the idea that a sponge body plan (notably featuring an aquiferous system with internalized choanocyte chambers and the pinacoderm) evolved in the stem line of the Porifera (Figures 2C and 3). The specialized collar apparatus of sponge choanocytes has often been assumed to be an ancient feature shared with choanoflagellates, based on phenotypic similarity [16]. However, many ultrastructural details of choanoflagellate and choanocyte cells are different, such as the length and spacing of the microvilli and the organization of the microtubule cytoskeleton. Their functional properties also differ, with the microvilli of choanoflagellates but not of choanocytes being contractile. Their similarity might thus represent convergence, with choanocytes being a synapomorphy (shared derived character) of Porifera. It is clear in any case that, rather than reflecting the ancestral animal form, adult sponges are better considered as highly specialized organisms, possibly having acquired a sedentary life style from a hypothetical pelagic ancestor. Notably, the absence of obvious symmetry in many adult sponges fuelled the popular idea that the last common metazoan ancestor lacked defined axial organization [18, 19]. In fact the adult bodies of hexactinellids, calcisponges, homoscleromorphs, and nonbilaterian eumetazoans are characterized by axial symmetry, as is the larval organization of sponges [20], ctenophores, and cnidarians. This suggests that the common ancestor of all animals may have showed symmetry around a single polarity axis [21], and thus that the asymmetry of the adult body in most demosponges and in *Trichoplax* is likely to be derived rather than ancestral (Figure 3).

Lessons from Relationships within the Porifera

In line with some previously published phylogenies (e.g., [6, 7, 11]), our analysis placed hexactinellids and demosponges together to form the Silicea Gray, 1867 [22] sensu stricto (with maximal bootstrap support in all analyses) characterized by siliceous spicules organized around a well-defined proteic axial filament [23] and by a particular class of membrane phospholipids known as demospongic acids [24]. Concerning the enigmatic Homoscleromorpha, our analyses clearly excluded



	multicellularity	body symmetry	poriferan aquiferous system	pinacocytes	siliceous spicules	basement membrane	gland cells in epithelia	neuro-sensory system	muscle cells	digestive system	mesoderm	bilaterality
choanoflagellates	-	-	-	-	-	-	-	-	-	-	-	-
calcisponges	+	+	+	+	-	-	-	-	-	-	-	-
homoscleromorphs	+	+	+	+	-/+	+	-	-	-	-	-	-
hexactinellids	+	+	+	+	+	-	-	-	-	-	-	-
demosponges	+	-/+	+	+	-/+	-	-	-	-	-	-	-
placozoans	+	-	-	-	-	-	+	-	-	-	-	-
cnidarians	+	+	-	-	-	+	+	+	+	+	-/+	-/+
ctenophores	+	+	-	-	-	+	+	+	+	+	+	-
bilaterians	+	+	-	-	-	+	+	+	+	+	+	+

Figure 3. Changes Affecting Important Body Plan Characters Traced onto the Topology Obtained from Our Molecular Analyses

Key to character changes: 1, acquisition of multicellularity and of a symmetrical body with a single axis of symmetry and polarity; 2, acquisition of the poriferan aquiferous system and of the pinacocytes; 3, acquisition of a well-developed basement membrane supporting epithelia (by convergence in the homoscleromorph sponges and in a cnidarian-ctenophore-bilaterian ancestor); 4, acquisition of siliceous spicules (by convergence in some homoscleromorph sponges and in a hexactinellid + demosponge ancestor, or independently in the hexactinellids and within the demosponges); 5, loss of body symmetry (by convergence in the stem-line of demosponges or within them, and in placozoans); 6, acquisition of gland cells in epithelia [17]; 7, acquisition of the neuro-sensory system, of the muscle cells, and of the digestive system; 8, acquisition of the mesoderm. Homology between the mesoderm of bilaterians, ctenophores, and some cnidarians is debatable; an alternative possibility being convergence of mesoderm-like germ layers between these three taxa; 9, acquisition of bilateral symmetry (by convergence in the Bilateria and in the cnidarian stem-line or within them in the Anthozoa). Parsimony optimization by Mesquite.

them from the demosponges and favored a sister group relationship to the Calcispongiae (with highest support of 91% [BS] obtained in the analysis with outgroup 2), in line with results from 18S rRNA analyses [25, 26] but in conflict with traditional classification schemes (see [27]). The siliceous spicules without defined axial filament found in some Homoscleromorpha [23] thus might have evolved independently from those of hexactinellids and demosponges (Figure 3). In addition, homology of siliceous spicules between the latter two taxa is uncertain because they are absent in the Dictyoceratida, represented here by *Carteriospongia foliascens*, the earliest-branching Demospongiae taxon in our phylogeny (Figure 1) (see [25, 28]). Whether the thick basi-epithelial basement membrane of homoscleromorph larvae and adults, which shares homologous biochemical components with eumetazoan basement membranes [29, 30], was inherited from a common metazoan ancestor and subsequently reduced or lost in most sponges and in *Trichoplax*, or acquired independently in homoscleromorphs and eumetazoans, cannot be decided from our analyses (Figure 3).

The Coelenterata Clade Revived

A recent phylogenomic analysis suggested that the ctenophores, a phylum of marine, mostly planktonic and gelatinous animals, diverged earlier than sponges [3]. This highly unorthodox hypothesis would see the dismantling of the clade Eumetazoa (ctenophores, cnidarians, and bilaterians), despite their sharing of many key characteristics such as nerve and muscle cells and a differentiated digestive system (absent in sponges and in *Trichoplax*). Polyphyly of eumetazoans would thus imply several independent acquisitions of these features, or their secondary loss in sponges and/or placozoans [31]. Our rooted analyses are not consistent with the basal position of ctenophores, but rather suggest the existence of a Coelenterata [32] (Ctenophora + Cnidaria) clade, placed within a monophyletic Eumetazoa (Figure 1). A recent study [11] also obtained the coelenterate grouping, but with low bootstrap support, and within a heterodox scheme of eumetazoan polyphyly. Historically, the coelenterate grouping [32] was based on certain anatomical resemblances between ctenophores and the cnidarian medusae (e.g., gelatinous body, tentacles, and "radial" symmetry) that were later considered convergences [33]. In fact, the complex body plan of ctenophores (with eight longitudinal rows of ciliated "comb rows," a ramified endodermal gastro-vascular system, a complex sensory apparatus located at the aboral pole, and a prevalence of biradial symmetry [19]) differs markedly from that of the cnidarians. Apart from some common embryological features (central yolk and similar unipolar cleavages; animal pole corresponding to adult mouth), there are no clear-cut morpho-anatomical synapomorphies supporting the Coelenterata.

The very long branch leading to the ctenophores (see Figure 1) makes their position prone to perturbation by the long-branch attraction (LBA) artifact [34]. The basal position of ctenophores suggested by Dunn et al. [3] might thus have resulted from attraction of the ctenophores by the distant outgroup taxa used to root the tree. This problem was alleviated in the present study by more comprehensive species sampling and by the use of the CAT model. That ctenophores are indeed attracted by distant outgroups is empirically demonstrated in our analyses by the observed increase in branch support for Coelenterata and Eumetazoa after partial or total removal of outgroup taxa (Figure 1). Thus, when distant outgroups (notably fungi) were used (as in [3]) (outgroup 1), the Coelenterata were moderately supported (BS = 78%) and the Eumetazoa were poorly supported (BS = 62%). With choanoflagellates as the only outgroup (outgroup 2), support for Coelenterata and Eumetazoa increased remarkably (BS = 93% and BS = 90%, respectively). Even higher support for the coelenterates was obtained by unrooted analysis (BS = 99%). We further checked that the position of ctenophores was not due to artifactual attraction by the long branch leading to medusozoan cnidarians (Hydrozoa + Scyphozoa) (see Figure 1), by an analysis excluding these species (Figure S4): ctenophores still grouped with anthozoan cnidarians (a short branch), with high support (BS = 91%).

Our results not only suggest that ctenophores are the sister group to cnidarians but also that eumetazoans are monophyletic, implying single acquisition during animal evolution of nerve and muscle cells and/or the digestive system, in line with conventional ideas. These findings are at odds with the schemes of eumetazoan polyphyly proposed in two other recent phylogenomic studies [3, 11], both of which used more limited taxonomic sampling of nonbilaterian metazoans and more phylogenetically distant outgroups. It is clearly premature to make a final conclusion on basal metazoan relationships, because not all our analyses yielded significant statistical support values, and the influence of outgroup taxon sampling on tree topology might indicate that there is conflict in the data. As additional data from more nonbilaterian species become available, the remaining doubts should finally be resolved. It should be noted that the position of the placozoan *Trichoplax* with respect to sponges and eumetazoans remains poorly supported in our analyses (Figure 1) and that recent investigations focused on placozoan relationships [11, 35] provided contradictory results, leaving this question unresolved.

Body Plan Evolution among the Eumetazoans

The proposed restoration of the Coelenterata implies that cnidarians and ctenophores are phylogenetically equally related to the bilaterians and has implications with respect to the origin of mesoderm and of bilateral symmetry. These body plan features have been classically thought to be evolutionary innovations of the Bilateria, but their origin has been suggested to date back to the common cnidarian-bilaterian ancestor from recent developmental gene evidence [36–38]. The mesoderm-like muscle cell lineage of ctenophores [37] might be homologous with the mesoderm of the Bilateria and with mesoderm-like derivatives previously identified in cnidarians [37, 39]. Concerning symmetry, parsimony optimization favors an independent evolution of anatomical bilaterality in the bilaterians and in anthozoan cnidarians (Figure 3), but the significance of the biradial anatomy of the ctenophores [21] remains to be evaluated, for instance through the study of the developmental regulatory genes unilaterally expressed in cnidarians and in the bilaterians [38].

Our new proposal of basal metazoan relationships provides a stimulating framework for furthering our understanding of early metazoan evolution. It suggests that several key features of metazoan body plans were affected by events of convergence or reversion (Figure 3), contrasting with the traditional conception of metazoan evolution dominated by a gradual increase in morphological complexity. It should motivate detailed exploration of many aspects of character transformations during evolution, development, and metamorphosis, as well as the relationships of larval to adult traits.

Experimental Procedures

EST Sequencing

Fresh samples of *Sycon raphanus, Oscarella lobularis*, and *Oopsacas minuta* were collected in the Mediterranean near Marseille (France). *Ephydatia muelleri* gemmules from Belgium were incubated in the lab until production of young adult sponges. Samples of *Heterochone calyx* were collected in British Columbia (Canada) and re-aggregated tissue was used as starting material. *Carteriospongia foliascens* was collected at Lizard Island (Great Barrier Reef, Australia) and *Leucetta chagosensis* at North Stradbroke Island (Australia). *Pleurobrachia pileus* adults were collected in Villefranche-sur-Mer (France). For *Clytia hemisphaerica*, the starting material was a strain cultured at the Marine Station in Villefranche-sur-Mer. Frozen samples, RNA Later (QIAGEN)-preserved, or extracted total RNA (depending on the species) were sent to Genome Express (*O. minuta*), RZPD (*S. raphanus, O. lobularis, E. muelleri*), Express Genomics (*P. pileus* and *C. hemisphaerica*), and the Max Planck Institute for Molecular Genetics in Berlin (Germany) (*H. calyx, C. foliascens, L. chagosensis*) for cDNA library construction. ESTs were sequenced at the Max Planck Institute for Molecular Genetics (Berlin, Germany) (*H. calyx, C. foliascens, L. chagosensis*) or at the Genoscope (Evry, France) (all other species). Numbers of sequenced ESTs were approximately 2,000 (*O. minuta, E. muelleri, S. raphanus, O. lobularis*), 4,000 (*H. calyx, C. foliascens, L. chagosensis*), 30,000 (*P. pileus*), and 90,000 (*C. hemisphaerica*). All these newly sequenced EST collections are publicly available in dbEST/GenBank (http://www.ncbi.nlm.nih.gov/dbEST/). The alignment used for phylogenetic analyses is provided as Supplemental Data.

Data Assembly

We built upon phylogenomic data sets previously assembled [13, 40]. These alignments were updated, via the protocol described in [41], with the addition of newly generated sequences, and of sequences publicly available from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) and the EST Database (http://www.ncbi.nlm.nih.gov/dbEST/) of GenBank at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). In addition, 23 new genes sampled for at least two main poriferan clades were added. All these genes are likely to be orthologs because they are in single-copy in most of the opisthokonts, few recent duplications being observed mostly in vertebrates and *Drosophila*. To further evaluate the possibility of hidden paralogy, we inferred single-gene phylogenies and looked for any strongly supported conflict with the super-matrix tree according to protocol described in [42]. At a bootstrap threshold of 70%, conflicts were observed for only 6.5% of the testable bipartitions, less than the expected error rate. None of these conflicts could be easily explained by hidden paralogy (see details on these analyses in the Supplemental Experimental Procedures).

As previously demonstrated [13], taxon sampling has a major impact in phylogenomic studies. In addition to the nine sponges, nine cnidarians, three ctenophores, and one placozoan available, we therefore selected 22 slowly evolving representative taxa among available Bilateria (based on previous branch length comparison). To reduce the potential impact of long-branch attraction (LBA) [34], we also incorporated all available ichthyosporeans and choanoflagellates (taxa hypothesized to be the closest unicellular relatives of Metazoa) to break the long-branch leading to the distantly related fungal outgroup (for which only the slow-evolving chytridiomycetes and zygomycetes were used).

Ambiguously aligned regions were removed with Gblocks [43]. Sequence selection and concatenation were performed with SCaFoS [44]. To reduce the amount of missing data in the final alignment, we discarded undersampled genes. Only genes sampled for at least two-thirds of the species (36 out of 55) were retained. The resulting gene selection (128 genes) yielded an alignment of 30,257 unambiguously aligned positions. For all but two genes, the four major diploblast lineages (Porifera, Cnidaria, Ctenophora, and Placozoa) were represented by at least one species; at least three of the main poriferan clades (Demospongiae, Hexactinellida, Homoscleromorpha, Calcispongia) were represented for 65% of the genes.

Chapter 4

OrthoSelect: A web server for selecting orthologous gene alignments from EST sequences

Citation

Fabian Schreiber, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A web server for selecting orthologous gene alignments from EST sequences.
Nucleic Acids Research (2009) 37, W185-W188.

Original Contribution

FS developed, implemented, tested the webserver, and wrote the manuscript.

Published online 2 June 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W185–W188***
doi:10.1093/nar/gkp434

OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences

Fabian Schreiber[1,2,*], Gert Wörheide[2] and Burkhard Morgenstern[1]

[1]Institut für Mikrobiologie und Genetik, Abteilung für Bioinformatik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077, Göttingen, and [2]Molecular Geo- & Palaeobiology, Department of Earth- and Environmental Sciences & GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Straße 10, 80333 München, Germany

Received February 13, 2009; Revised April 24, 2009; Accepted May 11, 2009

ABSTRACT

In the absence of whole genome sequences for many organisms, the use of expressed sequence tags (EST) offers an affordable approach for researchers conducting phylogenetic analyses to gain insight about the evolutionary history of organisms. Reliable alignments for phylogenomic analyses are based on orthologous gene sequences from different taxa. So far, researchers have not sufficiently tackled the problem of the completely automated construction of such datasets. Existing software tools are either semi-automated, covering only part of the necessary data processing, or implemented as a pipeline, requiring the installation and configuration of a cascade of external tools, which may be time-consuming and hard to manage. To simplify data set construction for phylogenomic studies, we set up a web server that uses our recently developed OrthoSelect approach. To the best of our knowledge, our web server is the first web-based EST analysis pipeline that allows the detection of orthologous gene sequences in EST libraries and outputs orthologous gene alignments. Additionally, OrthoSelect provides the user with an extensive results section that lists and visualizes all important results, such as annotations, data matrices for each gene/taxon and orthologous gene alignments. The web server is available at http://orthoselect.gobics.de.

INTRODUCTION

The rapid development of genome-sequencing techniques has led to the generation of complete genome sequences for >600 species. Most of these sequences belong to model organisms, covering only small portions of the tree of life. The generation of massive numbers of expressed sequence tag (EST) libraries that can now be sequenced inexpensively by third-generation sequencing is a cheap alternative to whole genome sequencing, and has also provided a wealth of phylogenetically relevant data. Several recent phylogenomic studies have used EST sequences to generate large data matrices (1–4).

These studies generated and assembled EST sequences, which were screened for orthologous sequence regions to build useful orthologous gene alignments. Orthologous sequences result from a speciation event, and are likely to have a conserved function, whereas paralogous sequences evolve through a gene duplication event within a species, and are less likely to maintain their original function, due to processes such as neo-/or subfunctionalization (5).

Orthologous and paralogous together are called homologues (6). Since the prime goal of building reliable phylogenetic trees is to decipher the evolutionary relationships among organisms based on their shared common ancestry, only orthologous sequences should be used.

A reliable protocol is needed to build sets of orthologous sequences from EST libraries for successive phylogenomic analyses. We recently proposed such a protocol, which we called OrthoSelect (Schreiber *et al.*, manuscript submitted). The workflow of the protocol is outlined in Figure 1. The main idea is to keep user interaction simple, by simultaneously using state-of-the-art methods for orthology assignment, EST translation and elimination of paralogues, as well as construction and automated refinement of multiple sequence alignments. OrthoSelect has been extensively tested and proven to be a useful tool for managing this complex task.

Here, we present a web interface to OrthoSelect, the first web-based EST analysis pipeline for constructing orthologous gene alignments from EST libraries. Our web

*To whom correspondence should be addressed. Tel: +49 (0) 551 3913884; Fax: +49 551 3914929; Email: Fschrei@gwdg.de server does not require any kind of installation or testing. The user simply uploads EST libraries and chooses parameters (or uses default settings) to conduct the analysis. OrthoSelect then provides the user with a dataset useful for subsequent phylogenetic analysis, as well as numerous helpful data and statistics, such as annotations, a data matrix showing EST assignments to the orthologous groups (OGs) and visualizations of the orthologous gene alignments.

WEB SERVER

The main purpose of our web server is the construction of orthologous gene alignments from assembled EST libraries or other nucleotide sequences. After the user uploads pools of EST sequences, ESTs are assigned to OGs and the sequences most likely to be orthologous—in case there were multiple sequences per species—are used to compute an alignment that is post-processed in a final step. The workflow of the pipeline is depicted in Figure 1, and is described in more detail in the next section.

METHODS

Using the OGs defined by the eukaryotic orthologous groups (KOG) database (7), each EST is assigned to the closest OG. The assignment is done using a reimplementation of BLASTO (8) that clusters hits from a similarity search of the EST against the KOG database. The similarity between a query sequence and an OG is defined as the mean *E*-value between the query and the sequences from the OG. ESTs are then translated using a standard six-frame translation method.

We then translate the ESTs using the tools ESTScan (9) and GeneWise, (10) to account for frame shift errors. Considering the best Blast hit of the EST as a reference sequence, our program selects the translated sequence that is most similar to the reference sequence. Only OGs with at least three taxa are further considered.

At this stage of the analysis, it is possible to preselect individual or groups of taxa. The set of OGs is then further reduced to contain only OGs containing all preselected taxa. Redundant (e.g. paralogous) sequences are removed from each OG. This is done by considering only the sequence from each species that maximizes a global alignment score as being most likely orthologous. All sequences from each orthologous group are then aligned using either Muscle (11), T-Coffee (12) or DIALIGN-TX (13). These alignments are used to build hidden Markov models (14) that will be used to search the EST libraries for additional hits. Gblocks (15) is subsequently used to remove ambiguously aligned alignment columns. Since EST sequences may only partially cover genes, there is an option to exclude sequences from the alignment that are too short. This procedure outputs gene alignments whose member sequences are the ones most likely to be orthologous, given the dataset.

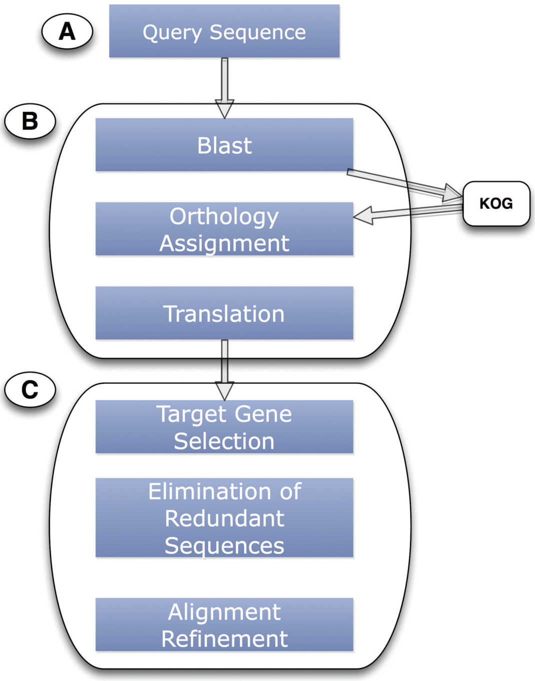


Figure 1. The main workflow of OrthoSelect: Each EST (**A**) is assigned to a pre-defined orthologous group (OG) by the KOG database, and translated (**B**). After all ESTs have been assigned to OGs, a subset of the OGs can be selected which will be further processed to exclude all redundant sequences, compute a sequence alignment and refine it in the last step (**C**).

