Extracting the evolutionary signal from genomes

Een wetenschappelijke proeve op het gebied van de Medische Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen in het openbaar te verdedigen
op maandag 15 oktober 2007 om 10:30 uur precies

door

Bas E. Dutilh

geboren op 2 mei 1976
te Utrecht

Promotor:

Prof. dr. Martijn A. Huynen

Co-promotor:

Dr. Berend Snel

Manuscriptcommissie:

Prof .dr. Gert Vriend (voorzitter)

Prof. dr. Paulien Hogeweg (Universiteit Utrecht)

Prof. dr. Yves van de Peer (Universiteit Gent)

ISBN 978-90-9022109-0

Contents

Introduction

Evolution and the tree of life

According to theory, evolution consists of two processes: mutation and selection. In principle, every organism inherits its genes from its parent(s), but mutation can change the genotype of the offspring. Selection then determines how fit the changed offspring is. Fitness is a complex property in which the genome and the environment in which the organism lives both play a crucial role but in short, what counts is whether the organism can survive and reproduce, in the short as well as in the long run. Mutations can easily decrease the fitness: the mutant only produces sterile offspring, becomes sterile itself, dies before it can reproduce, or does not get born at all. If a mutation sticks around in the population, it is mostly because it is neutral: it causes no increase or decrease in the fitness of the organism. Because in general, few mutations are accepted, and the generation time of species is long, evolution is a very gradual process, and the genomes in a population change slowly.

Sometimes, a new species emerges. Because evolution occurs so gradually, practically all speciation events happened in the past, and the most important research that looks into this process is theoretical. If we want to reconstruct the course of evolution, the only way to go is to infer it from its current-day products: the species. Given the fact that selection takes care of the continuity between related species, a logical approach to infer the evolutionary history of the species, or the tree of life, is to make a hierarchical clustering of species, based on some evolving property. Indeed, the first cytochrome *c* protein sequences that became available in the 1960s enabled the reconstruction of a phylogenetic tree (i.e. the evolutionary history of a single gene family) that showed a promising resemblance to the species tree (Fitch and Margoliash 1967; Zuckerkandl and Pauling 1965). Since then, many genes, proteins, fragments and even complete genomes have been sequenced, and in most cases, the trademark of evolution can be observed: the closer two sequences are related, the higher their similarity. Conversely, very distantly related sequences never look alike. This is a consequence of two processes. The first is the fact that sequences generally evolve randomly, the second is the high-dimensional nature of a sequence: every position in a sequence is a "dimension" that can occur in four (DNA) or twenty (protein) states. As a consequence, the paths of two evolving sequences will always diverge. Sequences that duplicated recently will always be more similar than sequences that duplicated longer ago, and two sequences that are not related will never become homologous. At the level of the protein function or the phenotype of the species (what the species looks like), this is different. Two analogous proteins can catalyze the same reaction even though they are not homologous, and two species can live a comparable ecological niche even though they are not related.

Although closely related sequences are more similar than distantly related ones, in some cases, the genes from distantly related species are unexpectedly similar, while the corresponding homologous genes from closely related species look more divergent. There are several possible explanations for this. Firstly, the gene may have duplicated in an ancient ancestor, and much later the gene copies have been differentially lost from different lineages. As a result, the remaining gene copies will look like orthologs (genes that diverged due to a speciation event), while they are actually paralogs (genes that diverged due to a gene duplication event). This process, known as unrecognized paralogy, can lead to distorted relationships between the genes in present-day species. It can be resolved if we find a species in which both the ancestral paralogs are still present. Secondly, two genes in distantly related species can be similar because of a process called horizontal gene transfer (Doolittle 1999b). This means that a gene is transferred from one species to another, and if these species are not directly related this will make them seem more similar than they actually are. Finally, it is not always easy to determine positively how genes are inter-related (Daubin et al. 2003; Gribaldo and Philippe 2002). This can be especially difficult for genes that diverged very long ago and have undergone many

mutations, but it can also happen that for some reason, selection suddenly allows high mutation rates for a certain gene in a certain species. Thus, although the study of sequences is very useful to infer the relationships between species, it is not always infallible.

Gene content trees

With the emergence of a great number of completely sequenced genomes since *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995), gene repertoires were also reported to contain a phylogenetic signal (Snel et al. 1999; Tekaia et al. 1999). This indicates that mutation and selection cause gene content to evolve similarly to gene sequences. However, as gene content is an intermediate between genotype and phenotype, and low-dimensional relative to sequences (every gene is a dimension that can occur in two states: present or absent) it has been argued that the gene repertoire can undergo convergence through selective pressures (Doolittle 1999a; Gogarten et al. 2002). Mutation can remove and alter the genes in the genome, and species can acquire genes by horizontal gene transfer (Figure 1). Together, these two processes could cause the gene content of species that live under comparable circumstances or in close proximity to one another to converge.

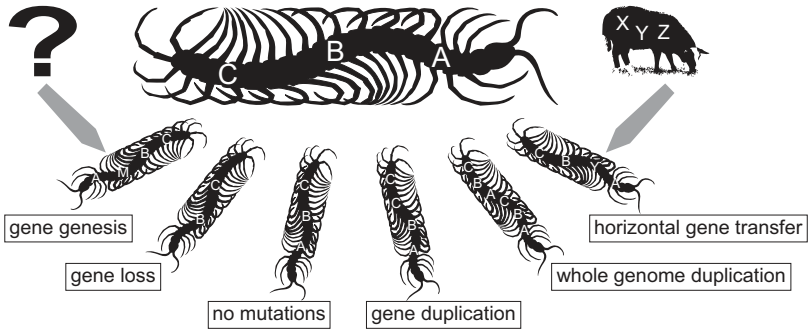

Figure 1. Examples of mutations that can occur at the gene content level. During the transfer of genes from parent to offspring, mutation can (from left to right) 'invent' a gene (new gene M), lose a gene (A is lost), do nothing, duplicate one or more genes (C is duplicated), or even duplicate the whole genome, and finally a gene may be obtained from a relatively unrelated species (Y is transferred) through Horizontal Gene Transfer (HGT). Note that in genomic research, gene invention (far left) is indistinguishable from HGT (far right) from an unknown donor.

In the chapter "The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise", we start with a data set of complete genomes, and filter out those genes whose distribution over the species forms a discordant signal in the gene content tree of life. We can remove up to 64% of the discordant genes, with very little change in the reconstructed phylogeny (Figure 7). The few shifts in the tree do not specifically affect organisms with shared phenotypic characters, e.g., parasites or hyperthermophilic species. Thus, as we do not see the effect of phenotype in the tree, such phenotypic convergence does not appear to be the cause or the result of large, systematic biases in the horizontal transfers. This is reassuring for the gene content tree of life, as it shows that the non-evolutionary acquisition or loss of genes can be considered noise on the scale of whole genomes, and complete genome data effectively averages out this noise.

Nevertheless, to obtain this result, one strong non-evolutionary signal has to be corrected for, i.e. the size of the genomes. Simply because of their size, large distantly related genomes will share more genes with one another than with closely related small genomes (Figure 3 and Figure 28). However, this genome size effect can be filtered out by a simple formula, see e.g. Equations 1 and 3 (Korbel et al. 2002).

In the chapter "Assessment of phylogenomic and orthology approaches for phylogenetic inference", we show that the Fungi contain another non-evolutionary signal that influences their gene content tree. Fungi can adopt a unicellular (yeast) or a multicellular (filamentous) lifestyle, and some species are

dimorphic (Table 2). As we show in Figure 12b, the topology of the fungal gene content tree contains a strong evolutionary signal, but it is also influenced by this lifestyle signal. In both the Ascomycota and the Basidiomycota, the filamentous fungi (Euascomycota and *P. chrysosporium*, respectively) are removed from their original positions and are drawn closer together. To our knowlede, this is the first time that such a bias has been shown to influence a gene content tree. We have not been able to identify which genes are responsible for this bias, and because lifestyle is a qualitative property, we can not filter out the effect in a way comparable to the genome size correction (Equations 1 and 3).