INPUT

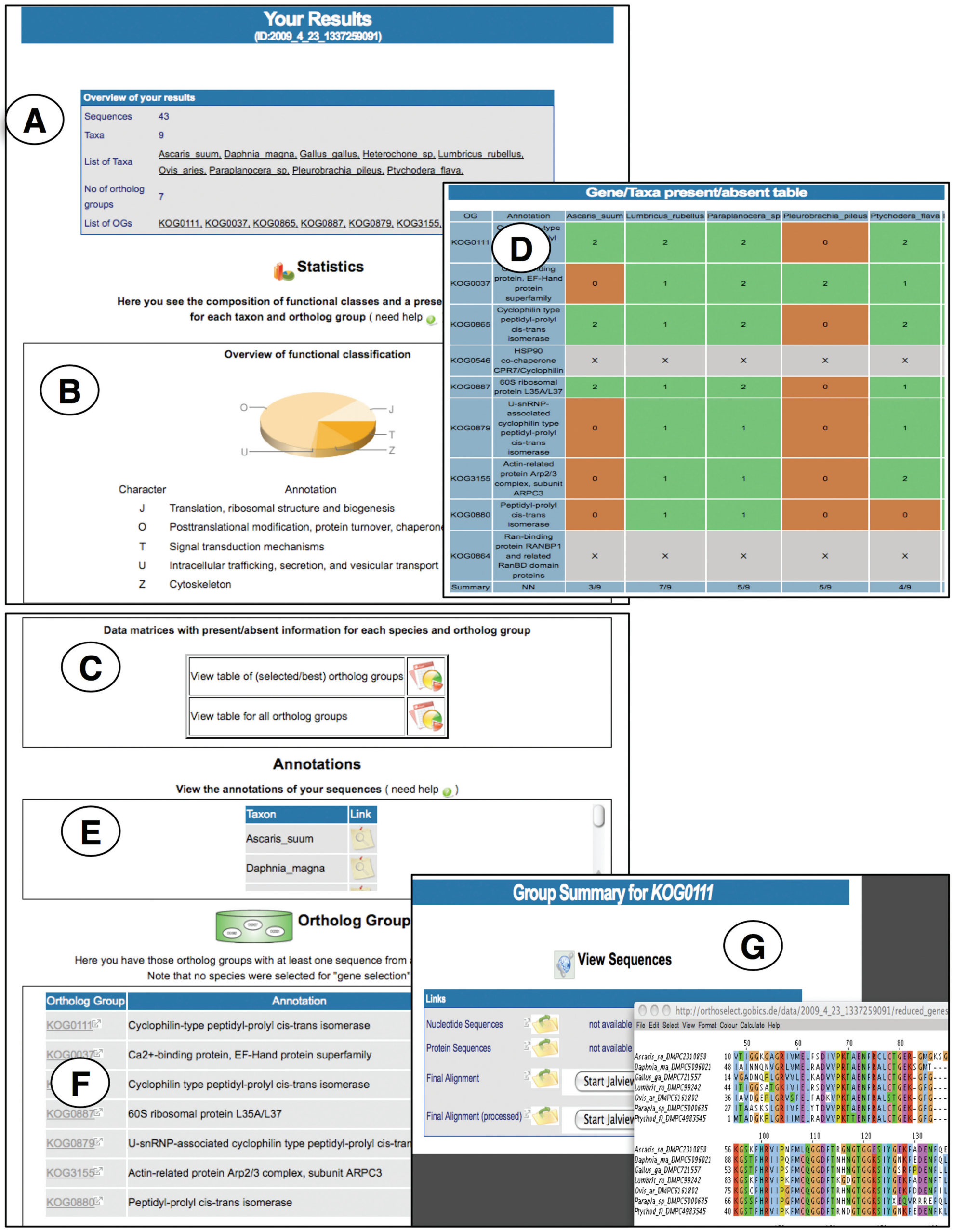
Our web server allows the use of OrthoSelect with default or adapted parameter values, e.g. the *E*-value for similarity searches using Blast (16) or methods for computing multiple sequence alignments. OrthoSelect accepts nucleotide sequences in FASTA format.

In the absence of a standard format for sequence identifiers in FASTA headers, sequence identifiers have to be adapted at some stage of a phylogenetic analysis to allow viewing taxa.

OrthoSelect requires the FASTA header to be in a certain format (the first word up to the first whitespace is taken as an accession number), and uploaded files have to match that format, or can be adapted using a converter supplied on the web page. Several syntax checks for the uploaded EST sequences have been implemented to ensure optimal performance from OrthoSelect. Our web interface offers the possibility to set up sequence identifiers (e.g. abbreviated taxon names) that will be used throughout the analysis.

Furthermore, the user can preselect one taxon or several taxa.

Our web server will then return a list of those OGs to which the submitted ESTs have been assigned, as well as a subset of those OGs containing the preselected taxa. The maximum number of input sequences is 30 000, and the maximum number of EST libraries is 10. An email address has to be supplied, since notification about the results will be sent via email.



Your Results
(ID:2009_4_23_1337259091)

Overview of your results	
Sequences	43
Taxa	9
List of Taxa	Ascaris_suum, Daphnia_magna, Gallus_gallus, Heterochone_sp, Lumbricus_rubellus, Ovis_aries, Paraplanocera_sp, Pleurobrachia_pileus, Ptychodera_flava,
No of ortholog groups	7
List of OGs	KOG0111, KOG0037, KOG0865, KOG0887, KOG0879, KOG3155,

Statistics

Here you see the composition of functional classes and a prese
for each taxon and ortholog group (need help)

Overview of functional classification

Character	Annotation
J	Translation, ribosomal structure and biogenesis
O	Posttranslational modification, protein turnover, chaperone
T	Signal transduction mechanisms
U	Intracellular trafficking, secretion, and vesicular transport
Z	Cytoskeleton

Gene/Taxa present/absent table

OG	Annotation	Ascaris_suum	Lumbricus_rubellus	Paraplanocera_sp	Pleurobrachia_pileus	Ptychodera_flava
KOG0111	C[illegible]-type [illegible]yl	2	2	2	0	2
KOG0037	[illegible]ding protein, EF-Hand protein superfamily	0	1	2	2	1
KOG0865	Cyclophilin type peptidyl-prolyl cis-trans isomerase	2	1	2	0	2
KOG0546	HSP90 co-chaperone CPR7/Cyclophilin	x	x	x	x	x
KOG0887	60S ribosomal protein L35A/L37	2	1	2	0	1
KOG0879	U-snRNP-associated cyclophilin type peptidyl-prolyl cis-trans isomerase	0	1	1	0	1
KOG3155	Actin-related protein Arp2/3 complex, subunit ARPC3	0	1	1	0	2
KOG0880	Peptidyl-prolyl cis-trans isomerase	0	1	1	0	0
KOG0864	Ran-binding protein RANBP1 and related RanBD domain proteins	x	x	x	x	x
Summary	NN	3/9	7/9	5/9	5/9	4/9

Data matrices with present/absent information for each species and ortholog group

View table of (selected/best) ortholog groups

View table for all ortholog groups

Annotations

View the annotations of your sequences (need help)

Taxon	Link
Ascaris_suum	
Daphnia_magna	

Ortholog Group

Here you have those ortholog groups with at least one sequence from
Note that no species were selected for "gene selection"

Ortholog Group	Annotation
KOG0111	Cyclophilin-type peptidyl-prolyl cis-trans isomerase
KOG0037	Ca2+-binding protein, EF-Hand protein superfamily
K[illegible]	Cyclophilin type peptidyl-prolyl cis-trans isomerase
KOG0887	60S ribosomal protein L35A/L37
KOG0879	U-snRNP-associated cyclophilin type peptidyl-prolyl cis-tran
KOG3155	Actin-related protein Arp2/3 complex, subunit ARPC3
KOG0880	Peptidyl-prolyl cis-trans isomerase

Group Summary for *KOG0111*

View Sequences

Links		
Nucleotide Sequences		not available
Protein Sequences		not available
Final Alignment		Start Jalview
Final Alignment (processed)		Start Jalview

http://orthoselect.gobics.de/data/2009_4_23_1337259091/reduced_genes

File Edit Select View Format Colour Calculate Help

```
                                 50        60        70        80
Ascaris_su_DMPC2310858    10 VTIGGKGAGRIVMELFSDIVPKTAENFRCLCTGER-GMGKSG
Daphnia_ma_DMPC5096021    48 IAINNQNVGRLVMELRADVVPRTAENFRALCTGEKSGMT---
Gallus_ga_DMPC721557      14 VGADNQPLGRVVLELKADVVPKTAENFRALCTGEK-GFG---
Lumbric_ru_DMPC99242      44 ITIGGSATGKIVIELRSDVVPKTAENFRALCTGEK-GFG---
Ovis_ar_DMPC6161802       36 IAVDGEPLGRVSFELFADKVPKTAENFRALSTGEK-GFG---
Parapla_sp_DMPC5000685    27 ITAASKSLGRIVFELYTDVVPKTAENFRALCTGEK-GFG---
Ptychod_fl_DMPC4983545     1 MTADGKPLGRIIMELRADVVPKTTENFRALCTGEK-GFG---

                                   100       110       120       130
Ascaris_su_DMPC2310858    56 KGSKFHRVIPNFMLQGGDFTRGNGTGGESIYGEKFADENFQE
Daphnia_ma_DMPC5096021    88 KGSTFHRIIPQFMCQGGDFTNHNGTGGKSIYGNKFEDENFQL
Gallus_ga_DMPC721557      53 KGSTFHRVIPSFMCQGGDFTNHNGTGGKSIYGSRFPDENFLL
Lumbric_ru_DMPC99242      83 KGSKFHRVIPKFMCQGGDFTKGDGTGGKSIYGEKFADENFTL
Ovis_ar_DMPC6161802       75 KGSCFHRIIPGFMCQGGDFTRHNGTGGKSIYGEKFDDENFIL
Parapla_sp_DMPC5000685    66 KGSSFHRIIPGFMCQGGDFTNHNGTGGKSIYXEQVRRREFQL
Ptychod_fl_DMPC4983545    40 KGSTFHRVIPKFMCQGGDFTRNDGTGGKSIYGNKFEDENFKL
```

Figure 2. The output of the OrthoSelect web server. Besides a general overview page of the results (**A**), our web server reports information about functional annotations (**B**), a gene/taxa presence/absence matrix (**C**, **D**), annotations for each taxon (**E**), as well as an overview of the orthologous groups (**F**). Additionally, for each orthologous group the resulting alignments are visualized using the Jalview (17) applet (**G**).

OUTPUT

Having generated EST libraries for the species under study, one of the main questions that arise is what genes those EST libraries have in common. These set of common genes can be used as a base for subsequent phylogenetic analysis. Our web server outputs those genes present in all EST libraries, but also provides additional information that will help the user to interpret the data and to decide which data are useful as input for phylogeny programs. The web interface offers a wide range of diagrams, charts, tables, etc. to supply the user with useful information (Figure 2). The most important part is the graphical representation of individual OGs with all assigned and translated EST sequences, and an overview of its taxonomical composition. Single sequences can be viewed along with their translation, as well as the computed multiple sequence alignment prior and subsequent to the final post-processing step in which the program Gblocks or Aliscore are used. The alignment is visualized using the Jalview (17) applet. The web server outputs an overview of the ESTs' functional classifications and OG assignments as a data matrix with presence/absence information for each gene and species in the study, and annotations for each species. The data matrix shows how many sequences from which taxa have been assigned to an OG. This way, the user can easily select OGs with all or a certain percentage of taxa present.

Besides an overview of all OGs with sequences assigned ('All orthologous groups'), OrthoSelect automatically builds a subset of OGs ('Best orthologous groups') that have either at least three different taxa or the predefined taxa present. The 'all orthologous groups' contain all orthologous groups to which sequences have been assigned, whereas the 'best orthologous groups' only contain one sequence per taxon (see Methods section).

The results page is intended to give the user an elaborate overview and useful information, but also provides all results to be downloaded for further examination and use in phylogenetic studies.

DESIGN AND IMPLEMENTATION

The OrthoSelect server consists of a web interface, a MySQL database management system (DBMS), and the core program OrthoSelect. The web interface for OrthoSelect has been constructed using Grails, which is a web application framework that uses the Groovy scripting language on the Java platform to help standardize the development of web interfaces (http://www.grails.org). Grails follows the idea of keeping data and web pages separate with a controller functioning as a mediator between them. All jobs are split into equal chunks to be computed in parallel on our computer cluster, and all data is stored in the DBMS. The average runtime for 20 000 ESTs with an approx. length of 500 bp is 5 h.

ACKNOWLEDGEMENTS

We thank Katharina Hoff for critically reading the manuscript, Dirk Erpenbeck for testing the web server and Rasmus Steinkamp for supporting the deployment of the web server.

FUNDING

German Research Foundation (DFG, Project Wo896/6-1,2) within DFG Priority Programme SPP 1174 'Deep Metazoan Phylogeny'. Funding for open access charge: Deptartment of Bioinformatics, Georg-August-Universität Göttingen.

Conflict of interest statement. None declared.

REFERENCES

1. Bourlat,S.J., Juliusdottir,T., Lowe,C.J., Freeman,R., Aronowicz,J., Kirschner,M., Lander,E.S., Thorndyke,M., Nakano,H. and Kohn,A.B. (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*, **444**, 85–88.
2. Dunn,C.W., Hejnol,A., Matus,D.Q., Pang,K., Browne,W.E., Smith,S.A., Seaver,E., Rouse,G.W., Obst,M. and Edgecombe,G.D. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.
3. Delsuc,F., Brinkmann,H., Chourrout,D. and Philippe,H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
4. Philippe,H., Derelleand,R., Lopez,P., Pick,K., Borchiellini,C., Boury-Esnault,N., Vacelet,J., Renard,E., Houliston,E. and Queinnec,E. (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biol*, **19**, 706–712.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Ann. Rev. Genet.*, **39**, 309–338.
6. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
7. Tatusov,R., Fedorova,N., Jackson,J., Jacobs,A., Kiryutin,B., Koonin,E., Krylov,D., Mazumder,R., Mekhedov,S. and Nikolskaya,A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
8. Zhou,Y. and Landweber,L.F. (2007) BLASTO: a tool for searching orthologous groups. *Nucleic Acids Res.*, **35**, W678–W682.
9. Lottaz,C., Iseli,C., Jongeneel,C.V. and Bucher,P. (2003) Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, **19(Suppl. 2)**, ii103–ii112.
10. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
11. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
12. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
13. Subramanian,A., Kaufmann,M. and Morgenstern,B. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.
14. Durbin,R., Eddy,S. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.
15. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402,.
17. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Chapter 5

OrthoSelect: A protocol for selecting orthologous groups in phylogenomics

Citation

Fabian Schreiber, Kerstin S. Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
OrthoSelect: A protocol for selecting orthologous groups in phylogenomics.
BMC Bioinformatics (2009) 10, 219.

Original Contribution

FS developed, implemented, tested *OrthoSelect*, and wrote the manuscript.

BMC Bioinformatics

BioMed Central

Software

Open Access

OrthoSelect: a protocol for selecting orthologous groups in phylogenomics

Fabian Schreiber*[1,2], Kerstin Pick[2], Dirk Erpenbeck[2], Gert Wörheide[2] and Burkhard Morgenstern[1]

Address: [1]Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany and [2]Department für Geo- und Umweltwissenschaften, Ludwig-Maximilians-Universität, Richard-Wagner-Str. 10, 80333 München, Germany

Email: Fabian Schreiber* - fab.schreiber@gmail.com; Kerstin Pick - kpick@uni-goettingen.de; Dirk Erpenbeck - erpenbeck@lmu.de; Gert Wörheide - woerheide@lmu.de; Burkhard Morgenstern - burkhard@gobics.de

* Corresponding author

Published: 16 July 2009

BMC Bioinformatics 2009, **10**:219 doi:10.1186/1471-2105-10-219

Received: 22 October 2008

Accepted: 16 July 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/219

Abstract

Background: Phylogenetic studies using expressed sequence tags (EST) are becoming a standard approach to answer evolutionary questions. Such studies are usually based on large sets of newly generated, unannotated, and error-prone EST sequences from different species. A first crucial step in EST-based phylogeny reconstruction is to identify groups of orthologous sequences. From these data sets, appropriate target genes are selected, and redundant sequences are eliminated to obtain suitable sequence sets as input data for tree-reconstruction software. Generating such data sets manually can be very time consuming. Thus, software tools are needed that carry out these steps automatically.

Results: We developed a flexible and user-friendly software pipeline, running on desktop machines or computer clusters, that constructs data sets for phylogenomic analyses. It automatically searches assembled EST sequences against databases of orthologous groups (OG), assigns ESTs to these predefined OGs, translates the sequences into proteins, eliminates redundant sequences assigned to the same OG, creates multiple sequence alignments of identified orthologous sequences and offers the possibility to further process this alignment in a last step by excluding potentially homoplastic sites and selecting sufficiently conserved parts. Our software pipeline can be used as it is, but it can also be adapted by integrating additional external programs. This makes the pipeline useful for non-bioinformaticians as well as to bioinformatic experts. The software pipeline is especially designed for ESTs, but it can also handle protein sequences.

Conclusion: OrthoSelect is a tool that produces orthologous gene alignments from assembled ESTs. Our tests show that OrthoSelect detects orthologs in EST libraries with high accuracy. In the absence of a gold standard for orthology prediction, we compared predictions by OrthoSelect to a manually created and published phylogenomic data set. Our tool was not only able to rebuild the data set with a specificity of 98%, but it detected four percent more orthologous sequences. Furthermore, the results OrthoSelect produces are in absolut agreement with the results of other programs, but our tool offers a significant speedup and additional functionality, e.g. handling of ESTs, computing sequence alignments, and refining them. To our knowledge, there is currently no fully automated and freely available tool for this purpose. Thus, OrthoSelect is a valuable tool for researchers in the field of phylogenomics who deal with large quantities of EST sequences. OrthoSelect is written in Perl and runs on Linux/Mac OS X. The tool can be downloaded at http://gobics.de/fabian/orthoselect.php

Background

DNA and protein sequences provide a wealth of information which is routinely used in phylogenetic studies. Traditionally, single genes or small groups of genes have been used to infer the phylogeny of a group of species under study. It has been shown, however, that molecular phylogenies based on single genes often lead to apparently conflicting tree hypotheses [1]. The combination of a large number of genes and species in genome-scale approaches for the reconstruction of phylogenies can be useful to overcome these difficulties [2]. This approach has been termed *phylogenomics* [3].

Since complete genome sequences are available only for a limited number of species, many phylogenomic studies rely on EST sequences. EST sequences are short (~200 – 800 bases), unedited, randomly selected single-pass reads from cDNA libraries that sample the diversity of genes expressed by an organism or tissue at a particular time under particular conditions. The relatively low cost and rapid generation of EST sequences can deliver insights into transcribed genes from a large number of taxa. Moreover, EST sequences contain a wealth of phylogenetic information. Several recent phylogenomic studies used EST sequences to generate large data matrices, e.g. [4-7]. Such studies start with the generation of EST libraries for a set of species. Overlapping EST sequences from single coding regions are then assembled into contigs and orthologous genes are identified as a basis for phylogenetic reconstruction. Homologous sequences are called orthologs if they were separated by a speciation event, as opposed to paralogous sequences, which were separated by a duplication event within the same species [8]. If the last speciation event predates the gene duplication event, homologous sequences are called inparalogs [9]. Orthologs are usually functionally conserved whereas paralogs tend to have different functions [10] and are less useful in phylogenetic studies. (because true genealogical relationships among taxa can only be reconstructed with great difficulty.) A typical protocol for detecting orthologs in phylogenomic studies should include (1) a similarity search using tools like BLAST [11], (2) a strategy to select a subset of hits returned by this search, (3) a criterion to identify sequences as potential orthologs, (4) a strategy for eliminating potential paralogs – in case several sequences from the same species have been assigned as potential orthologs to the same orthologous group.

Orthology assignment is a crucial prerequisite for phylogeny reconstruction as faulty assumptions about orthology – e.g. the inclusion of paralogs – can lead to an incorrect tree hypothesis [12]. Errors can result from similarity searches against non-specialized databases, e.g. NCBI's *nr* database, or from best-hit selection strategies such as *best reciprocal hit* [13] or *best triangular hit* that may lead to false positive orthology predictions. The similarity between a query and a database sequence stemming from a similarity search – expressed for example as a bit-score or expectation value (E-Value) – is usually taken as a criterion to predict an orthologous relationship. Since the results of these methods depend on the choice of a database and on the strategy to select sequences from similarity search hits, a more reliable protocol for ortholog predictions is needed.

Several databases and computational methods for predicting orthologs are available. Multi-species ortholog databases have been developed based on different sources of orthologous information. They include information about orthologous relationships between sequences. The OrthoMCL-DB database [14] and the KOG database [15] have been constructed from whole genome comparisons, HomoloGene [16] on the basis of synteny. HOVERGEN [17] and TreeFam [18] were constructed using the orthologous information from phylogenetic trees. Two of these databases, OrthoMCL-DB and KOG, explicitly define orthologous groups (OG) which can be used as a source for orthology assignment of unknown sequences using similarity searches.