Phylogenomics

The word phylogenomics has several definitions, but we use it as "inferring an evolutionary tree from complete genome data". Like gene content trees average out non-evolutionary biases by considering genome-scale data, this trick can also be applied to sequence-based phylogenies. Sequence-based phylogenomic approaches combine the sophisticated molecular phylogenetic methods with the power of numbers that is inherent to genome-scale analysis. That this idea works is shown in Figure 29: inclusion of more high-quality data yields phylogenies with a smaller standard deviation and higher similarity to the accepted tree of life. This does not mean that phylogenomics is the end of all conflict in species trees. For example, there are a lot of methods to turn genomic data into a species tree (Figure 10). In the chapter "Assessment of phylogenomic and orthology approaches for phylogenetic inference" we compare all these methods, and we find that the main dichotomy is the one between trees reconstructed using a sequence-based method, and trees reconstructed using gene content data (Figure 13). The best phylogenomic tree turned out to be a maximum likelihood superalignment tree based on selected well aligned positions of unambiguous cluster orthologs (Figure 12a). Because gene content trees were biased by the fungal lifestyle signal (see above), we had to conclude that they are less succesfull at reconstructing the accepted tree of life than sequence-based phylogenomic approaches.

There is one important footnote that has to be added to this conclusion, that regards the "accepted tree of life". In order to assess the performance of the phylogenomic methods in the chapter "Assessment of phylogenomic and orthology approaches for phylogenetic inference", we derived from the literature a gold-standard topology of the fungal phylogeny (Figure 11). Although a large number of references were included (Table 11), most of them are indeed based on some kind of sequence analysis, and it is important at least to realise the danger of a circular argument. One of the things we would like to do in future research is to compare phylogenomic inference approaches with other sources of evidence, such as fossil data, laboratory evolution experiments or computer simulations. The problem with fossil data is firstly that the analyses are limited to the small subset of species that fossilizes, i.e., mostly the higher eukaryotes with hard parts like plants or animals; and secondly that the record is never complete: organisms may always be missing. The disadvantage of laboratory evolution experiments and computer simulations is that the data is likely to be much simpler than the complex traces left in present-day genomes by centuries of mutation and selection.

Taxonomy

At many points, the tree of life is resolved with general consensus about the branching orders. Other points, however, remain uncertain either because the quick succession of divergence events has made resolution difficult, or because different inference methods predict different topologies. Throughout this thesis, we suggest several improvements for the tree of life. Below are examples of contributions we made to the phylogeny of the Eukaryota and of the Bacteria.

While there is much consensus about the phylogeny of the Fungi, we retained three unresolved nodes in our target phylogeny where the literature was ambiguous (Table 12). With the 54 phylogenomic trees we reconstructed, we obtained strong evidence for the resolution of these nodes (Table

3). The node that was recovered by the fewest phylogenomic trees is the basal position of the Archiascomycetes within the Ascomycota, but this is mainly due to the lifestyle effect observed in the fungal gene content trees. For the other unresolved fungal nodes, the evidence provided by the phylogenomic trees was quite unanimous, leading to the tree in Figure 12a as our ultimate fungal phylogeny.

In the topology of the two prokaryotic kingdoms, several points of dispute remain, especially among the ancient branches. For example, the ancestral position of the hyperthermophilic bacteria *Aquifex aeolicus* and *Thermotoga maritima* has led to the assumption that the origin of life took place in a hot environment (Woese 1987), although this has been challenged on the basis of the inferred rRNA G+C content of the last common ancestor (Galtier et al. 1999). In gene content trees, these bacteria are not found at the root of the tree of life; they do not even cluster together (e.g. Figure 7). In the chapter "The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise", we show that *A. aeolicus* is affiliated to the proteobacteria, and *T. maritima* clusters with the low G+C Gram-positives, even while phylogenetically discordant signals are removed from the data set. This clustering is very strong, as we show by composing an artificial tree that groups these two organisms together at the root of the bacterial kingdom, like they are often found in sequence-based phylogenies. Then, retaining those genes whose distribution is consistent with this altered phylogeny (according to a certain threshold), we recompose a gene content tree from the reduced data set. In this tree, the two hyperthermophiles immediately jump back to their original locations. In the chapter "Signature genes as a phylogenomic tool", we show that no signature genes exist for the hyperthermophilic cluster (Figure 15), while *A. aeolicus* does share signature genes with several proteobacterial taxa, and *T. maritima* is linked to Clostridia (low G+C Gram-positive) and Archaea (Table 4). As these examples show, gene content can be a valuable complement to sequence information for resolving taxonomic relationships.

Signature genes

Especially ancient branching taxa and other clades that may be confounded by mutation saturation can be difficult to resolve correctly using sequence information (Gribaldo and Philippe 2002). Ingenious methods are being developed that reduce the sensitivity of sequence trees to these biases. One example is the slow-fast method (Brinkmann and Philippe 1999). This method first separates the analysed species into a number of pre-defined clades based on "prior knowlegde". Then, those sites in the sequence alignment are selected that do not contain any mutations within the predefined clades, but that may differ between the clades. Thus, if any is present, the phylogenetic signal within these sites can only reveal the relationships between the clades, but it will be highly reliable because it is based on slowly evolving sites. By slowly allowing more mutations in the predefined groups, the resolution within these groups can be refined by gradually adding the information from less conserved sites.

Being a relatively small research area, gene content has had to do without comparable ingenuities thus far. One aspect that has hampered the wide use of gene content data in taxonomic research is the fact that classic gene content, to ascertain the absence of genes, requires completely sequenced genomes (Snel et al. 1999; Tekaia et al. 1999). In the chapter "Signature genes as a phylogenomic tool", we introduce signature genes as a method to use the phylogenetic power in gene content for incomplete genomes. Signature genes are widespread throughout a taxonomic clade, but virtually absent outside it. Thus, they can be considered "slowly evolving" at the level of gene repertoires, making them perhaps more reliable for phylogenetic inference in a way comparable to the slowly evolving sites in the slow-fast method above. Using an intuitive, applicable definition of signature genes based on the tree of life and many of the available complete genomes (Figure 14), we identified a large set of 8,362 signature genes for 112 taxa. Once again, these many signature genes emphasize the strength of the evolutionary signal that exists in gene content. In our subsequent analyses,

we show that the presence of signature genes in an uncharacterized sample can help to detect its taxonomic composition. For example, we identify the species present in several environmental samples (Tringe et al. 2005; Venter et al. 2004), reproducing the phylogenetic marker-based results of the original publications (Figure 16).

Expression context

A gene expression profile is a comparison of the expression values of a certain gene across several tissues or experimental conditions, and it is one of the bio-informatic estimates of gene function. Genes with correlating expression profiles are likely to have related functions, especially when this correlation is conserved across different copies of the genes or between different species (van Noort et al. 2003). To be able to compare gene expression profiles between species, the same tissues or experiments have to be available for both the species compared. Large scale, genome-wide expression data is availble for only a few species so far, and these species are not closely related (Stuart et al. 2003). In the chapter "A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation", we set out to develop an approach that would enable us to compare gene expression profiles between the four distantly related Eukaryota *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Homo sapiens* (human) and *Saccharomyces cerevisiae* (yeast). To do this, we used the genome in stead of the tissues as the context in which the genes are expressed. We interpreted the gene expression profile as the co-expression of a gene with all other genes, rather than as the expression of the gene across a range of tissues or experiments. Aligning the gene repertoires of two species on the basis of orthology then makes it possible to compare the expression contexts of genes in distantly related species, where equivalent tissues do not exist (Figure 20).

For all species pairs, the expression context is slightly more conserved between orthologs than between random genes (see Figure 21). This shows two things. Firstly, our interpretation of the expression context is a meaningful measure, that says something about the context in which the genes carry out their function. Secondly, it shows that there is an evolutionary constraint on the expression context; otherwise it would not have been correlated between these divergent species. We then show that the sequence conservation of the orthologs and the conservation of their expression contexts are not correlated, which means that these two properties evolve independently. This implies that sequence identity has a limited predictive quality for detailed gene function within an orthologous group, and that annotation of different expression contexts to orthologs should not be based on sequence similarity alone. And because expression profiles are our estimate of gene function, the last chapter is a bio-informatician's warning not to take the step from sequence similarity to functional similarity too lightly.