Most computational methods to identify orthologs are based on either a phylogenetic analysis, or on *all-against-all* BLAST searches [19]. The former approach is computationally expensive and usually requires manual intervention. *All-against-all* approaches use every sequence from the input data set as a query for BLAST searches against sequences from the respective other species. This generates OGs based on some similarity measure, e.g. using all best reciprocal hits. These OGs can further be processed to merge, delete, or seperate overlapping groups using a clustering algorithm, as implemented in e.g. OrthoMCL [20] or Inparanoid [21]. Zhou and Landweber [22] developed BLASTO, a different computational method for orthology prediction by including information from an orthologous database. Other important aspects in data set construction for phylogenetic analysis on a large scale are (1) correct identification of open reading frames in ESTs and their translation, (2) careful selection of target genes to maximize the phylogenetic information, (3) elimination of redundant sequences, and (4) a refinement step to select conserved blocks and remove homoplasy from multiple sequence alignments.

Nowadays, data sets in phylogenomic studies can easily contain dozens of taxa and hundreds of genes [6]. The construction of data sets of that size for phylogenomic studies is time-consuming and can hardly be achieved manually. To the best of our knowledge, no software pipeline is currently available that performs the above steps automatically. Herein, we present a software pipeline, called OrthoSelect, to process clustered EST sequences automatically for phylogenomic studies. Our goal is to give both non-bioinformaticians and bioinformatic experts a useful framework to carry out analyses on a phylogenomic scale. It integrates publicly available bioinformatic tools and manages data processing and storage. Although the software pipeline is designed to automate the construction of data sets for phylogenomic studies, the user can evaluate intermediate results at any time of the analysis. OrthoSelect produces automatically calculated and post-processed alignments that can be used as input for common phylogenetic reconstruction software. In a large-scale study, we applied OrthoSelect to a data set from metazoan species consisting of > 950, 000 ESTs belonging to 71 taxa (unpublished data). In order to assess the quality of OrthoSelect predictions in relation to results obtained from other methods, we compared OrthoSelect to the manually created and published phylogenomic data set by Dunn et al. [6]. Since our tool offers an increased functionality compared to other tools for orthology prediction (e.g. OrthoMCL), our tests focus on the assignment of orthology only, and do not cover the correct translation of ESTs, gene selection, alignment computation, and alignment postprocessing.

Implementation

Our software pipeline is written in *PERL* and uses BioPerl [23]. The main workflow is depicted in Figure 1. The entire analysis is guided by a configuration file and several *PERL* scripts. OrthoSelect can be run on a single desktop computer as well as on a computer cluster using a batch system, e.g. a Sun Grid Engine [24]. Required programs are *BLAST* for the similarity search, *ESTScan* [25] and *GeneWise* [26] for translating ESTs, and a software program for multiple sequence alignment. *ClustalW* and *MUSCLE* are needed for computing the pairwise sequence alignments. Our software supports multiple alignments computed by *MUSCLE* or *T-Coffee*, but it can easily be adapted to accept multiple alignments calculated by other programs. *Gblocks* [27], *Noisy* [28] and Aliscore [29] are used to select informative alignment columns. OrthoSelect offers the possibility to automatically download and install all missing required programs on the computer.

Program outline

In contrast to the above outlined methods for the identification of orthologs based on whole genome comparisons, we adopted an approach that compares EST sequences to predefined groups of orthologous genes. We developed a software pipeline that uses a reimplementation of BLASTO, an extension of BLAST that clusters BLAST hits using predefined orthologous groups from an ortholog database. Here, the similarity between a query sequence and an OG is defined as the mean E-value between the query and the sequences from the OG (see Figure 2). As input data, it takes a library of EST sequences together with a database of orthologous genes. We assume that the basic pre-processing steps such as end clipping and vector trimming have already been done and that the ESTs are already assembled into contigs. As a database of orthologs, either KOG or OrthoMCL-DB can be used.

Using the orthologous groups (OG) defined by KOG or OrthoMCL-DB as a basis, orthologous ESTs are detected by a similarity search of ESTs against the ortholog database and assigning them to the OGs using our reimplementation of BLASTO. The ESTs are then translated and stored. Redundant sequences within each OG are eliminated and an alignment of the remaining sequences is computed. In a last step, we use sophisticated post-processing methods to filter out non-informative or misleading information from the alignment (see Figure 1). The entire analysis is guided by a configuration file containing the main parameters and options for each external program.

Orthology Detection

The first step of the software pipeline comprises the detection of potential orthologs in EST libraries (see Figure 1, Point 1). This is a critical step, because false ortholog assignments can lead to serious errors in the resulting phylogenetic tree. Orthologs are detected by searching an ortholog database – either KOG or OrthoMCL-DB – with a query EST using *blastx* and subsequently the resulting hits are clustered according to an algorithm similar to that used in BLASTO. A standard BLAST search returns a list of hits ordered by their significance. By contrast, BLASTO calculates similarity values between the query sequence and entire groups of orthologs (OGs).

In BLASTO, the similarity between a query s and a OG g is defined as the average similarity between s and all sequences in g. In our approach, we modified this measure of similarity. For a query s and a OG g, we consider only the subset $g' \subset g$ that contains the best hit from each species. This is to compensate the many paralogs present in KOG [30], and to ensure a high probability of the EST sequence being orthologous to the sequences in the corresponding OG. The similarity score for a query s and an OG g is then calculated as

$$S_{g,s} = \sum_{f_i \in g'} -log(P_i) / |g'|$$

where

$$P_i = 1 - exp(-E_i)$$

Here, E_i is the E-value of the BLAST alignment of f_i with the query sequence s and $|g'|$ the number of species in g'. Finally, every EST sequence s is assigned to those orthologous groups g with a similarity score $S_{g,s}$ above a given

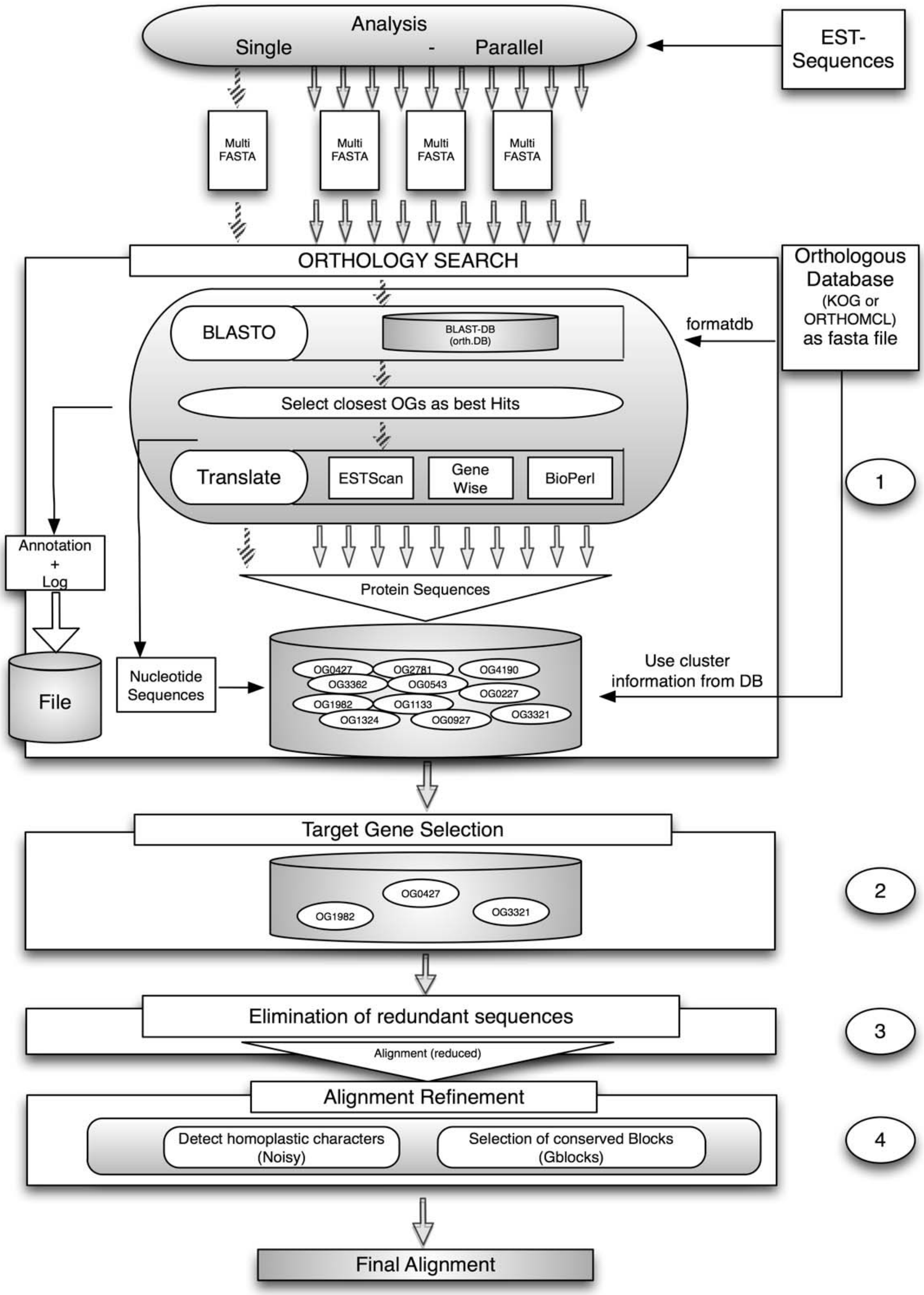


Figure 1

Workflow of OrthoSelect. The main workflow of the software pipeline to detect ortholog sequences in phylogenomic studies. Input are EST libraries and an ortholog database (either KOG or OrthoMCL) as multi-fasta files. The analysis comprises four parts. (1) The orthology detection – which can be performed on a single computer or a computer cluster – blasts each EST against the ortholog database, selects the closest ortholog group as the best hit and translates it and stored together with the nucleotide sequences in the corresponding OG. (2) Target genes can be selected. (3) The sequence most likely being an ortholog is selected by eliminating potential paralogs. (4) Informative alignment columns are selected to increase the phylogenetic signal.

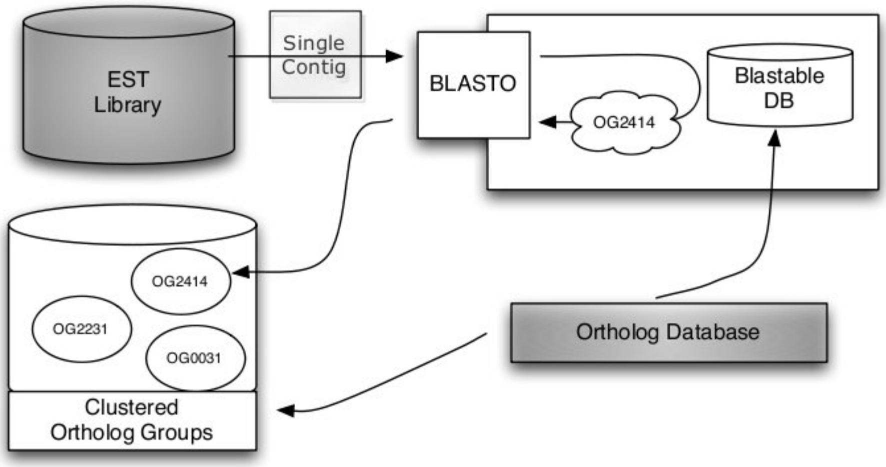


Figure 2
Workflow of orthology assignment. Workflow of our software pipeline. The two databases colored in green are to be supplied by the user. The ortholog database is converted into a BLAST database and clustered in ortholog groups. Each contig from the assembled EST library is assigned to the OG returned by a BLASTO search against the ortholog database.

threshold. We allow multiple assignments of a single EST, because ESTs can represent domains rather than full genes, and they should be assigned to all OGs containing that domain (E.g. the OGs KOG0100, KOG0101, KOG0102 of KOG all contain the same Pfam domain *HSP70*). All ESTs assigned to the same OG are now potential orthologous. Redundant sequences will be removed later (see section *Eliminating Redundancies*).

EST Translation

In the next step, potential coding regions in assembled EST sequences are detected and translated into proteins. By their nature, EST sequences often contain sequencing errors and may cover genes partially, only [31]. These errors result in e.g. reading frame shifts that make translation non-trivial. Several algorithms have been developed to overcome this problem. DIANA-EST [32] uses a combination of Artificial Neural Networks while ESTScan uses Hidden Markov Models. In contrast to this, DECODER [33] implements rule-based methods, and GeneWise uses a known protein as a template. In addition, combinations of these methods have been proposed to identify coding regions and to translate EST sequences correctly, e.g prot4EST [31]. We use a comparative approach of different well established programs for translation. Each EST is translated (using ESTScan, GeneWise, and a standard six-frame translation using BioPerl) and aligned to the best hit from the previous BLAST search using bl2seq [34]. The translated sequence with the lowest E-value is then chosen as the correctly translated sequence. This way, the probability of getting correctly translated ESTs is increased. Our goal was to fully automate the installation of all external programs. We did not include prot4EST since it requires additional programs and one of which is not freely available for download and therefore cannot be installed automatically.

Taxon/Gene Sampling Strategy

After the assembled EST sequences were assigned to predefined orthologous groups (OG) and translated into proteins, the next step consists of the proper selection of OGs suitable for phylogenetic analysis. Since EST libraries represent snapshots of expressed genes, not every OG will contain EST sequences from all species under study; some OGs may contain too few sequences and do not contain sufficient information for further consideration. We do not require every OG to contain all sequences of interest. There is no consensus about the influence of missing genes on the resulting phylogeny [35]. No reliable criterion, which OGs should be used for phylogenetic inference exists. Our software offers two alternative ways of selecting OGs:

1. The user selects a subset of individual species under study. In this case, those OGs will be selected that contain at least one EST from each of the user-selected species.

2. The user defines *groups* of species (e.g. groups that are thought to be monophyletic). Our tool will then select those OGs that contain at least one EST sequence for each of the specified groups.

The idea of these two methods is to select the maximal biclique of a graph with the nodes consisting of the OGs and the taxa – in case of option 1 – or monophyla – in case of option 2 [36]. The selection of genes according to these who methods focusses on maximising the phylogenetic signal in the dataset (see Figure 1, Point 2).

Eliminating Redundancies

Multiple divergent copies of the same gene and different levels of stringency during EST assembly can lead to a situation where OGs contain more than one sequence for each species (Depending on the size of the study, OGs can contain hundreds of sequences which makes manual elimination of redundant sequences impossible). It is also known that some of the orthologous groups contained in KOG contain not only orthologous genes but also paralogs [30]. In these cases, a fast and reliable method is needed to select the correct sequence for each species. We work with the assumption that a gene from one organism is often more similar to an orthologous gene from another organism than to paralogs from that organism. This seems plausible based on both the definition of orthology and the fact that orthologs typically retain the same function [10]. A scenario where a gene from one organism is more similar to a paralog rather than to its ortholog from another organism would require a considerable difference in the rate of paralog evolution [10]. Since this is more an exception than a rule and since OrthoSelect aims at the production of gene alignments containing only one sequence per species, we do not consider such cases.

All sequences belonging to the same OG are aligned in a pairwise manner to compute a distance matrix. Two types of distance matrices can be used to select the sequence from an organism that is most likely ortholog (see Figure 3):

1. An initial distance matrix as computed by alignment methods like *ClustalW* [37].

2. A normalized distance matrix selecting those sequences that have the highest percentage of matching positions in pairwise comparisons using *MUSCLE* [38,39].

The first option follows the idea that those sequences should be selected that optimize the alignment score in a global alignment. The second option takes into account that ESTs usually do not represent complete genes. Since a selection based on a standard distance matrix will penalize missing positions, longer paralogous sequences can be selected instead of shorter orthologous ones. The distance matrix used in the second option selects the sequence with the highest number of matching positions normalized by its length. The user can select one type of matrix to be used to eliminate redundant sequences (see Figure 1, Point 3). Based on that distance matrix, we want to select one sequence from each organism in such a way that the selected sequences are most probable to be ortholog to each other. Here, we use the following strategy: All sequences from one organisms are compared to all sequences from all other species. For each sequence *s* from a given species *S*, we count the number of species *S*' such that *s* has the shortest distance to a sequence from *S*' among all sequences from *S*' (if there are any such species *S*'). Formally, if the distance between sequences *s* and *s*' is denoted by *d*(*s*, *s*'), we count the number of species *S*' for which we have

$$d(s, s') = \min_{s \in S, s' \in S'} d(s, s')$$

For species *S*, we then select the sequence *s* for which this number is maximal (see Figure 3).

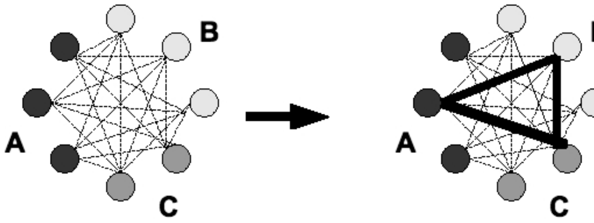


Figure 3
Eliminating redundant sequences. The figure shows how OrthoSelect eliminates redundant sequences. Here, we have an OG with three sequences from organism A and B and two sequences from organism C. All sequences are aligned in a pairwise manner to compute a distance matrix (left side). That sequence from an organism is selected that most often has the smallest distance to another organism, see section for details (right side).

Multiple Sequence Alignment

By default, the previously selected sequences are aligned using either *MUSCLE* or *T-Coffee* [40,41]. Other standard methods for multiple alignment can be used as well, e.g. *ProbCons* [42], *MAFFT* [43,44], *DIALIGN* [45,46] or *DIALIGN-TX* [47,48].

The computed alignments contain sequences that are most likely being orthologous given the data set. Nevertheless, there might be cases in which our comparative approach did not find the optimal translation (see section about EST Translation). To correct this, we use the software *hmmbuild* from the HMMER package to build profile hidden markov models (HMMs) from sequence alignments [49]. Additionally, the EST sequences from all taxa are translated using ESTScan. ESTScan is based on a HMM and was trained for species ranging from *Arabidopsis thal-*

iana to *Homo sapiens* by default. The translated sequence databases are then searched using hmmsearch from the HMMER package [50] and the HMM. The closest sequence from each taxon above a given threshold is taken as a hit. By this, we can find more similar as well as additional hits – hits that might have been overseen during the initial blastx search, because the EST sequence contained one or several frame shift errors. The workflow is depicted in Figure 4. The advantage of using a HMM is the possibility of finding that translated sequence that fits best to the whole existing alignment and not just to single sequences, as with standard Blast searches.

Once multiple alignments have been calculated for selected groups of ortholog EST sequences, these alignments can be further processed to exclude columns that are not suitable for phylogenetic analysis. Since not all parts of a gene evolve at the same rate, alignments typically contain highly conserved as well as less conserved sites. Alignment columns that are too conserved do not contain any phylogenetic signal. The same holds true for parts of the sequences that are too divergent to be correctly aligned. Another problem that confuses phylogenetic reconstruction is the presence of homoplasy caused by back- or parallel-mutation. Several programs have been developed to tackle these problems by automatically selecting sufficiently conserved blocks from alignments, for example *Gblocks* and *Aliscore*, or by eliminating potentially homoplastic sites, e.g. *Noisy*. *Gblocks*, *Aliscore*, and *Noisy* are incorporated in our software pipeline to allow a broad spectrum of alignment post-processing thereby increasing the accuracy of the subsequent phylogenetic analysis (see Figure 1, Point 4). Furthermore, alignments processed by *Gblocks* can be further filtered by discarding

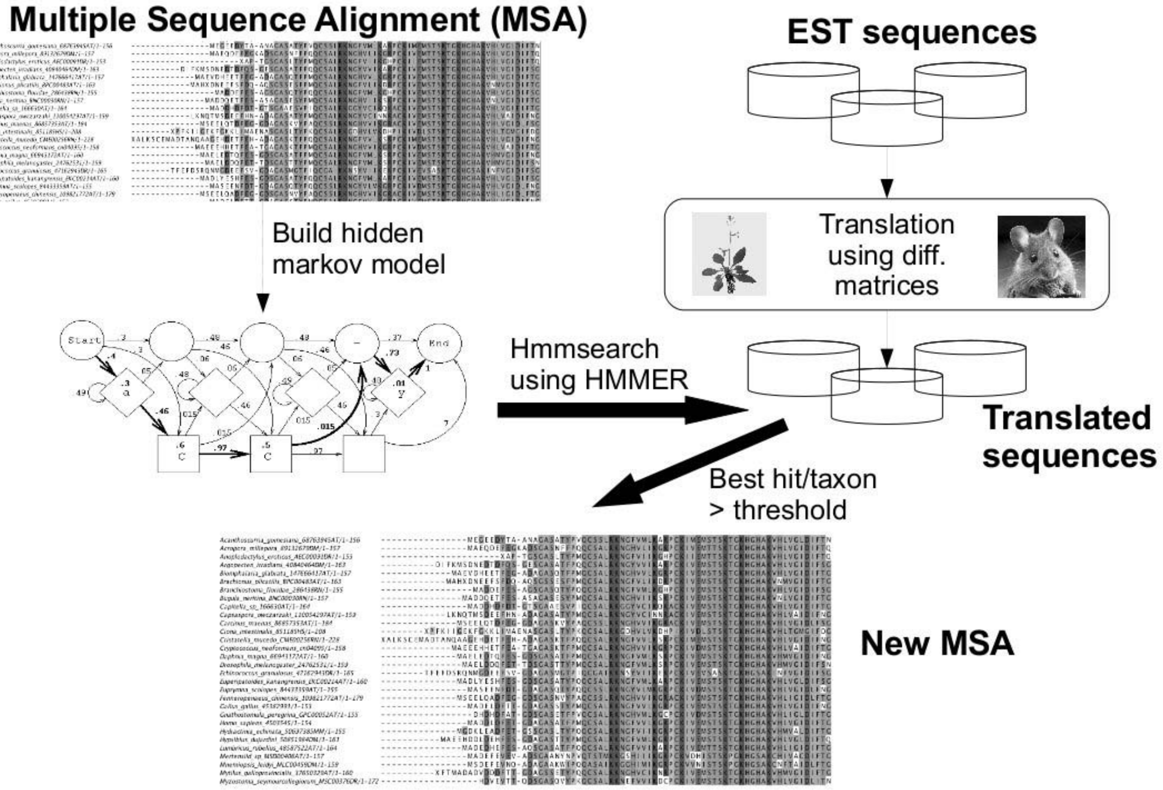

Multiple Sequence Alignment (MSA)
EST sequences
Build hidden markov model
Translation using diff. matrices
Start
End
Hmmsearch using HMMER
Translated sequences
Best hit/taxon > threshold
New MSA

Figure 4

Rebuilding the multiple sequence alignment. The figure illustrates how OrthoSelect refines the multiple sequence alignments (MSA) created so far. Based on the MSA a hidden Markov Model (HMM) is build. Additionally, all EST libraries are translated using ESTScan with different matrices (ranging from *Arabidopsis thaliana* to *Homo sapiens*). The software *hmmsearch* from the HMMER package then used the HMM to search all translated sequences and selecting the best hit from each taxon above a given threshold. From these hits the new MSA is then computed too short sequences from the alignment (e.g. sequences with > 50% missing characters).