Genome trees and the nature of genome evolution

Berend Snel, Martijn A. Huynen and Bas E. Dutilh
Annual Review of Microbiology (2005) 59: 191-209

Abstract

Genome trees are a means to capture the overwhelming amount of phylogenetic information that is present in genomes. Different formalisms have been introduced to reconstruct genome trees on the basis of various aspects of the genome, which we use to separate genome trees into five classes: 1) alignment-free trees based on statistic properties of the genome, 2) gene content trees based on the presence and absence of genes, 3) trees based on chromosomal gene order, 4) trees based on average sequence similarity, and 5) phylogenomics based genome trees. Despite their recent development, genome tree methods have already had some impact on the phylogenetic classification of bacterial species. However their main impact so far has been on our understanding of the nature of genome evolution and the role of horizontal gene transfer therein. An ideal genome tree method should be capable of using all gene families, including those containing paralogs, in a phylogenomics framework capitalizing on existing methods in conventional phylogenetic reconstruction. We expect such sophisticated methods to help us resolve the branching order between the main bacterial phyla.

Introduction

Phylogenies and genome trees

Baffled by the variety in life, one of man's first biological activities has been to classify it. Since Darwin's theory of evolution, the ultimate goal is to obtain a hierarchical classification that matches the evolutionary relations between species. This makes the construction of phylogenies one of the central activities of biologists, not only to reconstruct the history of life, but also to understand it, because "nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973). Traditionally, phylogenies were constructed from phenotypic characteristics, and phenotypic characteristics continue to play a dominant role in the analysis of data such as fossils. However, with the advent of sequencing technologies, it has become possible to construct trees on the basis of nucleotide and amino acid sequences as foreseen by Zuckerkandl & Pauling (Zuckerkandl and Pauling 1965). Sequence-based trees such as the ribosomal RNA molecules have become the golden standard in areas where phenotypic data are scarce, and are at least on equal footing in areas where we have phenotypic data as well as sequence data. Sequence-based analyses have yielded surprising observations, such as the close phylogenetic relationship between archaea and eukaryotes relative to bacteria (Gogarten et al. 1989), between fungi and animals relative to plants (Baldauf and Palmer 1993), and the monophyly of the Afrotheria (van Dijk et al. 2001). Furthermore, they can be used for organisms for which we do not have phenotypic data or for which we do not even know exist, as in the case of the environmental sampling of ribosomal RNA (Barns et al. 1994). Yet, the principle of constructing phylogenies on the basis of a single gene has been challenged (Doolittle 1999b), and for a gene such as ribosomal RNA many different phylogenetic trees have been published on the basis of different models of sequence evolution (Brochier and Philippe 2002; Olsen et al. 1994).

With the availability of complete genome sequences it has become possible to reconstruct phylogenies on the basis of much larger sets of data per species, allowing in principle a more reliable and representative inference of the tree of life. As complete genomes have been available only since 1995 (Fleischmann et al. 1995), and the methods discussed in this review are all relatively new, there is no consensus on what is the best way of integrating genome data or which genomic data should be used. Furthermore, the phylogenetic value of genome trees is not as commonly accepted as that of gene trees simply because the different parts of the genome do not necessarily have the same evolutionary history. This observation has led to the question whether it is possible to construct a phylogeny at the level of genomes (Doolittle 1999b). Given these arguments it is perhaps best to refer to a clustering of species on the basis of characteristics of complete genomes as a genome tree rather than a genome phylogeny. Genome trees then, are a means to capture and compare the overwhelming amount of information that is present in genomes and then combine this in a tree that can be interpreted as a phylogeny.

Not only are phylogenies interesting per se, all inferences in comparative biology depend on accurate estimates of evolutionary relationships (Kolaczkowski and Thornton 2004). For example, when we want to investigate how the HOX pathway evolved, we need to know the evolutionary relationships between the species in which it occurs. Similarly, comparing complete genomes using a phylogenetic tree allows researchers to study the evolution of genomic properties such as gene repertoire. Genome trees take an interesting intermediate position in this respect. In addition to being a means to derive genome-wide estimates of evolutionary relationships, they can also serve as a map on which to study the evolution of the genomes themselves. A genome tree is a direct "readout" of the processes that govern genome evolution, such as the rearrangement of chromosomal gene order. In this review we will discuss the various methods used to reconstruct genome trees, the new taxonomic insights they have given with respect to prokaryotic phylogeny, the controversies regarding their construction and the insights that they have allowed into the process of genome evolution.

Why there are so many ways to construct genome trees

A plethora of approaches to construct trees from complete genomes have been introduced (Brown et al. 2001; Daubin et al. 2002; Fitz-Gibbon and House 1999; Grishin et al. 2000; Henz et al. 2004; Huson and Steel 2004; Li et al. 2001; Otu and Sayood 2003; Qi et al. 2004b; Snel et al. 1999; Tekaia et al. 1999; Wolf et al. 2001; Yang et al. 2005). The reasons for this large variation in genome trees are twofold. Firstly, we cannot simply extend the classical approaches of sequence-based phylogenies to complete genomes. In classical molecular phylogenetics, the corresponding homologous characters in a multiple-sequence alignment, nucleotides or amino acids, are the basic elements used to infer the phylogeny. Extending that single gene phylogeny paradigm by making a long multiple-sequence alignment of genomes is not possible because evolutionary events such as gene order rearrangements, gene loss and gene duplication, occur at such high rates that even genomes from the same species cannot simply be aligned, as in *Escherichia coli*, for which the genome of different strains differ by as much as one megabase (Welch et al. 2002). Secondly, and more importantly, there are many more features to complete genomes than to genes. The sheer quantity of data, and types of data from any genome, has inspired researchers to develop new methods to cluster them. On the basis of the characters used to cluster genomes, genome trees can globally be divided into five classes (Figure 2): *(a)* alignment-free genome trees based on statistic properties of the complete genome, *(b)* gene content trees based on the presence and absence of genes, *(c)* genome trees based on chromosomal gene order, *(d)* genome trees based on average sequence similarity, *(e)* phylogenomic trees based on the collection of phylogenetic trees derived from shared gene families or on a concatenated alignment of those families.

In addition to the diversity of the information used to construct genome trees, various methods have been developed to translate the same genomic property into a tree. We classify the myriad genome trees that have appeared, on the basis of the type of genomic information, and discuss the variations in the precise phylogenetic methods that have been applied to each genomic property in the respective sections.

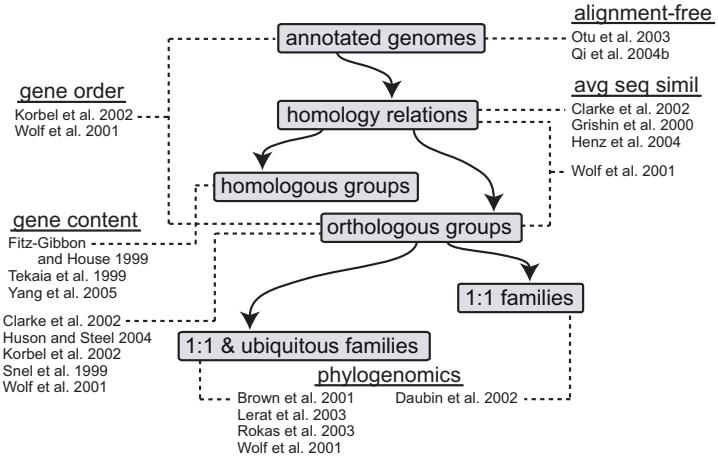

Figure 2. Classification of genome tree reconstruction methods. The genome tree publications are put in the context of the genomic property used to construct the tree. A paper that contains trees constructed with different methods is displayed in all the appropriate contexts. The amount of data available to construct a tree decreases from top (annotated genomes) to bottom (1:1 and ubiquitous families). "1:1 families" means gene families with a single copy in each genome.

Five classes of genome trees, and counting

Alignment-free genome trees: quick and dirty

Several genome tree reconstruction methods use a statistic of the entire genomic DNA, or of all encoded proteins in a genome to derive a distance between genomes that is then used to cluster them (Li et al. 2001; Otu and Sayood 2003; Qi et al. 2004a; Qi et al. 2004b). One class of alignment-free tree inference methods relies on word frequency, i.e. oligomers, K-strings, or n-mers in DNA or proteins (Vinga and Almeida 2003). The K-string method applies such a word-frequency-based method to complete genomes by simply counting the frequencies of all oligomers five or six amino acids in length in all the predicted protein sequences(Qi et al. 2004a; Qi et al. 2004b). The results are combined in a word-frequency vector, and the angle between two vectors represents the distance between two genomes. A distance-based clustering method is applied to generate the tree. This alignment-free method performs reasonably well. For example, it successfully clusters the proteobacteria as a monophyletic group (Qi et al. 2004b), unlike phylogenies based on single random, non-marker genes such as the glycolytic enzymes (Canback et al. 2002).

Another class of alignment-free methods is based on a concept from information theory (Vinga and Almeida 2003) called shared information, i.e. how much information is needed to obtain genome *a*, given that we know genome *b*. For something as complex as a genome, the specific implementations use algorithmic compression, such as Kolmogorov complexity (Li et al. 2001) or Lempel-Ziv complexity (Otu and Sayood 2003). The distance between two genomes is represented by the length of the shortest computer program to output genome *a* given the input genome *b*. These complexity measures have so far been applied to complete mitochondrial genomes from mammals, for which they accurately reconstruct the known phylogeny.

By not having to decide which genes from species *a* correspond to which genes from species *b*, these methods circumvent difficulties in orthology detection that arise from parallel gene loss and ancient gene duplications (Figure 2). They also avoid issues in single-gene tree phylogenetic reconstruction that are inherent to phylogenomic approaches, such as varying rates of evolution (see below). Furthermore, alignment-free methods are often computationally cheap and therefore one advantage is that they may provide a quick reference for obtaining the phylogenetic position of a genome or proteome as soon as it becomes available. Lastly, these are the only methods that really use all the information contained in the genome. The K-strings method uses the information from all protein coding genes, and the algorithmic compression uses the complete DNA sequence. In contrast, homology-based methods use information only from genes that have homologs in other species.

The fact that these alignment-free methods do not incorporate so much standard molecular evolutionary methodology and proven powerful evolutionary concepts, raises interesting questions, especially because they perform reasonably well. Why, for example, does the K-string complement of a proteome yield a tree that is similar to sequence-based trees? Is homology let in through the backdoor, in the form of well-conserved (i.e. identical) parts of proteins? In any case, further investigation is needed to establish which molecular evolutionary processes enable these methods to perform so well.

Genome trees based on shared gene content

A natural and convenient way to describe and analyze complete genomes is by their gene repertoire (Dandekar et al. 1998). Comparing genomes on the basis of the fraction of genes they share was one of the first comparative genomics activities to be developed with the availability of complete genome sequences (Koonin and Mushegian 1996). This worldview of genomes as bags of genes has allowed for many successful functional/evolutionary analyses, such as differential genomics (Bork et al. 1998) or phylogenetic profiles to predict protein function (Dandekar et al. 1998; Pellegrini et al. 1999).

Genome trees based on gene content are arguably the first type of genome trees of complete, organismal genomes that were published (Fitz-Gibbon and House 1999; Snel et al. 1999; Tekaia et al. 1999). Gene content trees show reasonable correspondence to the known species tree. Although this might seem trivial, given that organisms mostly inherit their genes from their parents, the concept of gene content trees has been questioned despite reports on "massive" horizontal gene transfer (HGT; also called lateral gene transfer), e.g. between archaea and hyperthermophilic bacteria (Nelson et al. 1999). Doolittle (Doolittle 1999b) argued that a unique organismal phylogeny is not conceivable unless organisms are construed as either less or more than the sum of their genes. In other words, a valid phylogeny may be derived from a gene family or from phenotypic characters, but the true map of organismal evolution cannot be represented by a tree. Rather, it should be represented by a network (Bapteste et al. 2004). Gene content trees provide a nice point of reference in this discussion, because here a genome is simply treated as the sum of its genes. That we can represent genomes in a tree is of course not an argument that genome evolution is tree-like, as any feature map can be clustered and turned into a tree. That shared gene content between genomes correlates well with evolutionary distance and that a gene content phylogeny is very similar to a sequence-based phylogeny is, however, a strong argument that genome evolution is predominantly tree-like.

As the sharing of genes is such a straightforward and logical approach to compare genomes, many different methods to make gene content trees have been introduced. The first difference between the methods is the use of orthology (Gu and Zhang 2004; Huson and Steel 2004; Korbel et al. 2002; Snel et al. 1999; Wolf et al. 2001) vs. homology (Fitz-Gibbon and House 1999; Tekaia et al. 1999; Yang et al. 2005)(Figure 2). Orthology is a more fine-grained definition of the sharing of a gene, and therefore arguably yields better trees. Today, the use of orthologs is favored over the use of homologs. In the absence of a more sophisticated method based on an explicit model of genome evolution, the tree reconstruction methods for the first gene content trees were distance based (mostly neighbor joining). Now that some consensus on the nature of genome evolution is emerging, more complicated tree reconstruction algorithms have been introduced, such as Dollo parsimony or maximum likelihood distances (Gu and Zhang 2004; Huson and Steel 2004; Wolf et al. 2001).

A major problem for gene-content-based trees is that in absolute terms large genomes of intermediate evolutionary distance, such as *E. coli* and *Bacillus subtilis,* share more genes than large genomes do with their more closely related but smaller cousins, such as *E. coli* with *Buchnera aphidicola* or *B. subtilis* with *Mycoplasma genitalium*. In fact this genome size effect is one of the strongest signals in shared gene content and thus deserves special attention when developing a distance measure (Snel et al. 1999; Yang et al. 2005). The number of genes that each eubacterium shares with a specific archaeum, such as *Sulfolobus solfataricus*, has a positive relation with very little spread (Figure 3). Most importantly, the number of shared genes saturates. For small bacterial genomes, their genome size is limiting for the number of shared genes, hence the rise, whereas for bigger genomes the archaeal genome size becomes limiting, hence the plateau. Thus, one way of correcting for this effect is to divide the number of shared genes by the number of genes in the smaller genome, the latter representing the maximum number of genes the two genomes can share. Not properly taking into account the genome size can result in gene content trees that reflect the phylogeny to a lesser extent, as they cluster, for example, small genomes together and the large genomes together (Tekaia et al. 1999; Wolf et al. 2001). Another way of handling the genome size effect is to simply leave out the small genomes (Fitz-Gibbon and House 1999). The genome size effect is intertwined with parallel gene loss, which is a major problem for gene content trees. The gene losses happen independently as well as in a coordinated fashion similar to the loss of many biosynthetic pathways in microbial organisms with a parasitic lifestyle, such as *B. aphidicola* or the mollicutes. This leads to a strong convergent signal, and although distance based methods have developed tools to manage this, it remains to be seen how well, for example, Dollo parsimony handles it. The application of Dollo parsimony by Wolf et al. (Wolf et al. 2001), and by Snel et al. (unpublished results), clusters the small genomes together. The potential of simpler methods to better cope with the issue of genome size

echoes a recent advance in conventional molecular phylogenetics. In the face of highly unequal rates of evolution, parsimony outperforms a more complicated method such as maximum likelihood (Kolaczkowski and Thornton 2004).

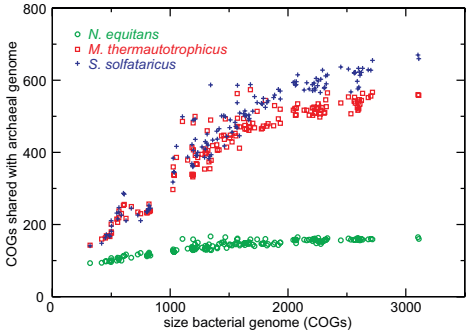

Figure 3. The genome size effect. The number of COGs shared between 45 bacterial genomes with three selected archaeal species: Nanoarchaeum equitans (323 COGs), Methanothermobacter thermoautotrophicus (1127 COGs), and Sulfolobus solfataricus (1421 COGs). Genes were assigned to orthologous groups as defined by the COG database (65). Note that there is a saturation in the number of shared COGs at larger bacterial genome sizes.