Results and Discussion

OrthoSelect is the first fully automated and freely available tool that covers the whole process of selecting orthologs from EST libraries to output orthologous gene alignments that can be used to build phylogenies. In the absence of a gold standard for benchmarking of orthology prediction and in order to evaluate the performance of our program, we designed the following tests: First, OrthoSelect was compared to the best-hit selection strategy using a set of sequences from JGI with KOG-annotations. Second, we evaluated the performance compared to the KOG database by re-annotating (re-assinging) ortholog database sequences. In the third and most powerful test we compared OrthoSelect tool to a manually created and published phylogenomic data set. In this context, we also compared our tool with OrthoMCL.

OrthoSelect vs. Best-hit selection strategy

To evaluate the performance of our software pipeline and the best-hit selection strategy regarding correct orthology assignment, we used a data set comprised of transcribed genes and annotation files from 4 different species as shown in Table 1. The best-hit selection strategy assigns the query sequence to that OG the best hit belongs to. As ortholog database, we used KOG. The annotation files contain KOG classification and therewith the functional annotation for each sequence. Sequences and annotations were downloaded from the Department of Energy Joint Genome Institute (JGI) [51]. Since OrthoSelect makes annotations by assinging sequences to OGs of KOG, we considered an assignment of a sequence to an OG to be correct if it matches the KOG classification provided by JGI. To evaluate the performance of our classification system, we calculated for each species and an E-value cut-off of $1e$ - 10 the ratio of correctly assigned OGs, i.e. the number of correctly assigned sequences divided by the number of assigned sequences. Table 2 shows the result of the analysis. Our software pipeline reaches a correct assignment rate of ~93%, whereas the best-hit selection strategy assigns the sequences in ~79% of the cases to the correct OG. OrthoSelect outperforms the best-hit selection strategy and its very high rate of correct ortholog prediction should provide a good basis for subsequent phylogenetic analyses.

Table 1: Species used.

Species	*Sequences*	*KOG Classifications*
Daphnia pulex	30940	15806
Ostreococcus tauri	7725	4733
Trichoderma virens	11643	6879
Xenopus tropicalis	27916	27617

The table shows species that we used in our test runs along with the number of sequences from each sequence and the corresponding KOG classifications.

Table 2: Results from orthology assignment: OrthoSelect vs. Best-hit selection strategy.

Species	*Predictions*	*OrthoSelect*	*Best-hit strategy*
Daphnia pulex	12696	98%	86%
Ostreococcus tauri	4742	91%	76%
Trichoderma virens	5886	99%	87%
Xenopus tropicalis	18556	84%	69%

The table shows species that we used in our test runs along with the number of predictions and percentage of correct predictions made by OrthoSelect and the best-hit selection strategy respectively.

OrthoSelect vs. KOG

In absence of a reference dataset for orthology prediction and due to the fact that our tool is mainly focused on the automation of a process rather than being a completely new method for orthology prediction, we compared OrthoSelect to the KOG database by re-annotating (re-assinging) ortholog database sequences. We performed the following: 5000 sequences were randomly chosen and masked out from the ortholog database. The remaining sequences were converted into a blastable database. We then ran OrthoSelect using each of the 5000 sequences as a query sequence against the masked database. Assuming the original ortholog group assignment in the ortholog database represents the correct orthology relation, we calculated in how many cases our orthology assignment matched the original assignment. We could assign the query sequences in 92% of the cases to the correct ortholog group.

OrthoSelect vs. manually created data set by Dunn et al

The goal of our tool is to automate the process of constructing data sets that can be used for subsequent phylogenetic analyses. To test our tool regarding this, we selected Dunn et al.'s data set (hereafter referred to as reference data set) published in *Nature* [6].

This data set consists of newly sequenced ESTs as well as publicy available ESTs and protein sequences, and has been generated using all-vs.-all BLAST searches, protein translations using prot4EST, grouping of the sequences into orthologous groups using TribeMCL [52] as well as manual curation and tree reconciliation (see [6] for more details).

The reference data set as well as the single EST and protein sequences were either downloaded from publicly available sources or provided by Casey Dunn. The initial data set consisted of 150 genes and 77 taxa. In order to guarantee comparable results, we mapped each sequence from each gene to the KOG database using the best BLAST-hit. Only genes where all sequences could be mapped to the same KOG were further considered. This led to a considerable decrease in the number of genes. Since some taxa were not available for download, we ended up with 70 out of the 77 taxa Dunn et al. initially used.

For prediction of orthologous sequences, we denote a true positive as a correctly predicted ortholog, a false positive as an incorrectly predicted ortholog, and a false negative as an overlooked sequence. To be more precise, we use the following measures of performance:

- *Taxon is present in both alignments:* If the percentage identity of both sequences is above a threshold (≥ 95%), the sequences are regarded as being equal and counted as a true positive. Else, both sequences are aligned to a hidden markov model (HMM) build from the alignment of the corresponding orthologous group (OG) using hmmsearch from the HMMER package. If the OrthoSelect sequence is closer to the HMM, it will be counted as a true positive, and otherwise it will be counted as a false positive.

- *Taxon is present in the reference alignment, but not in the OrthoSelect alignment:* It will be counted as a false negative.

- *Taxon is present in the OrthoSelect alignment, but not in the reference data set:* The sequence is aligned to the HMM of that OG. If it shows significant similarity, it will be counted as a true positive, and otherwise as a false positive.

Furthermore, we use the following formula to measure the specificity of our results:

- *Specificity:*

$$\frac{\mathit{True\ Positives}}{\mathit{True\ Positives}+\mathit{False\ Positives}}$$

We get the following results (see also Table 3): With respect to the reference data set, our tool receives a specificity of 98%. This means that the predictions about orthology our tool makes are almost always true and almost all orthologous sequences contained in the original reference data set could be found. The number of false predictions is considerably small. Although we missed 8% of the orthologous sequences, we could find additional hits for 270 sequences. 268 of those additional sequences showed significant similarity to the rest of the alignment and were counted as true positives. 2 sequences were falsely predicted as being orthologous. This equals an increase of +4% of orthologous sequences. Compared to the reference data set, we can make the following statements: Our tool selects orthologous sequences from EST libraries and other sources with very high accuracy. OrthoSelect correctly translates the sequences and receives a higher specificity by finding more true positives. In phylogenomics, the use of EST data can result in data matrices – where the rows are genes and the columns are taxa or vice versa – with most of the cells being empty. Although there is no consensus about the impact of missing sequences on the resulting phylogeny, the additionally found sequences will have a beneficial effect.

Table 3: Results from orthology assignment: OrthoSelect vs. reference data set

Value	*OrthoSelect*
Specificity	98%
Cases where OrthoSelect found better sequences	63%
Number of additional sequences found	270
Number of additional sequences found (good)	268
Number of additional sequences found (bad)	2
Number of sequences missed	197
Ratio of additional/missed sequences	+4%

OrthoSelect vs. OrthoMCL

In order to further assess the performance of OrthoSelect, we compared it with OrthoMCL, another tool for orthology prediction. OrthoMCL takes a set of sequences and clusters them into groups of orthologous and inparalogous sequences. In contrast to OrthoSelect, OrthoMCL only handles protein sequences and produces clusters of orthologous sequences rather than multiple sequence alignments (see Figure 5). These generated clusters can contain considerably more than one sequence per taxon, and subsequently build multiple sequence alignments would not be comparable to the ones produced by OrthoSelect and Dunn et al. Nevertheless, we are interested in the performance of our tool compared to OrthoMCL. The previous test revealed that clustering algorithms of OrthoSelect and the method by Dunn et al. perform similarly. To check if clusters build by OrthoMCL are in agreement with the OrthoSelect clusters and thus with the Dunn clusters, we used the following 6 taxa: *Cryptococcus neoformans*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Saccharomyces cerevisiae*, and *Suberites domuncula*. The dataset has been reduced to include only protein sequences, because OrthoMCL deals with protein sequences, only.

For each of the 60 previously compared gene clusters (see previous section), we checked whether OrthoMCL assigns sequences from the 6 taxa to the same OrthoMCL cluster or not. The results were, that all sequences belonging to the same alignment have been clustered together by

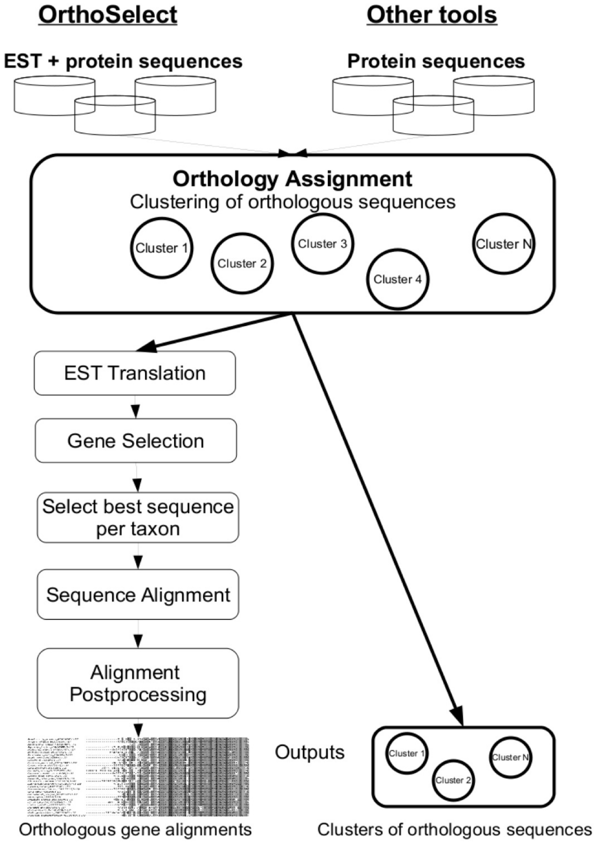


Figure 5

Overview of functionality of OrthoSelect compared to other tools. The figure illustrates the differences in functionality between OrthoSelect and other tool for orthology prediction. Both approaches have in common that they build clusters of orthologous sequences. Moreover, OrthoSelect can handle EST sequences and correctly translate them and further processes these clusters to select only one sequence per taxon, compute sequence alignments and refine them. In contrast to the other tools, OrthoSelect outputs orthologous gene alignments that can be directly used the subsequent phylogenetic analysis.

OrthoMCL. This means that the clustering algorithm of all methods produce similar results and converge.

Besides the additional functionality of OrthoSelect as compared to OrthoMCL and its usability for EST sequences, it is also much faster. It took OrthoMCL 24 hours to analyse the data set of 55.646 sequences. In contrast, our tool analysed the 1.000.000 sequences Dunn et al. used in about 6 hours.

Conclusion

OrthoSelect is a tool for finding ortholog groups in EST databases. It can be used by either installing it locally or via the OrthoSelect web server [53]. It automatically searches assembled EST sequences against databases of ortholog groups (OG), assigns ESTs to these predefined OGs, translates the sequences into proteins, eliminates redundant sequences assigned to the same OG, creates multiple sequence alignments of identified ortholog sequences and offers the possibility to further process these alignments in a last step. OrthoSelect performes better than the best-hit selection strategy and shows reliable results in re-annotating database member sequences of OrthoMCL-DB and KOG. Most importantly, we showed that our tool produces high quality data sets such as Dunn et al's data set, but with more selected sequences and therefore less missing data in the alignments. Furthermore, the results our tool produces are in absolut agreement with the results of OrthoMCL, but OrthoSelect offers additional funcionality, e.g. handling with EST sequences, computing sequences alignments, and refining them. Our method also showed a significant speedup in comparison to OrthoMCL. Correct orthology assignment is an important prerequisite for the construction of reliable data sets and OrthoSelect is capable of producing them. This makes a OrthoSelect a valuable tool for researchers dealing with large EST libraries focussing on constructing data sets for phylogenetic reconstructions. The tool can be downloaded at http://gobics.de/fabian/orthoselect.php or the web server accessed without local installation at http://orthoselect.gobics.de/.

Availability and requirements

Project name: OrthoSelect

Project home page: http://www.gobics.de/fabian/orthoselect.php

Operating system: Mac OS X, Linux

Programming language: Perl

Other requirements: BioPerl, BLAST, ESTScan, GeneWise, Clustalw, Muscle or T-Coffee, HMMER, Gblocks, Aliscore or Noisy

License: GNU GPL

Restrictions: none

Authors' contributions

FS developed, implemented, tested OrthoSelect, and wrote the manuscript. KP and DE evaluated the performance and usability of OrthoSelect and revised the manuscript. GW conceived and co-supervised the project, provided resources and revised the manuscript. BM co-supervised the project, provided resources and participated in writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Katharina Hoff for critically reading the manuscript, the anonymous reviewers for their valuable comments and Casey Dunn for kindly providing the EST sequences for our evaluation. This work was financially supported by the German Research Foundation (DFG, Project Wo896/6-1,2) within DFG Priority Programme SPP 1174 "Deep Metazoan Phylogeny", and by the German Federal Ministry of Research and Education (BMBF) project "MediGRID" (BMBF 01AK803G).

References

1. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** *Nature Reviews Genetics* 2005, **6(5)**:361-375.
2. Gee H: **Evolution: ending incongruence.** *Nature* 2003, **425**:798-804.
3. Eisen JA: **Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8(3)**:163-167.
4. Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES, Thorndyke M, Nakano H, Kohn AB: **Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida.** *Nature* 2006, **444(7115)**:85-88.
5. Delsuc F, Brinkmann H, Chourrout D, Philippe H: **Tunicates and not cephalochordates are the closest living relatives of vertebrates.** *Nature* 2006, **439(7079)**:965-968.
6. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452(7188)**:745-749.
7. Philippe H, Derelleand R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E, Silva CD, Wincker P, Guyader HL, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M: **Phylogenomics Revives Traditional Views on Deep Animal Relationships.** *Current Biology* 2009, **19(8)**:706-712.
8. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19(2)**:99-113.
9. Sonnhammer E, Koonin E: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genetics* 2002, **18**:619-620.
10. Koonin EV: **ORTHOLOGS, PARALOGS, AND EVOLUTIONARY GENOMICS.** *Annual Review of Genetics* 2005, **39**:309-338.
11. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25(17)**:3389-3402.
12. Zmasek C, Eddy S: **RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs.** *BMC Bioinformatics* 2002, **3**:14.
13. Mushegian AR, Garey JR, Martin J, Liu LX: **Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes.** *Genome Res* 1998, **8(6)**:590-598.
14. Chen F, Mackey AJ, Stoeckert J, Christian J, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucl Acids Res* 2006:D363-368.
15. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, Rao BS, Smirnov S, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
16. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *Journal of Computational Biology* 2000, **7(1–2)**:203-214.
17. Duret L, Mouchiroud D, Gouy M: **HOVERGEN: a database of homologous vertebrate genes.** *Nucl Acids Res* 1994, **22(12)**:2360-2365.
18. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Heriche JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R: **TreeFam: 2008 Update.** *Nucl Acids Res* 2008, **36(S1)**:D735-740.
19. Dolinski K, Botstein D: **Orthology and functional conservation in eukaryotes.** *Annual Review of Genetics* 2007, **41**:465-507.
20. Li L, Stoeckert J, Christian J, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9)**:2178-2189.
21. O'Brien KP, Remm M, Sonnhammer ELL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucl Acids Res* 2005:D476-480.
22. Zhou Y, Landweber LF: **BLASTO: a tool for searching orthologous groups.** *Nucl Acids Res* 2007:W678-682.
23. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl Toolkit: Perl Modules for the Life Sciences.** *Genome Res* 2002, **12(10)**:1611-1618.
24. Gentzsch T: *Sun grid engine: Towards creating a compute power grid* IEEE Computer Society Press; 2001.
25. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models.** *Bioinformatics* 2003, **19(Suppl 2)**:ii103-112.
26. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14(5)**:988-995.
27. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17(4)**:540-552.
28. Dress A, Flamm C, Fritzsch G, Grunewald S, Kruspe M, Prohaska S, Stadler P: **Noisy: Identification of problematic columns in multiple sequence alignments.** *Algorithms for Molecular Biology* 2008, **3**:7.

29. Misof B, Misof K: **A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments: A More Objective Means of Data Exclusion.** *Syst Biol* 2009, **58:**syp006.
30. Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH: **Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits.** *Nucl Acids Res* 2006, **34(11):**3309-3316.
31. Wasmuth J, Blaxter M: **prot4EST: Translating Expressed Sequence Tags from neglected genomes.** *BMC Bioinformatics* 2004, **5:**187.
32. Hatzigeorgiou AG, Fiziev P, Reczko M: **DIANA-EST: a statistical analysis.** *Bioinformatics* 2001, **17(10):**913-919.
33. Fukunishi Y, Hayashizaki Y: **Amino acid translation program for full-length cDNA sequences with frameshift errors.** *Physiol Genomics* 2001, **5(2):**81-7.
34. Tatusova T, Madden T: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiology Letters* 1999, **174(2):**247-250.
35. Wiens J: **Missing data and the design of phylogenetic analyses.** *Journal of Biomedical Informatics* 2006, **39:**34-42.
36. Changhui Yan JGB, Eulenstein O: **Identifying optimal incomplete phylogenetic data sets from sequence databases.** *Molecular Phylogenetics and Evolution* 2005, **35(3):**528-535.
37. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucl Acids Res* 2003, **31(13):**3497-3500.
38. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucl Acids Res* 2004, **32(5):**1792-1797.
39. Edgar R: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
40. Notredame C, Higgins DG, Heringa J: **T-coffee: a novel method for fast and accurate multiple sequence alignment.** *Journal of Molecular Biology* 2000, **302:**205-217.
41. Poirot O, O'Toole E, Notredame C: **Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments.** *Nucleic Acids Res* 2003, **31(13):**3503-3506.
42. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Research* 2005, **15(2):**330-340.
43. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nuc Acids Research* 2002, **30(14):**3059-3066.
44. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nuc Acids Research* 2005, **33(2):**511-518.
45. Schmollinger M, Nieselt K, Kaufmann M, Morgenstern B: **DIALIGN P: fast pair-wise and multiple sequence alignment using parallel processors.** *BMC Bioinformatics* 2004, **5:**128.
46. Morgenstern B, Prohaska SJ, Pöhler D, Stadler PF: **Multiple sequence alignment with user-defined anchor points.** *Algorithms for Molecular Biology* 2006, **1:**6.
47. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B: **DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment.** *BMC Bioinformatics* 2005, **6:**66.
48. Subramanian A, Kaufmann M, Morgenstern B: **DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment.** *Algorithms for Molecular Biology* 2008, **3:**6.
49. Eddy SR: **A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation.** *PLoS Comput Biol* 2008, **4(5):**.
50. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis* Cambridge, UK: Cambridge University Press; 2006.
51. **Department of Energy Joint Genome Institute** [http://genome.cshlp.org/cgi/content/abstract/12/10/1611]
52. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucl Acids Res* 2002, **30(7):**1575-1584.
53. Schreiber F, Wörheide G, Morgenstern B: **OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences.** *Nucl Acids Res* 2009:W185-188.

Chapter 6

Treephyler: Fast taxonomic profiling of metagenomes

Citation

Fabian Schreiber, Peter Gumrich, Rolf Daniel and Peter Meinicke.
Treephyler: fast taxonomic profiling of metagenomes.
Bioinformatics (2010), 26(7):960-961.

Original Contribution

FS developed, implemented, tested treephyler, and wrote the manuscript, except for the introduction, and set up the *Treephyler* webpage.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 7 2010, pages 960–961
doi:10.1093/bioinformatics/btq070

Genome analysis

Advance Access publication February 19, 2010

Treephyler: fast taxonomic profiling of metagenomes

Fabian Schreiber[1,2,3,*], Peter Gumrich[1], Rolf Daniel[4] and Peter Meinicke[1]

[1]Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, [2]Department of Earth- and Environmental Sciences, [3]GeoBioCenter^{LMU}, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 München and [4]Abteilung Genomische und Angewandte Mikrobiologie, Institut für Mikrobiologie und Genetik, Georg-August-Universität, Grisebachstrasse 8, 37077 Göttingen, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Assessment of phylogenetic diversity is a key element to the analysis of microbial communities. Tools are needed to handle next-generation sequencing data and to cope with the computational complexity of large-scale studies. Here, we present *Treephyler*, a tool for fast taxonomic profiling of metagenomes. *Treephyler* was evaluated on real metagenome to assess its performance in comparison to previous approaches for taxonomic profiling. Results indicate that *Treephyler* is in terms of speed and accuracy prepared for next-generation sequencing techniques and large-scale analysis.

Availability: *Treephyler* is implemented in Perl; it is portable to all platforms and applicable to both nucleotide and protein input data. *Treephyler* is freely available for download at http://www.gobics.de/fabian/treephyler.php

Contact: fschrei@gwdg.de

Received on December 18, 2009; revised on January 25, 2010; accepted on February 16, 2010

1 INTRODUCTION

Beyond the analysis of single species genomes of culturable organisms, metagenomics currently opens a new view on the exploration of microbial communities. Progress in sequencing technology enables broader and deeper genomic sampling of the biosphere which in turn puts new challenges for sequence analysis methods. Problems arise from the sheer mass and the short length of sequencing reads. Usually only a small fraction of reads can be assembled due to the phylogenetic diversity in the samples. In the first instance, large-scale analysis of short metagenomic sequencing reads has to provide an estimate of the phylogenetic distribution of the sample. Taxonomic profiling achieves this task by assigning sequencing reads to phylogenetic categories. The most common methods are based on homology to known genes.

The classical ‘gold standard’ approach to taxonomic profiling in metagenomics is focused on the 16S rRNA gene and relies on a sufficient number of sequences of that gene in metagenomic sequence data. Usually the number of reads containing sufficiently long stretches of 16S rRNA is small. Therefore, several researchers perform deep sequencing of that particular gene [see e.g. (Hamady and Knight, 2009)]. Although this approach efficiently overcomes the sparseness of 16S rRNA in metagenomic samples, the sequence data support taxonomic profiling only, without any explicit information about the functional inventory of microbial communities. Furthermore, 16S analysis does not apply to metatranscriptomics, an increasingly important approach to direct measurement of the metabolic activity of microbial communities. Another way to cope with the small proportion of 16 S rRNA in metagenomic data is to extend the set of marker genes to particular protein coding genes (Wu and Eisen, 2008). In von Mering *et al.* (2007), a set of 31 marker genes for metagenome analysis was proposed. This principle has been further extended in Krause *et al.* (2008) where all *PFAM* protein domains are used as potential markers. Their tool *CARMA* searches metagenomic sequences for *PFAM* domains and classifies them on the basis of pyhlogenetic trees built from the metagenome and reference sequences. Although computationally demanding for large-scale metagenome analysis, the *CARMA* approach shows the potential of a dual use of *PFAM* domain assignments which not only provides a basis for taxonomic profiling but also for functional profiling as well. In principle, also BLAST-based analysis [*MEGAN* (Huson *et al.*, 2007), *MG-RAST* (Meyer *et al.*, 2008)] can achieve both kinds of profiling at the same time because the detected homologies may provide information about functional and taxonomic relations. However, the known shortcomings of BLAST-based analysis in metagenomics include the requirement of a sufficient sequence length and the existence of close homologues in the reference database. In contrast to homology-based approaches, several methods pursue the direct classification of the DNA signature of single reads [*PhyloPythia* (McHardy *et al.*, 2007), *TACOA* (Diaz *et al.*, 2009), *Phymm* (Brady and Salzberg, 2009)]. While previous methods showed a rapidly decreasing classification performance for read lengths <1000 bp, more recent approaches also seem to perform reasonably well on short reads. Here, we present a new tool for community profiling in metagenomics and metatranscriptomics which is based on *PFAM* domain assignments. Previous methods like the *CARMA* approach are limited to small-scale analysis due to computational expense of homology search and tree inference. Here, we propose an approach which combines ultra-fast *PFAM* domain prediction as obtained from the *UFO* web server (Meinicke, 2009) with an efficient phylogenetic method based on fast tree inferences using approximate maximum likelihood trees (Price *et al.*, 2009).

2 METHODS

Our algorithm offers fast taxonomic profiling to investigate the community structure of metagenomes. Based on *PFAM* predictions, e.g. by *UFO*, pre-calculated profile Hidden Markov Models of all *PFAM* families are used to

[*] To whom correspondence should be addressed.

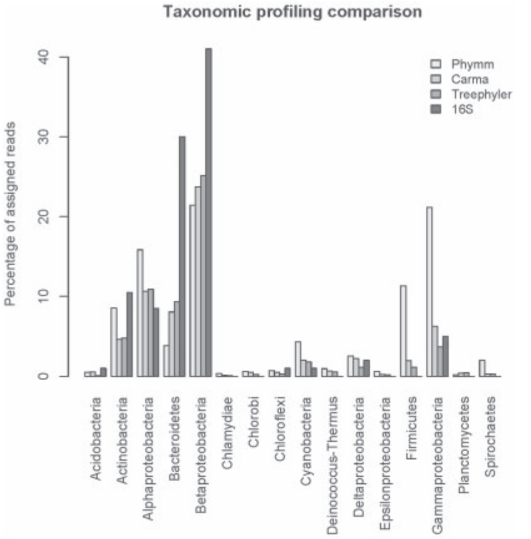


Fig. 1. The relative amount of assigned sequences is shown for each method as well as for each bacterial phylum for the glacial ice metagenome.

screen matching sequencing reads for significant hits. Reads are classified using a phylogenetic tree. For each *PFAM* family with a sufficient number of newly assigned sequences, approximate-maximum likelihood trees of the *PFAM* database sequences and the matching reads are computed using FastTree, which combines the speed of minimum-evolution methods with the accuracy of maximum likelihood methods. Once trees are computed, *Treephyler* uses the algorithm of (Nguyen *et al.*, 2006) to classify reads according to the phylogenetic placement in the tree (see also *Treephyler* web site). *Treephyler* offers an efficient way to balance the computation load on multi-core computers or computer clusters. By this, the runtime only depends on the computation of the largest trees. Similar to *CARMA*, *Treephyler* only computes trees for *PFAM* families with less than 3000 (assigned + reference) sequences.

3 RESULTS

The glacial ice dataset (Simon *et al.*, 2009) was taken as a reference because of its relatively short read length (∼200 bp), the availability of results from a 16S analysis and the moderate sample size (∼0.2 Gbp). We analysed the glacial ice dataset to assess the performance of *Treephyler* in comparison with the tree-based tool *CARMA* and the signature-based tool *Phymm*, and the 16S RNA reference analysis. The analysis was conducted on a single 2.4 GHz dual-core CPU AMD Opteron with 16 Gb RAM.

For runtime comparison, we randomly selected 1% of the glacial ice dataset to allow the comparison with *CARMA*. Both *Treephyler* and *Phymm* analysed the reduced dataset in ∼25 min, while it took *CARMA* 168 h to complete the analysis. On the full dataset, *Treephyler* needed only 12 h, while *Phymm* needed 30 h. The estimated runtime for *CARMA* is 696 h. The runtime of UFO for the reduced and the full dataset was 22 s and ∼30 m, respectively. Results on the full dataset of *Treephyler* and *CARMA* [taken from (Simon *et al.*, 2009)] are in good agreement with the 16 S analysis, expect for the phyla Bacteroidetes (*Phymm*: 3%, *CARMA*: 8%, *Treephyler*: 9%, 16S: 30%) and Betaproteobacteria (*P*: 21%, *C*: 24%, *T*: 24%, 16S: 41%), where all three methods differ from the 16S analysis (see Fig. 1). This may be the consequence of an uneven taxon sampling of *PFAM*. Remarkably, *Phymm* also disagreed on the phyla *Firmicutes* (*P*: 11%, 16S: 0%) and *Gammaproteobacteria* (*P*: 21%, 16S: 5%). Test data and additional results for the class level are available at the *Treephyler* web site.

4 CONCLUSION

We introduced *Treephyler*, a new tool for fast taxonomic profiling of metagenomes. We evaluated our method on real metagenomic data by comparison with previous approaches for taxonomic profiling. We could show a close correspondence between the predicted profiles of *Treephyler* and *CARMA*, while computational speed was increased by orders of magnitude. While speed is not necessarily an essential requirement in genome analysis, the increase of metagenomic sequence data urges for particularly efficient techniques, which also work with limited computational resources. Therefore, the approach we propose here is well prepared for next-generation sequencing technologies and large-scale studies like the exploration of the human microbiome.

Funding: DFG (German Research Foundation) Priority Program SPP1174 ‘Deep Metazoan Phylogeny’ (Project Wo896/6-1,2).

Conflicts of Interest: none declared.

REFERENCES

Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6**, 673–676.

Diaz,N.N. *et al.* (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, **10**, 56.

Hamady,M. and Knight,R. (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.*, **19**, 1141–1152.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Krause,L. *et al.* (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.

McHardy,A.C. *et al.* (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.

Meinicke,P. (2009) UFO: a web server for ultra-fast functional profiling of whole genome protein sequences. *BMC Genomics*, **10**, 409.

Meyer,F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Nguyen,T.X. *et al.* (2006) Phylogenetic analysis of general bacterial porins: a phylogenomic case study. *J. Mol. Microbiol. Biotechnol.*, **11**, 291–301.

Price,M.N. *et al.* (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.

Simon,C. *et al.* (2009) Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl. Environ. Microbiol.*, **75**, 7519–7526.

von Mering,C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.

Wu,M. and Eisen,J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, **9**, 10.

Chapter 7

Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships

Citation

Kerstin S. Pick[1], Hervé Philippe[1], Fabian Schreiber, Dirk Erpenbeck , Daniel J. Jackson, Petra Wrede , Wiens M., Alexandre Alié, Burkhard Morgenstern, Michael Manuel and Gert Wörheide.
Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships.
Molecular biology and evolution (2010), doi:10.1093/molbev/msq089.

Original Contribution

FS used OrthoSelect to expand the dataset of Dunn *et al.* (2008) by adding sequences from early-branching metazoans. FS participated in the data analysis.

[1] These authors contributed equally

MBE Advance Access published April 8, 2010

Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships

Letter

Pick, K.S.[1, §], Philippe, H.[3,§], Schreiber, F.[4], Erpenbeck, D.[1], Jackson, D.J.[2], Wrede P.[5], Wiens M.[5], Alié, A.[6], Morgenstern, B.[4], Manuel, M.[6], Wörheide, G.[1*]

[1] Department of Earth- and Environmental Sciences, Palaeontology and Geobiology & GeoBio-Center^{LMU}, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 München, Germany

[2] Courant Research Center Geobiology, Georg-August Universität Göttingen, Goldschmidtstr. 3, 37077 Göttingen, Germany

[3] Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada.

[4] Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany

[5] Department of Applied Molecular Biology, Institute for Physiological Chemistry and Pathobiochemistry, Duesbergweg 6, 55099 Mainz, Germany

[6] Univ Paris 06, UMR 7138 Systématique, Adaptation, Evolution, CNRS UPMC MNHN IRD, Case 05, Université Pierre et Marie Curie, 7 quai St Bernard, 75005 Paris, France.

[§] Both authors contributed equally to the manuscript.

*corresponding author, Email: woerheide@lmu.de, Tel:+49-89-2180 6718

Fax: +49-89-2180 6601

Keywords: Multigene analysis, EST, Metazoa, Ctenophora, Porifera, Long-branch Attraction, saturation

Running head: Metazoan Phylogenomics

Abstract length – 152 words

Total length of text – 10,075 Characters

Total page requirement – 3.5 pages

Number of references – 25

Abstract

Despite expanding data sets and advances in phylogenomic methods, deep level metazoan relationships remain highly controversial. Recent phylogenomic analyses depart from classical concepts in recovering ctenophores as the earliest-branching metazoan taxon, and propose a sister-group relationship between sponges and cnidarians (e.g. Dunn et al., 2008, Nature 452: 745). Here, we argue that these results are artefacts stemming from insufficient taxon sampling and long-branch attraction (LBA). By increasing taxon sampling from previously unsampled non-bilaterians and using an identical gene set to that reported by Dunn et al. (2008) we recover monophyletic Porifera as the sister group to all other Metazoa. This suggests that the basal position of the fast-evolving Ctenophora proposed by Dunn et al. was due to LBA and that broad taxon sampling is of fundamental importance to metazoan phylogenomic analyses. Additionally, saturation in the Dunn et al. character set is comparatively high, possibly contributing to the poor support for some non-bilaterian nodes.

Resolving the relationships of deep branching metazoan lineages is critical if we are to understand early animal evolution. Unravelling these relationships through the analysis of large scale molecular data sets has recently given birth to the field of phylogenomics (e.g. Philippe et al. 2005). Despite significant advances in this field, recent studies have generated contradictory results regarding relationships within and between early-diverging metazoan lineages: cnidarians, ctenophores (comb-jellies), sponges, placozoans (anatomically the simplest extant metazoans), and bilaterians. Placozoans have historically been regarded by some as relicts of the metazoan ancestor (see summary by Schierwater 2005), and some recent analyses place Placozoa at the base of a group of non-bilaterian animals (Dellaporta et al. 2006; Schierwater et al. 2009). However, recent whole genome (Srivastava et al. 2008) and phylogenomic (Philippe et al. 2009) analyses including *Trichoplax* recovered sponges as the sister-group to all other metazoans in accordance with morphological analyses (Ax 1996). Such contradictory hypotheses regarding non-bilaterian metazoan relationships prevent a consensus view of metazoan evolution, a goal that is of fundamental importance if we hope to fully understand the early evolution of animals (for an overview see Erpenbeck and Wörheide 2007).

A recent phylogenomic analysis adds further controversy to this debate (Dunn et al. 2008) (compare also Hejnol et al. 2009). Their outcome is highly unusual as sponges form a clade with the Cnidaria, while the ctenophores (despite being morphologically derived), are proposed to be the earliest branching metazoan taxon. As suggested by Philippe et al. (2009), we hypothesized that a long-branch attraction (LBA) artefact was responsible for these controversial findings due to insufficient ingroup sampling and an inappropriate choice of outgroup taxa. Furthermore, the Placozoa are conspicuously absent from the Dunn et al. (2008) data set, and sponges are represented by only one Demospongiae and one Homoscleromorpha with no representatives of the remaining two extant sponge classes: Calcarea (Calcispongiae or calcareous sponges) and Hexactinellida (glass sponges). Sparse taxon sampling is a common pitfall of phylogenetic analyses (Lecointre et al. 1993) and is largely responsible for the lack of a robustly supported non-bilaterian metazoan phylogeny (Erpenbeck and Wörheide 2007). With a largely different gene set (only 45 genes in common with the 150 gene set of Dunn et al. 2008) and an increased sampling of non-bilaterian species, Philippe et al. (2009) obtained monophyletic sponges as the first-diverging metazoan lineage, and a sister-group relationship between the Cnidaria and the Ctenophora.

To test whether insufficient sampling of non-bilaterian taxa and inappropriate outgroup choice adversely influenced the analyses performed by Dunn et al. (2008), we re-analysed their 64-taxon matrix cleared of instable taxa (leaf stability < 90%) and with the following major modifications (cf. Baurain, Brinkmann, and Philippe 2007):

1) Ingroup taxon sampling was increased by the addition of non-bilaterian EST and genomic sequences. These included: 12 additional sponge taxa representing all four major sponge lineages; one additional ctenophore; five additional cnidarians (see Supplementary Table 1) and *Trichoplax adhaerens* (Placozoa).

2) We removed outgroup taxa with long branches. Long branches in the outgroup can strongly influence the topology of early branching ingroups (Philippe and Laurent 1998; Rota-Stabelli and Telford 2008). The long branches of the fungal outgroup are not visible in the cladogram of the PhyloBayes analysis (CAT+Γ4) of Dunn et al. (see their Fig. 2), but are evident in their Supplementary Figure 1. Consequently, we analysed our dataset with two sets of outgroups. First using only choanoflagellates, the most likely sister-group to all Metazoa (Carr et al. 2008), consisting of *Monosiga ovata* (shortest branch of Dunn et al.'s outgroup taxa), *Monosiga brevicollis* (complete genome data) and *Proterospongia* sp.. Second, with more distant outgroups, such as those used by Dunn et al. (2008) (see Supplementary Figure 1, and Supplementary Data for a detailed taxon list and methods used).

Furthermore we eliminated errors (e.g., frameshifts) and refined the Dunn et al. (2008) alignment, e.g., by reducing missing data and removing 2,150 ambiguously aligned positions (see Supplementary Data for detailed procedures). Our extended data set with the choanoflagellate-only outgroup consists of 80 taxa and 19,002 characters. Using this dataset we performed Bayesian phylogenetic analyses under the CAT+Γ4 model (Lartillot and Philippe 2004) and subsequent non-parametric bootstrapping (cf. Philippe et al. 2009).

Contrary to Dunn et al. (2008), and also Hejnol et al. (2009), we recover sponges as the sister-group to all other metazoan taxa (Figure 1). This is in congruence with earlier morphological (Ax 1996), and phylogenomic analyses

(Philippe et al. 2009). In accordance with the latter study, we also recover sponges as a monophyletic group. The Homoscleromorpha, a taxon previously assigned to the Demospongiae (see Hooper and Van Soest 2002), are found to be the sister-group to Calcarea as suggested by van Soest (1984) and Grothe (1989) based on morphology, and subsequently by Dohrmann et al. (2008) based on rRNA data. Similarly, Hexactinellida and the remaining Demospongiae *sensu stricto* form a monophyletic group (Silicea *sensu stricto*).

The basal position of ctenophores proposed by Dunn et al. (2008) was probably caused by the attraction of ctenophores to distant outgroup species, particularly fungi. In comparison, our re-analysis of the updated Dunn et al. (2008) dataset with increased ingroup taxon-sampling and a refined alignment indicates that LBA is reduced, independent of whether we use the choanoflagellate-only outgroup or more distant outgroups (see Fig. 1 and Supplementary Figure 1). This indicates that ingoup-taxon sampling, and probably to a lesser extent data refinement, are the most important parameters affecting non-bilaterian relationships.

Results of our analyses indicate that sponges are the sister group to the remaining Metazoa, and Placozoa are sister to the Bilateria. We also recover both monophyletic Ctenophores and Cnidaria, but they are paraphyletic with respect to Placozoa+Bilateria (Fig. 1). This is in contrast to the findings of Philippe et al. (2009) that supported the "Coelenterata hypothesis" (c. f. Haeckel 1866), i.e., a monophyletic Cnidaria+Ctenophora clade and a sister-group relationship between Coelenterata and Bilateria. However, support values for the position of Ctenophora, Cnidaria and Placozoa in our analysis are either not significant (posterior probabilities < 0.9) or low (bootstrap support < 70%). We suspected that Dunn et al.'s character set contains a substantial amount of non-phylogenetic signal due to multiple substitutions. To test this, we conducted a saturation analysis of inferred substitutions against observed amino acid differences (Figure 2). This revealed a higher saturation in the original Dunn et al. (2008) character set (slope = 0.38x) compared to the Philippe et al. (2009) character set (slope = 0.46x) (Figure 2). From this we conclude that despite increasing the number of non-bilaterian taxa by a factor of 3 (from 9 to 27), multiple substitutions have partly masked phylogenetic signal contributing to the incongruent results reported here with those of Philippe et al. (2009). However, with the expanded and refined dataset reported here none of these incongruencies are statistically significant, indicating that non-phylogenetic signal has been reduced with respect to the original character set of Dunn et al. (2008). Furthermore, Dunn et al. (2008) recovered high support for the sister-group relationship of ctenophores to the remaining Metazoa – based on our analyses here this hypothesis should be rejected (with a bootstrap value of 91%).