Genome trees based on gene order

Gene order, like gene content, correlates fairly well with evolutionary distance, although it does evolve faster (Dandekar et al. 1998; Huynen et al. 2001). For example, E. coli and Haemophilus influenzae share 78% of their genes, while their gene order is only conserved for 36% (Huynen et al. 2001). As gene order evolves faster than gene content, it is in principle more suited for closely related species and should achieve a higher resolution at close distances. In addition, the rate at which gene order (synteny) evolves varies between taxa. Eukaryotic chromosomal gene order, for example, evolves much faster than prokaryotic gene order (Huynen et al. 2001). Within the prokaryotes, the mollicutes appear to have a relatively well-conserved gene order, possibly because they lack the chromosome rearrangement gene recG (Suyama and Bork 2001).

Gene order has only sparsely been applied for the reconstruction of genome trees of microbial genomes (Figure 2). Apart from the high rate of genome rearrangements which leads to a lack of resolution at large evolutionary distances, this is also due to the fact, that for a large part, gene order depends on gene content. To make a gene order tree, one needs a large-scale definition of orthology. In fact, the two publications on prokaryotic gene order genome trees of which we are aware also contain gene content genome trees (Korbel et al. 2002; Wolf et al. 2001). In the first publication, the tree based on conserved gene pairs is constructed with Dollo parsimony, but unlike gene content trees based on parsimony, the small genomes in this tree cluster with their big relatives (Wolf et al. 2001). Because the rate of gene order evolution is so much higher than the rate of gene content evolution, there may be many more shared-derived features in the form of lineage specific gene pairs than there are lineage specific genes. In the other effort, the gene order and gene content trees were similar to each other (Korbel et al. 2002). The gene order tree showed some improbable higher order affiliations, reflecting a lack of resolution for these longer evolutionary distances in which too many gene rearrangements have occurred. The gene content tree behaved normal for these distances.

In contrast to microbial genomes, gene order has been successfully applied to eukaryotic mitochondrial genomes and specifically to metazoan mitochondrial genomes (Blanchette et al. 1996; Boore and Brown 1998; Sankoff et al. 1992). In fact, trees based on mitochondrial gene order are arguably the first kind of genome trees, predating gene content trees of completely sequenced organismal genomes (Sankoff et al. 1992). In this area, real algorithmic progress has been made

such as the formal definition of re-arrangement distance based on inversions, translocations, and transversions (Blanchette et al. 1996). The dense sampling has also revealed cases of extreme rate variation in mitochondrial gene order evolution, such as the accelerated rate of mitochondrial gene order evolution in Echinodermata in contrast to the near stasis of Vertebrata and Hemichordata (Castresana et al. 1998)). Moreover, mitochondrial gene order is one of the few areas in which genome trees are specifically employed by genuine taxonomists to achieve a better picture of the phylogeny of certain species (Boore and Brown 1998).

Genome trees based on average sequence similarity

The approaches reviewed above do not use sequence information other than for the definition of orthologs. This knowledge is subsequently used to determine the number of shared genes or the extent of gene order conservation, from which a similarity measure is deduced. At the complete opposite of these approaches, lies a class of methods sometimes called blastology. Here, a distance matrix is calculated on the basis of the average sequence similarity between genomes or proteomes, explicitly neglecting any knowledge of orthology (Figure 2).

Henz et al. (2004) take the most basic approach imaginable. They make BLAST comparisons at the DNA level of 91 complete prokaryotic genomes and use the resulting heat shock proteins to compose a distance matrix (Henz et al. 2004). This approach then uses the average sequence similarity between two entire genomes as a similarity measure, making no distinction between coding and non-coding regions, although in prokaryotes most heat shock proteins can be expected to fall within the coding regions. A comparable method was introduced earlier by Grishin et al. (2000), who used only the coding sequences. Rather than comparing the entire genomic DNA, the authors compare 19 complete proteomes using BLAST (Grishin et al. 2000). They constructed a tree on the basis of the interprotein amino acid substitution rate distribution of all proteins with sufficient similarity ($e < 0.01$ in their data set). Another approach based on complete proteomes was presented by Clarke et al. (2002). They built a tree on the basis of the mean normalized BLAST scores for 37 species. Significant hits were normalized by dividing the e-value by the open reading frame's self-matching score, the average normalized score defining the distance between two species (Clarke et al. 2002). In the genome tree compilation of Wolf et al. (2001), the median percent identity of bi-directional best hits between two genomes is used as a similarity measure. The sequence similarities between genomes are transformed logarithmically to obtain a distance matrix, and subsequently, neighbor joining is used to build a tree (Wolf et al. 2001).

Although these methods are very straightforward to implement, and although they can be seen as an interesting intermediate between gene-content-based approaches and purely sequence-based approaches, the compilation of genome trees based on average sequence similarity has never had much follow-up. Researchers are reluctant to adopt the method because the approaches appear to combine the problems present in trees based on gene content as well as in trees based on sequence. By using the extra layer of information provided by the orthology assignment, researchers who implement gene content approaches can avoid some of the pitfalls present in naïve sequence analysis, such as convergence in nucleotide usage and codon usage. Phylogenomics approaches, on the other hand, use the sequence information in a phylogenetically superior way. They use proper multiple-sequence alignment rather than simply averaging BLAST scores. Furthermore, the average sequence similarity approaches do not allow inclusion of any evolutionary model and prohibit the construction of trees that use maximum parsimony or likelihood, methods that could add much value to approaches based on sequence comparison. Comparing homologous genes rather than orthologous genes, as is done in these methods, basically means introducing noise. Optimally, a filter should be applied to reduce the impact of non-orthologous homologs. In fact, the tree from Wolf et al. (2001) indeed uses only similarities between bi-directional best hits in order to include only orthologs, and this improves the topology. In contrast to other average sequence similarity

genome trees (Grishin et al. 2000; Henz et al. 2004), they successfully retrieve the proteobacteria as a monophyletic clade(Wolf et al. 2001).
Meanwhile, articles that present new genomes often present a list of species for which such a new genome has a large fraction of its best BLAST hits. This practice, which is related to the tree-building method outlined above, provides a fast indication of the taxonomic neighbors of a species.

Genome trees based on gene trees: phylogenomics, supertrees and concatenated sequences

Because we cannot use traditional sequence alignment tools to compare the sequences of complete genomes, it is a logical step to at least use traditional sequence alignment tools where possible (Figure 2). The advantage is that we can use the entire toolbox of sophisticated phylogenetic reconstruction methods. One approach is to make trees of gene families that are represented in the genomes of interest. The first effort in this direction dates back to 1999 and immediately ran into the issue that trees from different genes have a different topology (Teichmann and Mitchison 1999).
To overcome such incongruent gene trees, one can simply concatenate the homologous sequences from the different gene families, as a concatenated alignment automatically yields a single tree (Brown et al. 2001). This method has had some success, but faces difficulty for evolutionary divergent organisms. The concatenated genes not only have to be present in all genomes compared, they should also have a single copy in each genome to make sure that they are indeed orthologous to one another. With the increasing number of sequenced genomes, the number of genes present with exactly one copy in all organisms shrinks dramatically. In closely related species that share many genes this method has been applied successfully, e.g. in a phylogenomic study on the _-proteobacteria (Lerat et al. 2003). Rather than (or at the same time as) making a concatenated alignment and escaping the issue of what to do with all these different phylogenies, one can also compare them, and obtain some consensus, for example by using approaches comparable to bootstrapping in single-gene trees (Lerat et al. 2003; Rokas et al. 2003). The advantage of calculating individual gene trees is that one can separate trees that are relatively different from one another, for example, because of unrecognized paralogy or HGT (Brown et al. 2001). One can even use the orthologous groups that have not been filtered out to construct a new concatenated alignment. There appears to be no straightforward answer to the question whether it is better to concatenate sequences or to integrate individual trees (Figure 4), but the differences can be striking. For example, we find that the concatenated alignment yields the same topology in neighbor joining as in maximum likelihood, while this is not true for the consensus of the phylome (Figure 4). The choice of integration can thus be more influential than the choice of precise phylogenetic method, even with methods as different as neighbor joining and maximum likelihood. On the one hand, concatenation prevents each gene family to be treated with parameters that are specific for this family. The issue with individual gene trees, on the other hand, seems to be that we do not know how to nicely integrate them, other than using a strict consensus. However, as noted, in practice both methods are applied, often in comparison to the same dataset.
The main limitation to the above methods, however, remains that they require one gene per genome per gene family (1:1 family). By relaxing this criterion, one can in principle obtain much more information from the genomes and their phylogenetic position; however, one must implement methods that compare trees with different numbers of species. One such method is the supertree method (Daubin et al. 2002). Although this method still requires a gene family to be present not more than once in each genome (to assure unambiguous orthologous relations), it eases the demand that a gene family should be present in every genome. To handle the different species compositions of the various trees, the authors created a new alignment of co-occurrence of species in all the partitions of each tree. From this new alignment, a distance matrix is created that is then fed into the neighbor-joining algorithm. This final step may be open to improvement, because it seems *ad hoc*, like many of the gene content genome tree methods. Nevertheless, the resulting tree is of excellent quality (all established prokaryotic taxa such as the Euryarchaea or the Proteobacteria are

monophyletic), possibly because of the aforementioned increase in the amount of data on which it is based.