The inclusion of additional taxa has little influence on the relationships within and between bilaterian crown groups. Three of the four differences between the findings of Dunn et al. (2008) and our results affect the relationships of a single sequence within their well-defined clades (*Euprymna* within Mollusca, *Paraplanocera* within Platyhelminthes and *Anoplodactylus* among the chelicerate arthropods). None of these splits were strongly supported in the original Dunn et al. (2008) analysis. Additionally, we do not recover

Panarthropoda due to a difference in the position of Tardigrada. Panarthropoda was also weakly supported in the Dunn et al. (2008) analysis (PP values under WAG and CAT models were 0 and 0.86 respectively, and RAxML bootstrap support under the WAG model with 64 and 77 taxa was 4% and 2% respectively).

Our results highlight the sensitivity of phylogenomic studies to ingroup taxon sampling, and demonstrate the need for great care in the analysis and interpretation of large data sets. Character-rich analyses are thought to outperform character-poor analyses, and have been suggested to be of greater importance than increased taxon sampling with regard to recovering robust metazoan phylogenies (Rokas and Carroll 2005). However, our analyses demonstrate the strong influence of taxon sampling, even though non-bilaterian taxa still remain under-represented (Cnidaria: no Octocorallia, Ceriantharia, Cubozoa or Staurozoa; Ctenophora: no Platyctenida, Beroida, Cestida; just one placozoan strain etc.). The phylogenomic approach promises to reveal a well resolved consensus metazoan tree, but it should not be assumed that a large dataset will automatically produce a strong or correct phylogenetic signal (Jeffroy et al. 2006). A wide range of factors, such as saturation, LBA, the best fitting evolutionary model and appropriate outgroup choice (Philippe et al. 2005) need to be carefully addressed before a fully resolved and robust animal tree of life will be realized.

Acknowledgements: This study was supported by the German Research Foundation (DFG) through the Priority program SPP 1174 "Deep Metazoan Phylogeny" (Projects Wo896/6-1, 2; Wi 2116/2-1, 2). M. Kube and his team at the Max-Planck Institute for Molecular Genetics (Berlin, Germany) are acknowledged for library construction and EST sequencing, as well as I. Ebersberger and his team at the Center for Integrative Bioinformatics (Vienna, Austria) for their bioinformatic processing. We thank C. Eckert for some tissue samples. HP gratefully acknowledges financial support by NSERC, the Canadian Research Chair Program and the Université de Montréal, and the Réseau Québecois de Calcul de Haute Performance for computational resources. M.M. acknowledges the French Ministry of Research (“ACI jeunes chercheurs” and ANR NT_NV_52 Genocnidaire), the Consortium National de Recherche en Génomique, the Genoscope and the GIS Institut de la Génomique Marine for financial support. DJJ is supported by DFG funding to the Courant Research Centre for Geobiology, Göttingen.

Literature Cited:

Ax P. 1996. Das System der Metazoa. ein Lehrbuch der Phylogenetischen Systematik. Stuttgart: Gustav Fischer Verlag.

Baurain D, Brinkmann H, Philippe H. 2007. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? Mol Biol Evol 24:6-9.

Carr M, Leadbeater BS, Hassan R, Nelson M, Baldauf SL. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. Proc Natl Acad Sci USA 105:16641-16646.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol 17:540-552.

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci USA 103:8751-8756.

Dohrmann M, Janussen D, Reitner J, Collins A, Wörheide G. 2008. Phylogeny and evolution of glass sponges (Porifera: Hexactinellida). Syst Biol 57:388-405.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.

Erpenbeck D, Wörheide G. 2007. On the molecular phylogeny of sponges (Porifera). In: Zhang Z-Q, Shear WA, editors. Linnaeus Tercentenary: Progress in Invertebrate Taxonomy. Zootaxa 1668. Magnolia Press. p 107-126.

Grothe F. 1989. On the phylogeny of homoscleromorphs. Berl Geowiss Abh A 106:155-164.

Haeckel EH. 1866. Generelle Morphologie der Organismen. Berlin: G. Reimer.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc R Soc Lond B Biol Sci 276:4261-4270.

Hooper JNA, Van Soest RWM. 2002. Systema Porifera. Guide to the Supraspecific Classification of Sponges and Spongiomorphs (Porifera). New York: Plenum.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225-231.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286.

Lecointre G, Philippe H, Van Le HL, Le Guyader H. 1993. Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol 2:205-224.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Deniel E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M. 2009. Phylogenomics restores traditional views on deep animal relationships. Curr Biol 19:706-712.

Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-623.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Ann Rev Ecol Syst 36:541-562.

Rokas A, Carroll SB. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Mol Biol Evol 22:1337-1344.

Rota-Stabelli O, Telford MJ. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol 48:103-111.

Schierwater B, Eitel M, Jakob W, Osigus H, Hadrys H, Dellaporta S, Kolokotronis S, Desalle R, Penny D. 2009. Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern "Urmetazoon" Hypothesis. PLoS Biology 7:e20.

Schierwater B. 2005. My favorite animal, *Trichoplax adhaerens*. BioEssays 27:1294-1302.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS. 2008. The *Trichoplax* genome and the nature of placozoans. Nature 454:955.

van Soest RWM. 1984. Deficient *Merlia normani* from the Curaçao reefs, with a discussion on the phylogenetic interpretation of sclerosponges. Bijdragen tot de Dierkunde 54:211-219.

Figures:

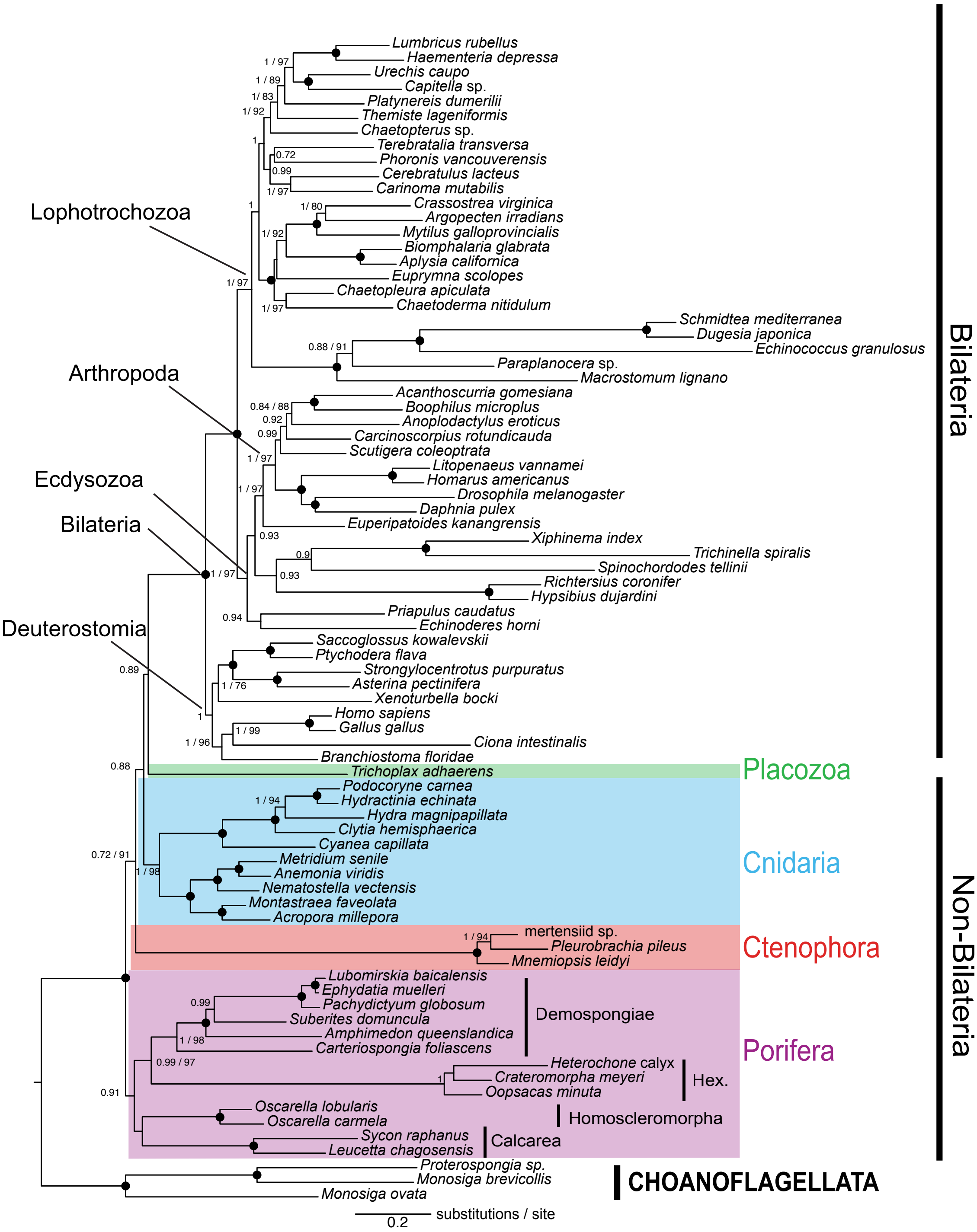


Figure 1: Phylogenetic tree based on refinements to the Dunn et al. (2008) 64-taxon set reconstructed with PhyloBayes (Lartillot, Lepage, and Blanquart 2009) under the CAT+Γ4 model. Choanoflagellates were set as outgroup and an additional 18 non-bilaterian taxa included. Posterior probabilities > 0.7 are indicated followed by bootstrap support values > 70. A large black dot indicates maximum support in posterior probabilities and bayesian bootstraps (= 1 / 100).

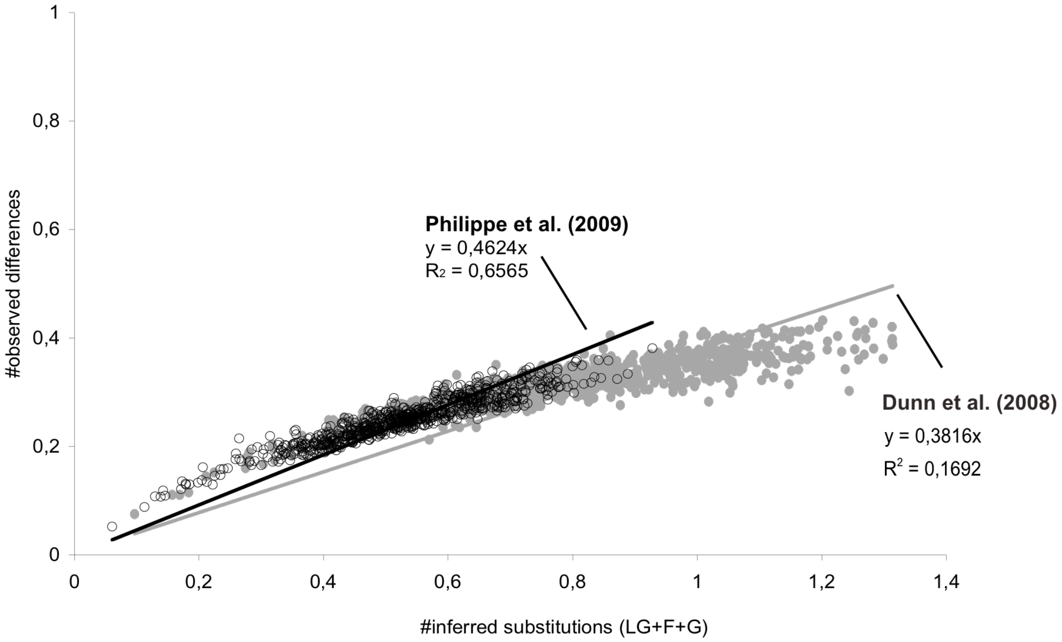


Figure 2: Saturation plot of character sets. See supplementary materials for method details. Grey line and filled dots: Dunn et al. (2008). Black line and open dots: Philippe et al. (2009).

Chapter 8

Results and Discussion: Automated orthology search, monophyletic sponges and scalable taxonomic profiling

The aim of this study was to solve long standing issues of early animal evolution. For this purpose, we needed a new tool to automatically build phylogenomic datasets from our newly sequenced ESTs as well as publicly available sequences. Furthermore, we wanted to develop a new method suitable for large-scale analysis of metagenomes.
The results and discussion section is divided into the following parts:

- ◇ Tool for orthology search (see section 8.1),
- ◇ Solving long-standing issues regarding basal animal evolution using phylogenomics (see section 8.2),
- ◇ Taxonomic profiling in metagenomics (see section 8.3).

8.1 What we learned about orthology search

EST sequences are routinely used in phylogenetic studies to answer evolutionary questions. A first crucial step in these studies is the assembly of a dataset of orthologous genes from large sequence databases. These orthologous genes will be processed to keep appropriate target genes and eliminate redundant sequences to obtain a set of sequences suitable for the subsequent tree reconstruction. Since the manual generation of such datasets is very time-consuming, our goal was to develop a software tool that carries out all steps automatically. Furthermore, a web server should be developed to allow an easy access to the software without installation.

Summary of results

We have developed the easy-to-use software pipeline *OrthoSelect* (Schreiber *et al.*, 2009a), that constructs orthologous gene alignments for phylogenomic analyses. Phylogenomic studies are usually based on large sets of assembled ESTs and sequences from public databases. *OrthoSelect* searches those large sets for orthologous sequences, assigns the translated sequences to existing orthologous groups, keeps the orthologous copy of the gene - in case there are multiple copies of a gene for a taxon -, uses popular alignment tools to build alignments, and offers post-processing (the removal of ambiguously aligned alignment colums, the removal of homoplastic sites, and the removal of phylogenetic misleading information) afterwards. Our software pipeline can be extended and runs on desktop machines and computer clusters. Furthermore, we set up a web server to provide an easy-to-use interface to *OrthoSelect* (Schreiber *et al.*, 2009b).

Main point

OrthoSelect is the first tool that builds phylogenomic datasets in a fully automated way and can also deal with EST sequences. In contrast to existing methods, *OrthoSelect* not only assigns new sequences to (existing) orthologous groups, but also selects the sequences most likely to be orthologous from each taxon. The web server provides the results of the analysis, but also shows numerous statistics to give researchers additional useful information.

Evaluation

We successfully evaluated the performance of our tool by comparing the results of *OrthoSelect* to the following two other approaches that predicts orthologs as well as a published reference dataset:

- ◇ Best-hit selection strategy (Mushegian *et al.*, 1998),
- ◇ Predictions by the KOG algorithm (Tatusov *et al.*, 2003),
- ◇ The reference data set by Dunn *et al.* (2008).

The tests showed that our tool was superior in terms of speed and accuracy to the best-hit selection strategy and the KOG algorithm. Furthermore, *OrthoSelect* was able to automatically construct the manual curated reference dataset. *OrthoSelect* is capable of producing datasets for phylogenomic studies.

Concluding remark

In phylogenomics, there was previously no tool available capable of searching EST databases and building orthologous gene alignments ready for phylogenetic tree reconstruction software. By developing *OrthoSelect*, we have filled this gap. *OrthoSelect* is a valuable tool for researchers dealing with large libraries of EST sequences focussing on the assembly of large datasets for phylogenetic analyses. That makes it applicable to almost any phylogenetic study dealing with the *Tree of Life*.

Phylogenomic methods: Summary of our aims

- ◇ First fully automated tool for dataset construction in EST-based phylogenomics → $\surd$
- ◇ Evaluating performance of tool using real data → $\surd$
- ◇ Easy-to-use web interface → $\surd$

Open question

Once, *OrthoSelect* has assigned sequences to predefined orthologous groups, it selects that sequence from each taxon that is most likely an ortholog. However, in cases where the ortholog for a taxon is missing - due to gene loss or absence from the EST database - *OrthoSelect* will select a sequence in any case. This might lead to the inclusion of paralogs. A possible approach to deal with this is outlined in section 9.

8.2 What we learned about early animal evolution

Previous studies dealing with the relationships between basal metazoan taxa produced contradictory trees with low support. Major issues such as the phylogenetic status of sponges (monophyletic or paraphyletic) and the position of Ctenophora and Placozoa were unresolved. These inconsistencies may be due to insufficient taxon sampling of non-bilaterian phyla. We used a phylogenomic approach and increased taxon sampling of non-bilaterian phyla to try to answer these issues.

We extended two existing datasets by newly generated as well as publicly available sequences, we have assembled two different datasets with high taxon sampling for non-bilaterian taxa.

Position of Placozoa The position of Placozoa in the basal metazoan tree could not be resolved in Philippe *et al.* (2009) (basal to metazoans, but with virtually no support: Bootstrap support (BS) $\approx$ 60%). In Pick *et al.* (2010), we recovered Placozoa as a sister group to all Bilateria with moderate support (posterior probability (pp) = 0.89). The recent study by Schierwater *et al.* (2009) sees Placazoa as the first branching metazoan taxon (BS = 100), whereas Placozoa was placed between paraphyletic sponges in Hejnol *et al.* (2009) with BS = 89. With these contradictory results, the question of the position of Placozoa remains unresolved.

Position of Ctenophora Our results could not confirm a previously proposed hypothesis that Ctenophora are the earliest branching metazoan taxon (Dunn *et al.*, 2008; Hejnol *et al.*, 2009) (BS $\geq$ 90% and BS = 100, respectively). This hypothesis supports polyphyletic Eumetazoa and the independent innovation of eumetazoan synamoporphies such as nerve and muscle cells as well as a differentiated digestive system (Miller and Ball, 2008).

In Philippe *et al.* (2009), Ctenophora formed a sister group with Cnidaria (Coelenterata, BS $\approx$ 93%), the coelenterate clade. This is in congruence with Schierwater *et al.* (2009), although this clade has virtually no support (BS = 27%). The Coelenterate clade is based on anatomical similarities between cnidarian medusae and ctenophores (e.g. gelatinous body, tentacles, and "radial" symmetry), which was later considered as convergent evolution (Harbison, 1985). Although, ctenophores and cnidarians share some embryological features, their body plan notably differs. The phylogenetic position of ctenophores is uncertain as its long branch is prone to the long branch attraction artefact. This fact is supported by the increased support for a coelenterate clade after the removal of the fungi outgroup in Philippe *et al.* (2009). A similar result is in Pick *et al.* (2010), where Ctenophora are basal to a clade Cnidaria + Placozoa + Bilateria (pp = 0.91). Therefore, we attribute the basal position of Ctenophora in Dunn *et al.* (2008) and Hejnol *et al.* (2009) to the attraction of ctenophores to distantly related outgroup taxa.

Phylogenetic origin of sponges

Our analyses support the hypothesis of a monophyletic origin of sponges: In Philippe *et al.* (2009) with BS $\approx$ 95% and in Pick *et al.* (2010) with pp = 0.91. However, support values are not significant for sponge monophyly in (Pick *et al.*,

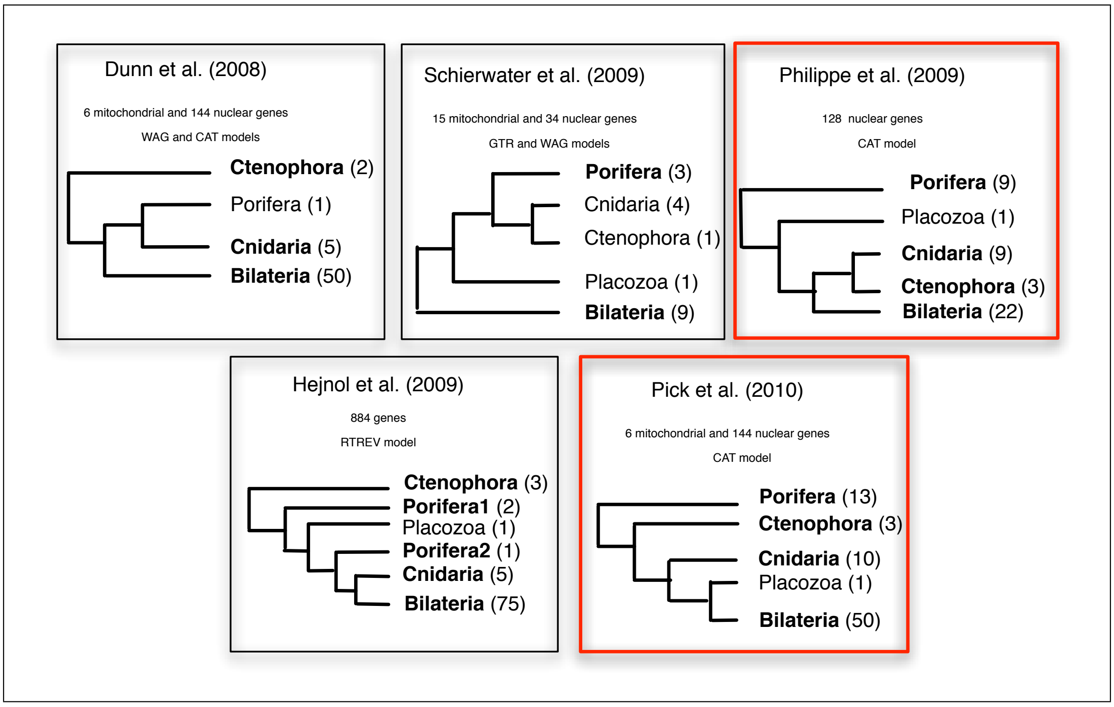


Figure 8.1: The picture shows the results from most recent studies about the animal *Tree of Life* in chronological order. The red boxes are results from our own studies Philippe *et al.* (2009) and Pick *et al.* (2010), and other studies are from Dunn *et al.* (2008), Schierwater *et al.* (2009), and Hejnol *et al.* (2009).

2010) as well as in Philippe *et al.* (2009) when using (a too distantly-related) *Fungi* outgroup. The monophyly of sponges is also present in Schierwater *et al.* (2009), but not supported (BS = 53%). In Hejnol *et al.* (2009), sponges are paraphyletic, but with virtually no support (BS = 31%). A possible scenario of sponge paraphyly that would imply that Eumetazoa are derived from sponge-like ancestors (Borchiellini *et al.*, 2001; Nielsen, 2008; Peterson, 2001) can be ruled out. The hypothesis of sponge monophyly is strongly supported by morphological characters shared by all sponge lineages (e.g. an aquiferous system with choanocyte chambers and the pinacoderm). There is still some uncertainty about whether the choanocytes of sponges are an ancient feature shared with choanoflagellates or they are the product of convergence. In any case, sponges do not reflect a metazoan ancestor, but should be seen as a specialized taxon.

Within sponge classes The relationships of the four sponge classes were identical in both of our studies: Demospongiae formed a clade with Hexactinellida (BS = 100% in Philippe *et al.* (2009) and pp = 0.97 in Pick *et al.* (2010)) in a clade called Silicea (Gray, 1867) sensu stricto. This grouping is supported by the presence of siliceous spicules organized around a well-defined proteic axial filament (Uriz *et al.*, 2003) as well as demospongic acids, a particular class of membrane phospholipids (Thiel *et al.*, 2002). In Schierwater *et al.* (2009), Calcarea has a sister group relationship (BS = 100%) to the clade Hexactinellida + Demospongiae (BS = 98%).

Homoscleromorpha formed a clade with Calcarea (BS = 90% in Philippe *et al.* (2009) and pp = 1.0 in Pick *et al.* (2010)). Although these findings are in conflict with traditional views (Hooper and van Soest, 2002), they are supported by recent phylogenetic studies based on the 18S rRNA (Borchiellini *et al.*, 2004; Dohrmann *et al.*, 2008). The silicious spicules that can be found in Homoscleromorpha might have been independently evolved from those present in hexatinellids and demosponges. No conclusion can be drawn regarding the evolution of the basi-epithelial basement membrane found in larvae and adult homoscleromorphs. This membrane either was present in common metazon ancestor and lost in placozoans and most sponges or independently acquired in homoscleromorphs and eumetazoans.

Concluding remark

Despite the use of large, newly-generated sequences, the relationships between basal metazoa are still controversial. However, we found support for the monophyly of sponges as well as for the branching pattern of sponge lingeages.

We could partially answer our leading questions from section 1.3.2:

◇ Relationships between basal *Metazoa*? → still controversal (why? see chapter → 9)

◇ What is the position of *Placozoa* in the basal metazoan tree? → still controversal

◇ Are *sponges* monophyletic or not? → monophyly: √

◇ Relationship within *sponge* classes? → (Hexactinellida + Demospongiae),(Calcarea + Homoscleromorpha)?

8.3 Metagenomics - a quick overview of the taxonomic composition of metagenomes

We developed *Treephyler* (Schreiber *et al.*, 2010), a tool for fast taxonomic profiling of metagenomes. We evaluated *Treephyler* using real metagenomic data by comparison with existing methods for taxonomic profiling.

We could show that the predicted profiles by *Treephyler* are in close correspondence with those of *CARMA* (Krause *et al.*, 2008), while computational speed was increased by orders of magnitude.

While speed is not an essential requirement in genome analyses, efficient algorithms that work with limited computational resources are needed to cope with the increase of metagenomic sequence data. Therefore, *Treephyler* is well prepared for next generation sequencing technologies and large-scale studies like the exploration of the human microbiome (Turnbaugh *et al.*, 2007).

Chapter 9

Outlook: How to avoid the bottleneck in phylogenomics

> **"The current molecular phylogenetic paradigm still reconstructs gene trees to represent the species tree."**
>
> Liu and Pearl (2007)

The situation with phylogenomic methods

The current situation in phylogenomics is that the use of different datasets can lead to different, contradicting results. An example for this was our study of long-held issues regarding the evolution of basal metazoan taxa. Our studies Philippe *et al.* (2009) and Pick *et al.* (2010) as well as the studies from Dunn *et al.* (2008), Schierwater *et al.* (2009), and Hejnol *et al.* (2009) all resulted in different hypotheses about the branching pattern of Ctenophora, Cnidaria, Placozoa, and Porifera. These findings disprove the hypothesis of Miyamoto and Fitch (1995) that the more data we use, the closer we get to the true tree.

The reason for incongruencies

From a theoretical point of view, our partially contradicting results from chapter 8.2 are not a surprise. It has been reported that evolutionary trees from different datasets/genes can have conflicting topologies (Tajima, 1983; Hudson, 1983; Neigel and Avise, 1986; Pamilo and Nei, 1988; Nichols, 2001; Pollard *et al.*, 2006). In general, there is incongruence between the evolution of genes and the evolution of species.

Reasons for incongruencies can be artefactual, when we fail to recover the correct tree due to insufficient sequence length (stochastic error) or violation of model assumptions (systematic error) (Jeffroy *et al.*, 2006) or a biological reason (Galtier and Daubin, 2008). The biological reasons for such incongruencies are incomplete lineage sorting[1], hidden paralogy, and horizontal gene transfer. All these processes lead to incongruencies between gene and species tree.

Hemiplasy

> **"The topological discordance between a gene tree and a species tree attributable to lineage sorting of genetic polymorphisms that were retained across successive nodes in a species tree."**
>
> Avise and Robinson (2008)

In case of a large effect of incomplete lineage sorting and/or horizontal gene transfer, many morphological or ecological defined species should be para-/ or polyphyletic, a scenario that is rarely observed. The phylogeny of early-branching metazoa is a good case for ancient incomplete lineage sorting, as short internal branches[2] are common (Whitfield and Lockhart, 2007).

A possible solution: Quality not quantity

All of the above mentioned processes lead to the same result: The gene tree does not match the species tree. This is a chicken and egg issue. We need to know the species tree to decide whether a gene tree matches a species tree or not. But, the reconstruction of the species tree is our goal.

The relaxed species tree

One way to overcome this problem is to use a relaxed species tree when comparing gene trees with species trees. A relaxed species tree is a consensus of our current knowledge about the branching order of a certain group of taxa. For parts of the tree, where we have conflicting hypotheses, the tree does not contain any information and is polytomous. The relaxed species trees will not include any hypotheses that the current study is going to test. Given the hypothesis that a

[1] also termed: ancestral polymorphism, deep coalescence, or incomplete coalescence

[2] these branches are usually short, because of the short time frame of high diversification during the cambrian explosion (Giribet, 2009)

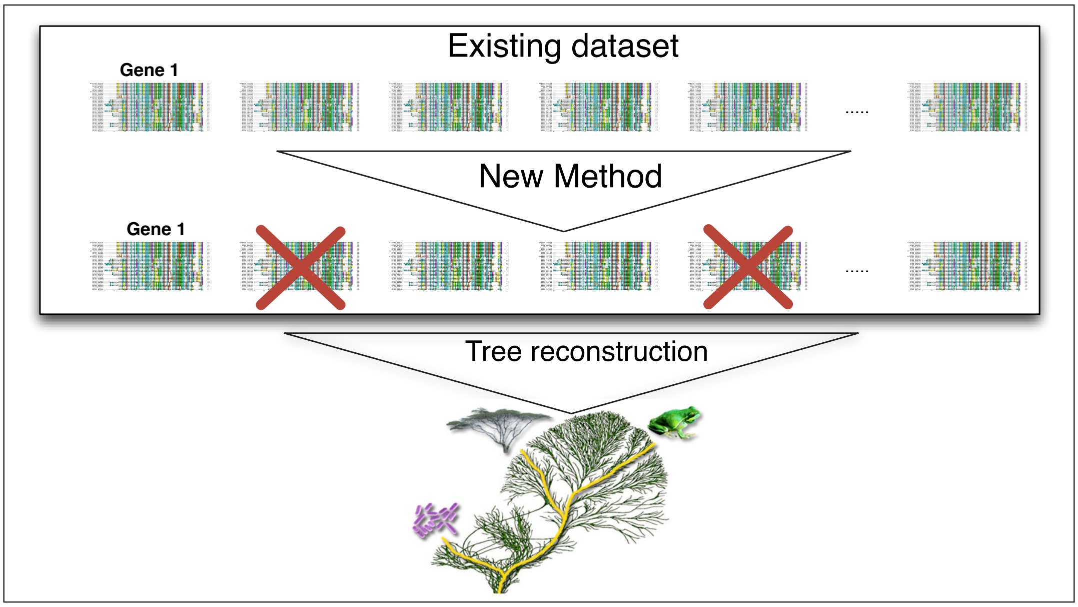


Figure 9.1: The picture shows the idea behind the use of a relaxed species tree. Each gene from an existing dataset is compared to the known tree and either discarded or kept depending on some similarity threshold.

species tree should be constructed using congruent genes, a relaxed species tree can be used as a reference.

Disregarding bad genes

In a first instance, one could compare each gene tree with the relaxed species tree and disregard those genes that are below a certain similarity threshold, e.g. that are not similar enough to the relaxed species tree.

Choosing the best alignment columns

This method can then be extended and applied to single alignment columns instead of whole genes. This is possible because most statistical models for reconstructing phylogenies assume independence of the alignment columns. The evolution of each alignment column can be compared to our current knowledge of the evolution of species. This is done by constructing a phylogenetic tree for each column and compare it to the relaxed species tree. If the alignment column tree is not similar enough to the relaxed species tree, it will be discarded. This leaves only those alignment columns that are congruent with our current knowledge of species evolution.

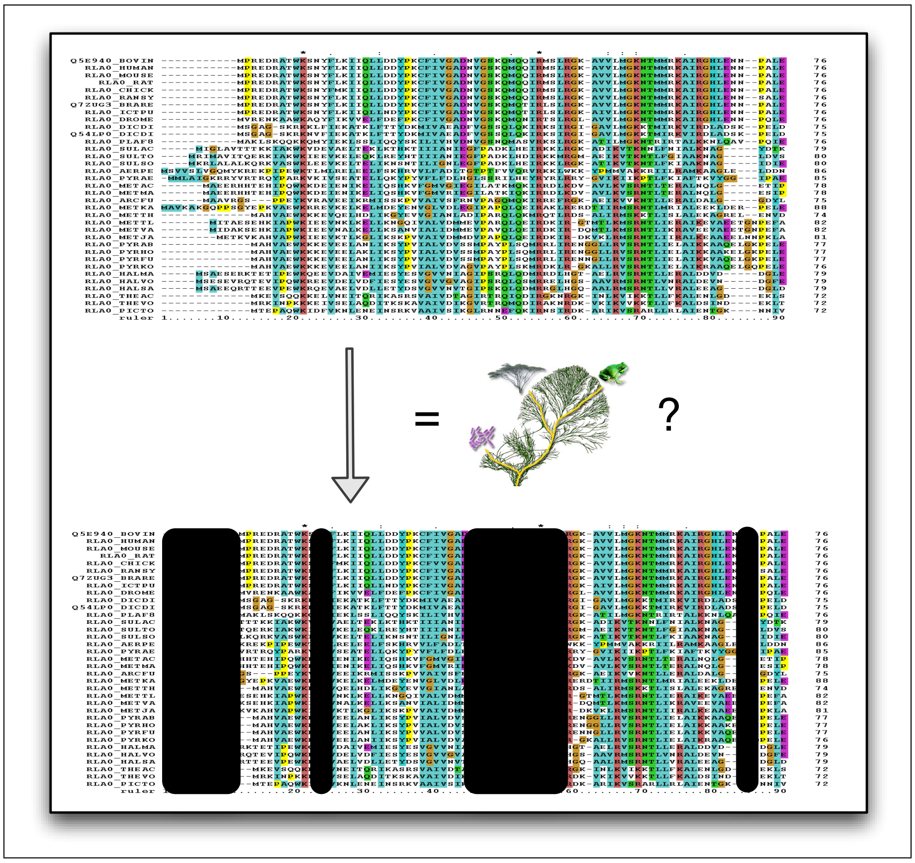


Figure 9.2: The picture shows the alignment columns could be either discarded or accepted for further analysis based on their similarity to a relaxed species tree.

Future of orthology search - back to orthology definition

While these two approaches are only applicable to already existing datasets, the idea can also be applied to build phylogenomic datasets by incorporating the approaches into *OrthoSelect*. Based on orthologous gene alignments from orthologous databases (e.g. *OMA* (Schneider *et al.*, 2007) or *OrthoMCL* (Chen *et al.*, 2006)) new sequences will be added one at a time, a phylogenetic tree build, and the resulting tree compared to some relaxed species tree. By this, orthologous sequences can be clearly distinguished from paralogous. This way of predicting orthology resembles more closely the original definition of orthology (Fitch, 1970) than all existing methods.

Bibliography

Alexeyenko, A., Lindberg, J., and Pérez-Bercoff, Ĺ. (2006). Overview and comparison of ortholog databases. *Drug Discovery Today: Technologies*, **3**(2).

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Computational Biology*, **5**(1), e1000262.

Altschul, S., Madden, T., Schaffer, A., and Zhang, J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Avise, J. C. and Robinson, T. J. (2008). Hemiplasy: A New Term in the Lexicon of Phylogenetics. *Systematic Biology*, **57**(3), 503–507.

Ax, P. (1996). Das System der Metazoa: Ein Lehrbuch der phylogenetischen Systematik. *Gustav Fischer Verlag*.

Baldauf, S., Roger, A., Wenk-Siefert, I., and Doolittle, W. (2000). A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. *Science*, **290**(5493), 972.

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Durufle, 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. *PNAS*, **99**(3), 1414–1419.

Baurain, D., Brinkmann, H., and Philippe, H. (2007). Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Molecular Biology and Evolution*, **24**(1), 6–9.

Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N., and DeLong, E. F. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**(5486), 1902–6.

Blair, J., Ikeo, K., Gojobori, T., and Hedges, S. B. (2002). The evolutionary position of nematodes. *BMC Evolutionary Biology*, **2**(1), 7.

Borchiellini, C., Manuel, M., Alivon, E., Boury-Esnault, N., Vacelet, J., and Parco, Y. L. (2001). Sponge paraphyly and the origin of Metazoa. *Journal of Evolutionary Biology*, **14**(1), 171–179.

Borchiellini, C., Chombard, C., Manuel, M., Alivon, E., Vacelet, J., and Boury-Esnault, N. (2004). Molecular phylogeny of Demospongiae: implications for classification and scenarios of character evolution. *Molecular Phylogenetics and Evolution*, **32**(3), 823–837.

Bourlat, S., Juliusdottir, T., Lowe, C., and Freeman, R. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature*, **444**, 85–88.

Boury-Esnault, N., Ereskovsky, A., Bezac, C., and Tokina, D. (2003). Larval development in the Homoscleromorpha (Porifera, Demospongiae). *Invertebrate Biology*, **122**(3), 187–202.

Brady, A. and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, **6**(9), 673.

Cavalier-smith, T. and Chao, E. (2003). Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution. *Journal of Molecular Evolution*, **56**(5), 540–563.

Cavalier-smith, T., Allsopp, M. T. E. P., Chao, E. E., Boury-Esnault, N., and Vacelet, J. (1996). Sponge phylogeny, animal monophyly, and the origin of the nervous system: 18S rRNA evidence. *Canadian Journal of Zoology*, **74**(11), 2031–2045.

Chen, F., Mackey, A., Jr, C. S., and Roos, D. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**, 363–368.

Chen, F., Mackey, A., Vermunt, J., and Roos, D. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **4**, 3383.

Collins, A. G. (1998). Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. *PNAS*, **95**(26), 15458–63.

Cummings, M., Otto, S., and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Molecular Biology and Evolution*, **12**(5), 814.

Darwin, C. (1859). The Origin of Species by Means of Natural Selection. *Murray*.

Dellaporta, S. L., Xu, A., Sagasser, S., Jakob, W., Moreno, M. A., Buss, L. W., and Schierwater, B. (2006). Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. *PNAS*, **103**(23), 8751–8756.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**, 361–375.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.

DeLuca, T., Wu, I.-H., Pu, J., Monaghan, T., Peshkin, L., Singh, S., and Wall, D. P. (2006). Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**(16), 2044–2046.

Diaz, N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T. (2009). TACOA -taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, **10**, 56.

Dohrmann, M., Janussen, D., Reitner, J., Collins, A., and Worheide, G. (2008). Phylogeny and Evolution of Glass Sponges (Porifera, Hexactinellida). *Systematic Biology*, **57**(3), 388.

Dolinski, K. and Botstein, D. (2007). Orthology and functional conservation in eukaryotes. *Annual Review of Genomics and Human Genetics*, **41**, 465–507.

Driskell, A., Ane, C., Burleigh, J., Mcmahon, M., O'meara, B., and Sanderson, M. (2004). Prospects for Building the Tree of Life from Large Sequence Databases. *Science*, **306**(5699), 1172.

Dunn, C., Hejnol, A., Matus, D., Pang, K., and Browne, W. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Dutilh, B., van Noort, V., and van der Heijden, R. (2007). Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, **23**(7), 815–824.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C., and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Eisen, J. (2007). Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, **5**(3), e82.

Eisen, J. A. and Fraser, C. M. (2003). Phylogenomics: intersection of evolution and genomics. *Science*, **300**(5626), 1706–7.

Ender, A. and Schierwater, B. (2003). Placozoa Are Not Derived Cnidarians: Evidence from Molecular Morphology. *Molecular Biology and Evolution*, **20**(1), 130.

Felsenstein, F. (2004). Inferring phylogenies. *Sinauer Associates*.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Zoology*, **27**(4), 401.

Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, S., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., and Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, **36**(Database issue), D281.

Finnerty, J. R. and Martindale, M. Q. (1997). Homeoboxes in Sea Anemones (Cnidaria; Anthozoa): A PCR-Based Survey of *Nematostella vectensis* and *Metridium senile. The Biological Bulletin*, **193**(1), 62.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**(2), 99–113.

Gadagkar, S. R., Rosenberg, M., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology*, **304B**(1), 64–74.

Galtier, N. and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society*, **363**, 4023–4029.

Gatesy, J., Matthee, C., Desalle, R., and Hayashi, C. (2002). Resolution of a Supertree/Supermatrix Paradox. *Systematic Biology*, **51**(4), 652.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, **312**(5778), 1355–9.

Giribet, G. (2009). Perspectives in Animal Phylogeny and Evolution. *Systematic Biology*, **58**(1), 159.

Giribet, G., Edgecombe, G. D., and Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, **413**(6852), 157.

Goodstadt, L. and Ponting, C. P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Computational Biology*, **2**(9), e133.

Gray, J. E. (1867). Notes on the arrangement of sponges, with the description of some new genera. *Proceedings of the Zooological Society of London*, **2**, 492–558.

Graybeal, A. (1998). Is it Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Systematic Biology*, **47**(1), 9.

Haeckel, E. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der Organischen Formen– Wissenschaft, Mechanisch begründet durch die von Charles Darwin reformirte Descendenz–Theorie. *Georg Reimer Verlag*.

Haen, K. M., Lang, B. F., Pomponi, S. A., and Lavrov, D. V. (2007). Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? *Molecular Biology and Evolution*, **24**(7), 1518–27.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, **68**(4), 669–685.

Hansen, S. K., Rainey, P. B., Haagensen, J. A. J., and Molin, S. (2007). Evolution of species interactions in a biofilm community. *Nature*, **445**(7127), 533–6.

Harbison, G. R. (1985). On the classification and evolution of Ctenophora. *The origins and relationships of lower invertebrates*, **28**, 78–100.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Muller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., and Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society*, **276**(1677), 4261–4270.

Hillis, D., Pollock, D., Mcguire, J., and Zwickl, D. (2003). Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? *Systematic Biology*, **52**(1), 124.

Hooper, J. and van Soest, R. (2002). Systema Porifera. A Guide to the Classification of Sponges. *New York: Kluwer Academic/Plenum Publishers*.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucleic Acids Research*, **35**(Database issue), D610–7.

Hudson, R. R. (1983). Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. *Evolution*, **37**(1), 203–217.

Huson, D., Auch, A., Qi, J., and Schuster, S. (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Hwang, U. W., Friedrich, M., Tautz, D., Park, C. J., and Kim, W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature*, **413**(6852), 154.

Inagaki, Y., Susko, E., Fast, N. M., and Roger, A. J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. *Molecular Biology and Evolution*, **21**(7), 1340–9.

Iseli, C., Jongeneel, C., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*.

James, W. and Mark, B. (2004). prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics*, **5**(187).

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, **22**(4), 225–231.

Johnson, T. (2007). Reciprocal best hits are not a logically sufficient condition for orthology. *Arxiv preprint*.

Kolaczkowski, B. and Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**(7011), 980–984.

Koonin, E. (2001). An apology for orthologs-or brave new memes. *Genome Biology*, **2**(4).