Leaving out all restrictions on the species distribution of homologs altogether, one can simply create the phylogenies of all the genes from one genome, the phylome (Sicheritz-Ponten and Andersson 2001). Many insights other then purely phylogenetic ones can be gained from these collections of trees. One can reconstruct the metabolism of the ancestor of the mitochondria (Gabaldon and Huynen 2003) or predict functional relations between genes (Ramani and Marcotte 2003). Nevertheless, these massive phylomes have not yet been integrated into a single hypothesis on the phylogenetic relationships between all genomes.

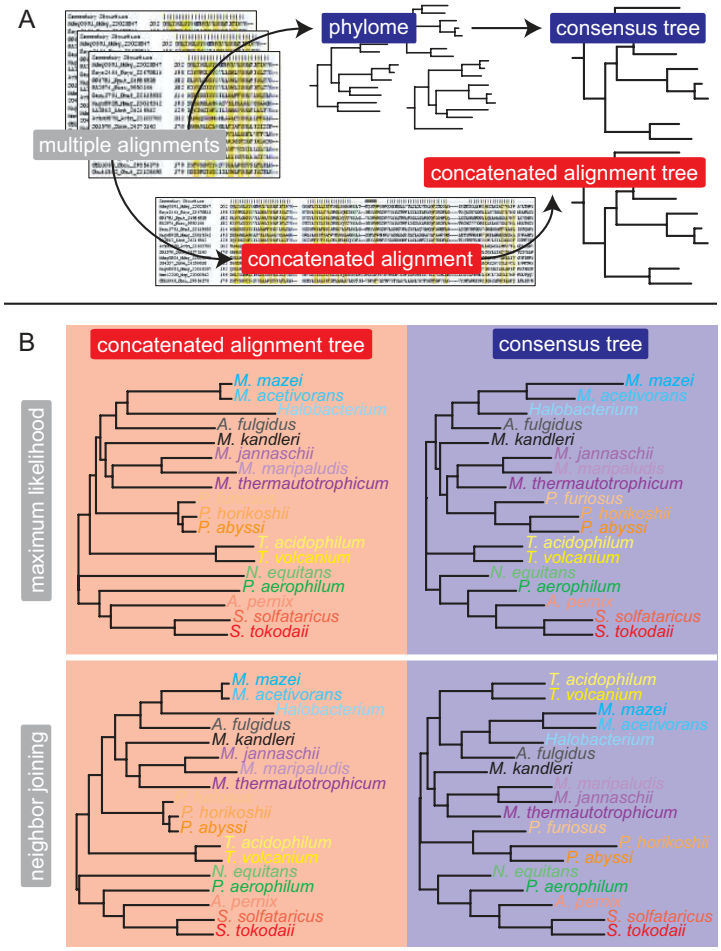

Figure 4. Concatenation versus phylome. Panel a shows the outline of the two basic methods to obtain a single tree from multiple-gene family, multiple-sequence alignments. The first strategy, phylome and consensus, first constructs trees for each individual family and subsequently extracts the strict consensus single tree from these families. The second strategy, concatenation, makes one big alignment from the multiple-sequence alignments and subsequently constructs a single tree on the basis of this alignment. Panel b shows four phylogenomics trees based on the same dataset of archaeal sequences. The dataset is the collection of 90 COGs present with a single copy in each species. For each COG, a multiple-sequence alignment was constructed by MUSCLE (20). The maximum likelihood trees were obtained by PHYML (30) at default settings. The neighbor-joining trees were reconstructed by QuickTree (33). The concatenated alignment was obtained by concatenation of the 90 alignments. The consensus trees are the strict consensus from the 90 trees as obtained by running CONSENSE from the PHYLIP package (22). Note that the concatenated alignment yields the same tree topology with neighbor joining as with maximum likelihood. This is not the case for the consensus tree.

New developments in the construction of genome trees

Filtering for inconsistent signals

HGT as well as ancient gene duplications followed by gene loss (unrecognized paralogy) lead to gene phylogenies that are inconsistent with the species phylogeny. For each class of genome trees, methods have therefore been developed to filter out genes with inconsistent histories. Phylogenomic methods have in fact almost without exception applied such filters (Brown et al. 2001). As it turns out, the incongruence in gene trees rarely has biological reasons like the above but derives mainly from varying rates of evolution and incorrect alignments (Daubin et al. 2003). Nevertheless, in sequence-based trees the filtering is generally reported to make a substantial improvement (Brown et al. 2001).

Filtering for inconsistent signals has not been so often applied for other types of genome trees. It has been applied once for gene content trees and once for average protein similarity trees (Clarke et al. 2002; Dutilh et al. 2004). Methodologically it is more difficult to define inconsistent signals for both of these types of genomic information than it is for phylogenomics methods. In both approaches, the improvements in the quality of the trees were minor. In fact it seems, at least for gene content, that the phyletic distribution of virtually each gene contains at least *some* phylogenetic information (Dutilh et al. 2004). One explanation for this observation is that genes that are horizontally transferred still behave phylogenetically concordant before and after the transfer event.

Modeling genome evolution

Phylogenies can in principle be improved by using more information than just the pair-wise distances between genes or genomes as is done in clustering methods such as neighbor joining. Not only do maximum likelihood methods include information about the rates of various processes during gene evolution, but more importantly they explicitly calculate the probability that a certain sequence alignment is produced by a specific phylogenetic tree and a specific model of evolution (Felsenstein 1981). Such an explicit model of evolution can include the rates at which certain point mutations occur. The tree that is most likely to have produced the alignment then, given the model of sequence evolution, is the maximum likelihood tree. In practice, experimentally generated, known phylogenies of bacteriophage T7 have been better reconstructed by maximum likelihood methods than by neighbor joining (Cunningham et al. 1998).

Recently, approaches that incorporate more explicit models of genome evolution have been applied to genome trees. For example, the simulation of gene content evolution in artificial genomes seems to yield a reliable maximum likelihood distance for the reconstruction of gene content trees (Gu and Zhang 2004; Huson and Steel 2004). Such simulations also suggest that a more explicit description of gene content evolution in the form of Dollo parsimony should show excellent performance. Nevertheless, the only implementation so far of Dollo parsimony on real genome data results in a tree that suffers from clustering of unrelated small genomes and paraphyly of established clades such as the γ-proteobacteria (Wolf et al. 2001). Similarly, the application of maximum likelihood on actual genome data results in clustering of small genomes (Gu and Zhang 2004).

Evolutionary insights from genome trees

New phylogenetic/taxonomic findings

A few new phylogenetic affiliations have been uncovered or were partly resolved with genome trees. One feature almost unanimously supported by genome trees, is that firmicutes (low G+C Gram-positives) and actinomycetes (high G+C Gram-positives) are polyphyletic. According to the current NCBI taxonomy, these two groups of Gram-positive bacteria are paraphyletic at the root

of the Eubacteria, whereas previously they were grouped together in one taxon, based on the 16S rRNA phylogeny. The alternative grouping is now accepted, but mainly because of other discoveries (Ahmad et al. 1999). Genome-scale phylogenetic analyses have declared a close evolutionairy relationship for various methanobacteria (Slesarev et al. 2002; Snel et al. 1999), which until now were considered polyphyletic on the basis of 16S rRNA analysis (Olsen et al. 1994). The relegation of *Fusobacterium* from a separate bacterial division to a member of the firmicutes is one of the conclusions of genome-scale analysis, which are a part of the original publication of its genome sequence (Kapatral et al. 2002; Mira et al. 2004). A more hypothetical theme, which nonetheless recurs in many gene content papers from our group and others, is that the hyperthermophilic eubacteria are not primitive, rather *Aquifex* seems to be affiliated with the proteobacteria and *Thermotoga* with the Firmicutes (Dutilh et al. 2004; Korbel et al. 2002; Qi et al. 2004b; Wolf et al. 2001). The special link of gene order with mitochondrial genome trees is reflected in the contribution of mitochondrial gene order trees to the resolution of the four basal Arthropod lineages (Boore and Brown 1998).