Kopp, A. and True, J. (2002). Phylogeny of the Oriental *Drosophila melanogaster* Species Group: A Multilocus Reconstruction. *Systematic Biology*, **51**(5), 786.

Krause, L., Diaz, N., Goesmann, A., Kelley, S., Nattkemper, T., Rohwer, F., Edwards, R., and Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, pages 1–10.

Kruse, M., Leys, S. P., Müller, I. M., and Müller, W. E. (1998). Phylogenetic position of the Hexactinellida within the phylum Porifera based on the amino acid sequence of the protein kinase C from *Rhabdocalyptus dawsoni*. *Journal of Molecular Evolution*, **46**(6), 721–8.

Lartillot, N. and Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, **21**(6), 1095–1109.

Lartillot, N. and Philippe, H. (2008). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philosophical Transactions of the Royal Society*, **363**, 1463–1472.

Lecointre, G., Philippe, H., Le, H. L. V., and Guyader, H. L. (1993). Species Sampling has a major impact on phylogenetic inference. *Molecular Phylogenetics and Evolution*, **2**(3), 205–224.

Lee, B., Hong, T., Byun, S., Woo, T., and Choi, Y. (2007). Estpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. *Nucleic Acids Research*, **35**, 159–162.

Lerat, E., Daubin, V., and Moran, N. A. (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the beta-Proteobacteria. *PloS Biology*, **1**(1), 101.

Li, L., Jr, C. S., and Roos, D. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, **13**, 2178–2189.

Lin, Y.-H., Mclenachan, P., Gore, A., Phillips, M., Ota, R., Hendy, M., and Penny, D. (2002). Four New Mitochondrial Genomes and the Increased Stability of Evolutionary Trees of Mammals from Improved Taxon Sampling. *Molecular Biology and Evolution*, **19**(12), 2060.

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. *Systematic Biology*, **56**(3), 504–14.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, **19**(1), 1–7.

Löytynoja, A. and Milinkovitch, M. C. (2001). Molecular phylogenetic analyses of the mitochondrial ADP-ATP carriers: The Plantae/Fungi/Metazoa trichotomy revisited. *PNAS*, **98**(18), 10202–10207.

Maddison, D. R., Schulz, K.-S., and Maddison, W. P. (2007). The Tree of Life Web Project. *Zootaxa*, **2007**, 19–40.

Maddison, W. (1997). Gene Trees in Species Trees. *Systematic Biology*, **46**(3), 523.

Maddison, W. P. and Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**(1), 21–30.

Madsen, O., Scally, M., Douady, C. J., Kao, D. J., Debry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W., and Springer, M. S. (2001). Parallel adaptive radiations in two major clades of placental mammals. *Nature*, **409**(6820), 610.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–80.

Martin, A. and Burg, T. (2002). Perils of Paralogy: Using HSP70 Genes for Inferring Organismal Phylogenies. *Systematic Biology*, **51**(4), 570.

Martindale, M. and Henry, J. Q. (1997). Reassessing embryogenesis in the Ctenophora: the inductive role of e1 micromeres in organizing ctene row formation in the 'mosaic' embryo, *Mnemiopsis leidyi. Development*, **124**, 1999–2006.

Martindale, M., Finnerty, J., and Henry, J. Q. (2002). The Radiata and the evolutionary origins of the bilaterian body plan. *Molecular Phylogenetics and Evolution*, **24**(3), 358–265.

Martinez, D., Bridge, D., Masuda-Nakagawa, L. M., and Cartwright, P. (1998). Cnidarian homeoboxes and the zootype. *Nature*, **393**(6687), 748.

Mason-Gamer, R. and Kellogg, E. (1996). Testing for Phylogenetic Conflict Among Molecular Data Sets in the Tribe Triticeae (Gramineae). *Systematic Biology*, **45**(4), 524.

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, **4**(1), 63–72.

Mclysaght, A. and Huson, D. H. (2005). OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements. *RECOMB 2005 Workshop on Comparative Genomics*, **3678**.

Meinicke, P. (2009). UFO: a web server for ultra-fast functional profiling of whole genome protein sequences. *BMC Genomics*, **10**, 409.

Mering, C. V., Hugenholtz, P., Raes, J., Tringe, S., Doerks, T., Jensen, L., Ward, N., and Bork, P. (2007). Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science*, **315**(5815), 1126.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Miller, D. J. and Ball, E. E. (2008). Animal evolution: Trichoplax, trees, and taxonomic turmoil. *Current Biology*, **18**(21), R1003–5.

Minelli, A. (2009). Perspectives in Animal Phylogeny & Evolution. *Oxford University Press*, **1**.

Miyamoto, M. M. and Fitch, W. M. (1995). Testing Species Phylogenies and Phylogenetic Methods with Congruence. *Systematic Biology*, **44**(1), 64–76.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**(6820), 614.

Mushegian, A., Garey, J., Martin, J., and Liu, L. (1998). Large-Scale Taxonomic Profiling of Eukaryotic Model Organisms: A Comparison of Orthologous Proteins Encoded by Human, Fly, Nematode, and Yeast Genomes. *Genome Research*, **8**, 590–598.

Neigel, J. E. and Avise, J. C. (1986). Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. *Evolutionary Processes and Theory*, pages 515–534.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, **16**(7), 358–364.

Nielsen, C. (2008). Six major steps in animal evolution: are we derived sponge larvae? *Evolution & Development*, **10**(2), 241–257.

Page, RDM, Holmes, and EC (1988). Molecular Evolution: a Phylogenetic Approach. *Blackwell Synergy*.

Pagel, M. and Meade, A. (2004). A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology*, **53**(4), 571.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, **5**(5), 568–583.

Peterson, D. K. J. (2001). Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evolution & Development*, **3**(3), 170–205.

Peterson, K. J. and Butterfield, N. (2005). Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record. *PNAS*, **102**(27), 9547–9552.

Peterson, K. J. and Eernisse, D. J. (2001). Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evolution & Development*, **3**(3), 170–205.

Philippe, H. (1997). Rodent Monophyly: Pitfalls of Molecular Phylogenies. *Journal of Molecular Evolution*, **45**, 712–715.

Philippe, H. and Germot, A. (2000). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Molecular Biology and Evolution*, **17**(5), 830–4.

Philippe, H., Chenuil, A., and Adoutte, A. (1994). Can the cambrian explosion be inferred through molecular phylogeny ? *Development*, pages 15–25.

Philippe, H., Snell, E., Bapteste, E., Lopez, P., Holland, P., and Casane, D. (2004). Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. *Molecular Biology and Evolution*, **21**(9), 1740.

Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene Analyses of Bilaterian Animals corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution*, **22**(5), 1246.

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**(8), 706–12.

Pick, K., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D., Wrede, P., Wiens, M., Alie, A., Morgenstern, B., Manuel, M., and Wörheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships. *Molecular Biology and Evolution*.

Poe, S. and Swofford, D. L. (1999). Taxon sampling revisited. *Nature*, **398**(6725), 299.

Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genet*, **2**(10), e173.

Qiu, Y.-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., Zimmer, E. A., Chen, Z., Savolainen, V., and Chase, M. W. (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature*, **402**(6760), 404.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**(5), 1041–52.

Rieger, R. M. (1976). Monociliated epidermal cells in Gastrotricha: Significance for concepts of early metazoan evolution. *Zeitschrift für zoologische Sytematik und Evolutionsforschung*, **14**.

Rokas, A., King, N., Finnerty, J., and Carroll, S. (2003a). Conflicting phylogenetic signals at the base of the metazoan tree. *Evolution & Development*, **5**(4), 346–359.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003b). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**(6960), 798.

Rokas, A., Kruger, D., and Carroll, S. (2005). Animal Evolution and the Molecular Signature of Radiations Compressed in Time. *Science*, **310**(5756), 1933.

Rosenberg, M. and Kumar, S. (2003). Taxon Sampling, Bioinformatics, and Phylogenomics. *Systematic Biology*, **52**(1), 119.

Sanderson, M., Driskell, A., Ree, R., Eulenstein, O., and Langley, S. (2003). Obtaining Maximal Concatenated Phylogenetic Data Sets from Large Sequence Databases. *Molecular Biology and Evolution*, **20**(7), 1036.

Satta, Y., Klein, J., and Takahata, N. (2000). DNA archives and our nearest relative: the trichotomy problem revisited. *Molecular Phylogenetics and Evolution*, **14**(2), 259–75.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **38**(Database issue), D5–16.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Dellaporta, S. L., Kolokotronis, S.-O., and Desalle, R. (2009). Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biology*, **7**(1), e20.

Schmid, R. and Blaxter, M. (2008). annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, **9**(180).

Schneider, A., Dessimoz, C., and Gonnet, G. (2007). OMA Browser Exploring orthologous relations across 352 complete genomes. *Bioinformatics*, **23**(16), 2180–2182.

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G., and Morgenstern, B. (2009a). OrthoSelect: a protocol for selecting orthologous groups in phylogenomics. *BMC Bioinformatics*, **10**, 219.

Schreiber, F., Wörheide, G., and Morgenstern, B. (2009b). OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences. *Nucleic Acids Research*, **37**(Web Server issue), W185–8.

Schreiber, F., Gumrich, P., Daniel, R., and Meinicke, P. (2010). Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, **26**(7), 960–961.

Shafer, P., Lin, D., and Yona, G. (2006). EST2Prot: mapping EST sequences to proteins. *BMC Genomics*, **7**(41).

Signorovitch, A. Y., Dellaporta, S. L., and Buss, L. W. (2005). Molecular signatures for sex in the Placozoa. *PNAS*, **102**(43), 15518–15522.

Smith, R., Buchser, W., Lemmon, M., and Pardinas, J. (2008). EST Express: PHP/MySQL based automated annotation of ESTs from expression libraries. *BMC Bioinformatics*, **9**(186).

Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L., Signorovitch, A. Y., Moreno, M. A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I. V., Buss, L. W., Schierwater, B., Dellaporta, S. L., and Rokhsar, D. S. (2008). The Trichoplax genome and the nature of placozoans. *Nature*, **454**(7207), 955.

Stiller, J. W. and Hall, B. D. (1997). The origin of red algae: Implications for plastid evolution. *PNAS*, **94**, 4520–4525.

Strahm, Y., Powell, D., and Lefèvre, C. (2006). EST-PAC a web package for EST annotation and protein sequence prediction. *Source Code for Biology and Medicine*, **1**(2).

Syed, T. and Schierwater, B. (2002). The evolution of the placozoa: A new morphological model. *Senckenbergiana lethaea*, **82**(1), 315–324.

Tajima, F. (1983). Evolutionary Relationship of DNA Sequences in finite Populations. *Genetics*, **105**(2), 437.

Tatusov, R., Koonin, E., and Lipman, D. (1997). A Genomic Perspective on Protein Families. *Science*, **278**(5338), 631.

Tatusov, R., Fedorova, N., Jackson, J., and Jacobs, A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(41).

Thiel, V., Blumenberg, M., Hefter, J., Pape, T., Pomponi, S. A., Reed, J., Reitner, J., Wörheide, G., and Michaelis, W. (2002). A chemical view of the most ancient metazoa – biomarker chemotaxonomy of hexactinellid sponges. *Naturwissenschaften*, **89**, 60–66.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, **449**(7164), 804.

Uriz, M.-J., Turon, X., Becerro, M. A., and Agell, G. (2003). Siliceous spicules and skeleton frameworks in sponges: Origin, diversity, ultrastructural patterns, and biological functions. *Microscopy Research and Technique*, **62**(4), 279–299.

van Niel, C. B. (1955). Perspectives and Horizons in Microbiology. *Rutgers University Press*, pages 3–12.

Voigt, O., Collins, A. G., Pearse, V. B., Pearse, J. S., Ender, A., Hadrys, H., and Schierwater, B. (2004). Placozoa — no longer a phylum of one. *Current Biology*, **14**(22), R944–R945.

Wallberg, A., Thollesson, M., Farris, J. S., and Jondelius, U. (2004). The phylogenetic position of the comb jellies (Ctenophora) and the importance of taxonomic sampling. *Cladistics*, **20**(6), 558–578.

Wang, X. and Lavrov, D. V. (2007). Mitochondrial genome of the homoscleromorph *Oscarella carmela* (Porifera, Demospongiae) reveals unexpected complexity in the common ancestor of sponges and other animals. *Molecular Biology and Evolution*, **24**(2), 363–73.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–62.

Whelan, S., Lio, P., and Goldman, N. (2001). Molecular phylogenetics: state-of-the- art methods for looking into the past. *Trends in Genetics*, **17**(5), 262–272.

Whitfield, J. B. and Lockhart, P. J. (2007). Deciphering ancient rapid radiations. *Trends in Ecology & Evolution*, **22**(5), 258–65.

Wiens, J. (2003). Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology*, **52**(4), 528.

Wikipedia (2010a). Wikipedia: *Bathocyroe fosteri. Website.*

Wikipedia (2010b). Wikipedia: *Chrysaora fuscescens. Website.*

Wikipedia (2010c). Wikipedia: *Trichoplax adhaerens. Website.*

Wikipedia (2010d). Wikipedia: *Xestospongia testudinaria. Website.*

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, **51**(2), 221.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS*, **74**(11), 5088–5090.

Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2004). Coelomata and Not Ecdysozoa: Evidence From Genome-Wide Phylogenetic Analysis. *Genome Research*, **14**, 29–36.

Woollacott, R. R. M. (1995). Flagellar basal apparatus and its utility in phylogenetic analyses of the Porifera. *Journal of Morphology*, **226**(3), 247–265.

Wu, M. and Eisen, J. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, **9**, R151.

Xu, H., Yang, L., Xu, P., Tao, Y., and Ma, Z. (2007). cTrans: generating polypeptide databases from cDNA sequences. *Proteomics*, **7**, 177–179.

Yang, Z., Goldman, N., and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum- likelihood phylogenetic estimation. *Molecular Biology and Evolution*, **11**(2), 316.

Zmasek, C. and Eddy, S. (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *feedback*, **3**(14).

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, **8**(2), 357–366.

Curriculum vitae

Fabian Schreiber +49(163)735466778
Zimmermannstr. 3, App. 56 Fab.Schreiber@gmail.com
37075 Göttingen, Germany

Education

- **Ludwig Maximilians University** Munich, Germany
  PhD Studies of Molecular Biology *April 2009 - Dec. 2009*
- **Georg-August University** Göttingen, Germany
  PhD Studies of Molecular Biology *July 2007 - present*
- **Georg-August University** Göttingen, Germany
  M.Sc., Applied Computer Science (Very Good (1.5)) *May. 2007*
  - Master Thesis: A phylogeny pipeline and its application to contribute to resolving the phylogeny of sponges (phylum Porifera)
- **Turun Yliopisto (University of Turku)** Turku, Finland
  Studies of Biology *Sep. 2006 - Dec. 2006*
- **Georg-August University** Göttingen, Germany
  Master Studies of Applied Computer Science (Major: Biology) *Sep. 2005 - May. 2007*
- **Georg-August University** Göttingen, Germany
  B.Sc., Applied Computer Science (Good (1.9)) *July. 2005*
  - Bachelor Thesis: Ein HMM zur Detektion von Transkriptionsfaktorbindestellen: Integration von PWM- und phylogenetischen Informationen - Predicting transcription factor binding sites with a Hidden Markov Model
- **Georg-August University** Göttingen, Germany
  Bachelor Studies of Applied Computer Science (Major: Biology) *Sep. 2002 - July. 2005*

Work Experience

- **Dept. of Bioinformatics (Biology) (Univ. of Göttingen)** Göttingen, Germany
  Student assistant *May 2007 - July 2007*
  - Designed a tool to find orthologues sequences in (EST) databases.
  - Working on a tool for the detection of pseudogenes.
- **Dept. of Bioinformatics (Biology) (Univ. of Göttingen)** Göttingen, Germany
  Student assistant *July 2006 - Aug. 2006*
  - Implemented a GUI for BLAST searches against local databases.
  - Minor scripting in Perl.
- **CAS-MPG Partner Institute for Computational Biology** Shanghai, China
  Internship *March 2006 - May. 2006*

  - Comparison of non-alignment and alignment-based methods for phylogenetic trees using HIV data.

- **Dept. of Bioinformatics (Medicine) (Univ. of Göttingen)** Göttingen, Germany
  Student assistant *Sept. 2005 - March 2006*
  - Evaluation and refactoring of the HMM tool, developed during my bachelor thesis.
- **Transcriptome Analysis Laboratory (Univ. of Göttingen)** Göttingen, Germany
  Internship *Aug. 2004 - Oct. 2004*
  - Set up a database in PL/SQL and a GUI in Perl/TK for the lab members.
  - Worked on implementing a webinterface in Perl/PHP for searching local databases.
- **Clinical Center Göttingen** Göttingen, Germany
  Student assistant *January 2004 - Dec. 2004*
  - Maintaining and curation of local databases.
  - Support for evaluation programs.

Scholarships

GGNB Travel Grant	2010
E-Fellows online scholarship	since 2007
Erasmus scholarship for studying one semester abroad	2006

Professional Activities

Referee service: *Bioinformatics*, *Proteome Science*, *Molecular Phylogenetics and Evolution*

Elected student member of examination board	2005-2007
Member of search committee for a professorship for computational neuroscience	2004
Elected as student represenative to the faculty parlament of the faculty of mathematics	2003-2004

Papers in Peer Reviewed Journals

- Hervé Philippe, Romain Derelle, Philippe Lopez, Kerstin Pick, Carole Borchiellini , Nicole Boury-Esnault, Jean Vacelet, Emmanuelle Deniel, Evelyn Houliston, Eric Quéinnec, Corinne Da Silva, Patrick Wincker, Hervé Le Guyader, Sally Leys, Daniel J. Jackson, Bernard M. Degnan, **Fabian Schreiber**, Dirk Erpenbeck, Burkhard Morgenstern, Gert Wörheide, and Michal Manuel.
  Phylogenomics restores traditional views on deep animal relationships.
  Current Biology (2009) 19, 706-712.

- **Fabian Schreiber**, Gert Wörheide and Burkhard Morgenstern.
  OrthoSelect: A web server for selecting orthologous gene alignments from EST sequences.
  Nucleic Acids Research (2009) 37, W185-W188.

- **Fabian Schreiber**, Kerstin Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
  OrthoSelect: A protocol for selecting orthologous groups in phylogenomics.
  BMC Bioinformatics (2009) 10, 219.

- Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke.
  HIV Classification using Coalescent Theory.
  Bioinformatics (2010), doi:10.1093/bioinformatics/btq159.

- Kerstin S. Pick[1], Hervé Philippe[1], **Fabian Schreiber**, Dirk Erpenbeck, Daniel J. Jackson, Petra Wrede, Mathias Wiens, Alexandre Alié, Burkhard Morgenstern, Michael Manuel, and Gert Wörheide.
  Broader phylogenomic sampling improves the accuracy of non-bilaterian relationships.
  Molecular biology and evolution (2010), doi:10.1093/molbev/msq089.

- **Fabian Schreiber**, P. Gumrich, R. Daniel, and P. Meinicke.
  Treephyler: fast taxonomic profiling of metagenomes.
  Bioinformatics (2010), 26(7):960-961.

Posters at Conferences

- Katharina J. Hoff, **Fabian Schreiber**, Maike Tech, Peter Meinicke
  The effect of sequencing errors on metagenomic gene prediction
  Presented at the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB) 2009, Stockholm, Sweden

- **Fabian Schreiber**, Kerstin Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
  OrthoSelect: A protocol for selecting ortholog groups in phylogenomics.
  Presented at the Conference *Celebrating Darwin: From the Origin of Species to Deep Metazoan Phylogeny* (2009), Berlin, Germany

- Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke
  Classification of HIV-1 Using Coalescent Theory
  Presented at the German Conference on Bioinformatics, GCB 2008, Dresden, Germany

- Ingo Bulla, Anne-Kathrin Schultz, **Fabian Schreiber**, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke
  Classification of HIV-1 Using Coalescent Theory
  Presented at the European Conference on Computational Biology, ECCB 2008, Cagliari, Italy

- **Fabian Schreiber**, Kerstin Pick, Dirk Erpenbeck, Gert Wörheide and Burkhard Morgenstern.
  OrthoSelect: A protocol for selecting ortholog groups in phylogenomics.
  Presented at the Göttingen Graduate School for Neurosciences and Molecular Biosciences opening (2008).

Skills

Computer Languages: C/C++, LaTeX, Groovy, Java, SQL, PL/SQL, HTML, PHP, Perl, XML, XSLT, XPath, XQuery

Operating Systems: Linux (Ubuntu), UNIX, MacOS X, Windows 95/98/XP

Web Development: HTML, PHP, Javascript, CSS, Grails

Applications: LaTeX, OpenOffice, MS Office XP, Photoshop, Emacs, Eclipse, Visual C++, Vim, several bioinformatic tools (e.g. Paup, MrBayes, Phylip)

[1] These authors contributed equally

Interests

Academic: Molecular evolution, molecular phylogeny, Orthology selection, Construction of software pipelines, Metagenomics

Sports: Fitness, playing hockey, table tennis, tennis and soccer

Other: Languages (English (TOEFL Score: 620), French (2 Years), Latin (Latin proficiency certificate), Spanish (current)).