A limited role of horizontal gene transfer in microbial evolution

One cannot construct genome trees without wondering how we can produce a tree in the presence of HGT. Before the availability of complete genome sequences, HGT was not attributed a quantitatively major role in evolution. The sequencing of complete genomes has drastically changed this view. Publications of genome sequences, studies that specifically targeted HGT occurrence in published genomes and phylogenomic investigations, report massive levels of HGT (Nelson et al. 1999; Ochman et al. 2000; Teichmann and Mitchison 1999). Although genome trees are not direct assays of the frequency of HGT, papers describing genome trees are almost unanimous in their surprise at how well their tree based on any given genomic property matches the known species phylogeny. Most genome tree papers therefore report HGT as being quantitatively of small importance (Daubin et al. 2002; Snel et al. 1999). Part of the argument is semantic: Can we call an alien origin of 12% of the genes in *E. coli* "massive" (Ochman et al. 2000)? Yet, the frequency of HGT has even been argued to preclude the existence of a species tree, making the contradictory statements more than simply different perspectives on the same data (Doolittle 1999b).

To resolve this discussion, the investigations into genome evolution have moved beyond studies that only targeted HGT or studies that only constructed genome trees. First, we can try to estimate the occurrence of processes that affect gene content (such as gene loss, gene duplication, HGT, and the appearance of new gene families), on the basis of the distribution of current gene families in a reliable species phylogeny. Such analyses are challenging because patchy phyletic distributions of genes (in which a gene is sparsely distributed over a taxon) can be explained by HGT as well as by differential gene loss. Most approaches solve this by introducing a "cost" in terms of gene losses for each HGT event. When comparing different possible explanations for a patchy phyletic pattern, the cost of a HGT event is weighed against the cost of the of gene losses that are no longer necessary if we introduce this HGT event. Such methods allow a broad scope on genome evolution and they reveal substantial continuity in genome evolution: On any given branch, most genes are transmitted vertically even when using low costs for HGT (Snel et al. 2002) (Canback et al. 2004; Kunin and Ouzounis 2003; Mirkin et al. 2003).

The second approach studies the extent of aberrant sequence evolution of single genes due to HGT. In a seminal paper by Daubin et al. (2003) a comprehensive collection of quartets of unambiguous orthologous genes were tested for their support of the species phylogeny as defined by rRNA. The results showed that few (sometimes even zero) quartets support the two other possible trees in the case of four sequences, implying a limited role for HGT. Interestingly, they also showed that the quartet alignments that do support HGT have long terminal branches compared with the internal branches, suggesting that these might still be the result of errors in tree reconstruction (Daubin et al. 2003). It has been argued that the low level of HGT found in this study, and in a subsequent, more detailed report (Lerat et al. 2003), is the result of using only unambiguously orthologous genes, i.e.

no paralogy whatsoever (Zhaxybayeva et al. 2004). An independent and equally impressive effort, applied a similar question to the more loosely defined clusters of orthologous groups (COGs), which do contain many paralogs (Novichkov et al. 2004). Here, the sequence similarities of individual genes were compared with the average sequence similarity between two genomes. Most (70%) of the genes did not show aberrant levels of sequence similarity potentially caused by HGT. Explicit phylogenetic analysis of the remaining 30% indicated that only half of these could be due to HGT, while the other half was due to lineage-specific acceleration of evolution (Novichkov et al. 2004).

From both broad-scope views of genome evolution it can be concluded that by far most of the genes, and thus the genome, have evolved by normal vertical transmission. These explicit studies of the evolutionary dynamics of genome evolution have brought at least some researchers with opposing world views together: Gene content tree builders explicitly acknowledge the need for HGT to explain present-day genomic gene repertoires (Daubin et al. 2003; Snel et al. 2002) and to filter out its effect (Dutilh et al. 2004), while HGT hunters discovered that other tree-like processes such as vertical inheritance and gene invention are, at least quantitatively, more important for genome dynamics than HGT is (Kunin and Ouzounis 2003; Novichkov et al. 2004). Disregarding the proposal that the similarity between gene content trees and rRNA trees results from the HGT of the rRNA molecule (Gogarten et al. 2002), the outlines of a consensus on the nature of microbial genome evolution thus seem to be emerging from the literature: a quantitatively modest, but qualititatively important role for HGT and a large role for tree-like processes such as gene loss.

Conclusions

Challenges ahead: data and computation

Apart from large amounts of relatively clean sequence data in the form of complete genome sequences, the fact that DNA sequencing has become much easier has paradoxically also led to a dramatic increase in noisy data. One important development has been the emergence of metagenomics and environmental sequencing. In these techniques pieces of DNA are sequenced from uncultured samples, such as a drop of ocean water (Venter et al. 2004) or a sample of liquid from an acid mine (Tyson et al. 2004). The results of such studies are not complete genome data, but they do contain an invaluable amount of phylogenetic information that needs to be classified without using rRNA, because it is unknown which sequenced reads belong with which rRNA. In the first instance, such a sequence read equals the genome and the species. Supertree and other genome tree approaches can thus provide a phylogenetic framework for the sequences in the absence of rRNA.

Another source of growing amounts of noisy sequence data are incomplete genomes. These data are generated for prokaryotic genomes, because they provide an easy and cheap method to answer certain microbial questions (Overbeek et al. 2003). At the same time, semicomplete eukaryotic genomes are emerging because even with the current relative ease of sequencing many model species (e.g. *Gallus gallus* (chicken), *Fugu rubripes*) cannot receive the intense attention that was put into for example the human genome project (Aparicio et al. 2002; Hillier et al. 2004). These data are problematic for genome trees based on gene content, because the absence of genes can be explained as easily by not having been sequenced as by a genuine loss from the genome. However, genome tree methods that rely on sequence similarity can still be applied here.

Note that all the metagenomics data and the data from incomplete genomes are deposited in comprehensive sequence databases such as EMBL or Genbank. An interesting study in this light is the effort to build the tree of life from two of such databases, namely Swiss-Prot and a subset of Genbank (Driskell et al. 2004). From all these data, a super-matrix is compiled from all groups of phylogenetically informative homologs that are present with a single copy in every genome. This method is a good approach to dealing with these data, although taxonomic labeling of the genes is

required, which is exactly what is missing from metagenomics. Nevertheless, this review shows that genome trees can be encompassed in methodologies that integrate ever more data.

Another challenge is computational, especially in light of the speed of DNA sequencing as mentioned above. The computational demands of comparative studies logically always increase faster than the already exponential increase in data. Even given Moore's law (Moore 1965), a solution will not simply or be found only in faster computers. In addition, the understandable preference for a more phylogenomic approach to genome trees means handling computationally intensive problems such as multiple-sequence alignment and phylogenetic tree reconstruction. Solutions will come in many different shapes and sizes. Some will be algorithmic, such as those already developed for fast and reliable multiple-sequence alignments (e.g. MUSCLE (Edgar 2004a)) or maximum likelihood inference of phylogeny (e.g. PHYML, (Guindon and Gascuel 2003)). Another solution is to use the data selectively. A representative and reliable selection of genes or species will be used as references to construct a backbone for the tree of life at higher taxonomic levels. More sequences can be subsequently selected to fill in the details for lower taxonomic levels, which are established from the higher-level backbone. Such a procedure is akin to the existing supertree methods that summarize phylogenetic findings from different papers or collections of trees with different species samples (Driskell et al. 2004). As a solution for the all-against-all comparison of sequences, which promises to become a computational nightmare, profile database searches could play an important role. Profile searches do not need to be redone, as their reliability does not depend on database size and profiles also automatically give a reliable (profile anchored) multiple-sequence alignment.

In contrast to the other methods, alignment-free methods will remain computationally inexpensive. They may thereby provide an independent reference. All these efforts are bioinformatic challenges and they will bring genome trees and phylogenomics closer to a tight integration with existing sequence databases.

The future for genome trees is phylogenomics

Phylogenomics approaches are popular for good reasons. They incorporate the best of both worlds. These approaches use sophisticated existing (and continuously being developed) conventional molecular phylogenetic methods for the reconstruction of single-gene trees while they apply the power of numbers that is inherent to genome-scale analysis. As orthologous genes are formally defined by the relation of a gene tree to a species tree, phylogenomics is also part of the solution for defining orthology, which is an important challenge in gene content and gene order genome trees. Phylogenomics could thus incorporate these two other types of genome trees. For example, a gene content genome tree that would be based on an orthology derived from single-gene phylogenies is basically nothing more than an alternative to constructing a supertree from a phylome.

The continued application of phylogenomics approaches faces certain hurdles. First, there is the computational hurdle mentioned above. Secondly, most phylogenomics studies require that the genes are present in all genomes under consideration, but with more genomes this will be the case for fewer and fewer genes. This hurdle has already been overcome in part by the supertree approach (Daubin et al. 2002), but the way in which the trees and their different species compositions are integrated is open to improvement (see above). The remaining obstacle might be the requirement of one ortholog/gene per family per genome. This obstacle is related to the integration of trees with different species compositions. One solution could be sought in a more dynamic definition of orthology from the gene tree itself: using sub-trees of a gene tree with paralogs to construct a supertree.

Summarizing conclusions

The main result of genome trees, from trees based on word frequencies to trees based on single gene phylogenies, is that they are all similar to each other and reflect the known species phylogeny regardless of the various specific genomic properties used or the method used to create the

phylogeny. Perhaps we find this coherence because these properties, to various extents, depend on each other. Genome trees have yielded the fundamental insight that genome evolution is largely a matter of vertical transmission. Although the dominance of vertical transmission and thus the quantitatively minor role of HGT became a controversial claim, currently, thanks in part to genome trees, the consensus on microbial genome evolution that emerges is that gene repertoires largely follow phylogeny.

Apart from their contribution to a consensus view of genome evolution, genome trees have made some impact on the phylogeny per se. As genome trees are in line with the undisputed parts of the tree of life, they can also be treated as a line of evidence for phylogenetic relationships in inconclusive parts of the single-gene-based species phylogeny. Their impact so far has been on ancient branching points between higher-level taxa, such as the position of the Fusobacteria, and the divergence between the high and low G+C Gram-positives. So far, these contributions have been infrequent, as might be expected given the nascent state of this line of research. Yet, the relative branching orders in the bacterial as well as the eukaryotic divisions, and thus relationships between many higher-level taxa, still need resolution. Here, the quick succession of divergence events has made resolution difficult and until now the order of these events has not been resolved with single-gene phylogenies. In principle, genome trees are in a position to contribute to resolving these points. They are the means to use the maximum amount of data available to solve these tough problems and to thereby solidify the backbone of the bacterial phylogeny.

A number of hurdles remain on the path ahead. Finding solutions to these hurdles will remain a challenge to our creativity as the recognition of the added value of genome phylogenies grows. Thus far, the main contribution has been the recognition that the classic view of genome evolution by vertical inheritance is indeed quantitatively the most important. The promise of more important phylogenetic discoveries will continue to stimulate researchers in this field.

The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise

Bas E. Dutilh, Martijn A. Huynen, William J. Bruno and Berend Snel
Journal of Molecular Evolution (2004) 58: 527-539

Abstract

Phylogenetic trees based on gene repertoires are remarkably similar to the current consensus of life history. Yet it has been argued that shared gene content is unreliable for phylogenetic reconstruction because of convergence in gene content due to horizontal gene transfer and parallel gene loss. Here we test this argument, by filtering out as noise those orthologous groups that have an inconsistent phylogenetic distribution, using two independent methods. The resulting phylogenies do indeed contain small but significant improvements. More importantly, we find that the majority of orthologous groups contain some phylogenetic signal and that the resulting phylogeny is the only detectable signal present in the gene distribution across genomes. Horizontal gene transfer or parallel gene loss does not cause systematic biases in the gene content tree.

Introduction

With the availability of complete genome sequences, it has become possible to use the information contained in whole genomes to infer phylogenies (for a review see (Wolf et al. 2002)). Genome trees are created in an attempt to combine all the phylogenetic messages in all the genes. The main idea is that one can obtain a more representative phylogeny by averaging out the confounding signals in single gene trees. It has been argued that the gene repertoire is a phenetic character (Doolittle 1999a; Gogarten et al. 2002) and that gene content can undergo convergence through selective pressures. Thus, some of the processes that impair single gene trees, such as horizontal transfer (Doolittle 1999b) and parallel loss of related genes (Snel et al. 2002; Wolf et al. 2002), can, when frequent enough, also affect genome trees. For example, some phenotypic characteristics, such as a parasitic lifestyle, are reflected in a similarity in the functional classes of genes in the genome (Zomorodipour and Andersson 1999).

It is true that gene content is a more phenotypic character than gene sequence. After all, the gene repertoire determines the phenotype of an organism. We have argued that gene content phylogenies take a position intermediate to phylogenies based on single genes and phylogenies based on phenotypic characteristics (Snel et al. 1999). However sequence evolution can also reflect the phenotype, e.g., thermophily is reflected in the amino acid content of a genome (Cambillau and Claverie 2000; Kreil and Ouzounis 2001; Suhre and Claverie 2003), and in general sequence-based phylogenetics can suffer from homoplastic events. Fast-evolving positions create a problem when inferring ancient phylogenetic relationships, adding noise rather than signal to the data (Gribaldo and Philippe 2002). Unless parallel gene loss and horizontal gene transfer occur along demarcated transfer routes, these processes will also only add noise. Sequence analysis has developed tools to identify and remove this noise (Bruno et al. 2000; Goldstein and Pollock 1994). Because the gene presence/absence profile is a binary sequence in all organisms, we can use similar tools to remove noise from genome phylogenies (Brown et al. 2001; Clarke et al. 2002). Clarke et al. (2002) suggested an implementation in which they rid the genome of phylogenetically discordant signals (PDSs) by applying a filter that identified horizontal transfers as sequences with an irregular ranking of the BLAST expectancy values of their orthologs. Removing these PDSs did improve bootstrap support for basal nodes in the phylogeny but, aside from that, altered hardly any topological features.

In the current investigations, we reduce the impact of noise in gene content phylogenies by two schemes that treat the presence/absence profiles as sequence alignments. As we identify PDSs by examining their species distribution, we avoid the pitfalls inherent in sequence analysis, unlike Clarke et al. (2002), who reverted to sequence comparison for identification of the PDSs. By using both orthology and sequence information, Clarke et al. try to combine possibly inconsistent sources of information. This approach can be expected to erroneously identify sequences in rapidly evolving lineages as phylogenetically discordant. Our approach should be less sensitive to this long-branch artifact, as the orthology assignment (Tatusov et al. 2001; von Mering et al. 2003) suffers from this problem only to a small extent. We identify as PDSs instances of horizontal gene transfer, and contrary to Clarke et al. (2002), our schemes also identify parallel gene loss as PDS. As we take orthologous groups as the starting material, orthologous gene displacement within an orthologous group is not identified as a discordant signal.

The coding of genomes as binary sequences allows our approaches to deal with noise from fast-evolving positions. First, we use a method that finds PDSs in a reconstructed genome phylogeny and removes them from the data set. To properly incorporate changes, construction of phylogenies, and identification and removal of PDSs are repeated iteratively until the trees converge. To further test whether the phylogenetic signal in gene content is the only dominant signal, we also used this approach to determine to which topology our trees converge from 100 different random initial topologies.

Second, we use an adapted method that was originally developed for assessing amino acid sequences (Bruno 1996). Assigning high weights to the clade specific genes, and low weights to genes that evolve rapidly, we were able to scale down the impact of noisy signals and infer a filtered phylogeny.

Methods

Orthology

To be able to compare genomes based on their gene content, it is first necessary to identify which genes are shared between genomes, i.e., which genes are orthologs. Orthologs are genes in different species that are directly related by vertical inheritance (Fitch 1970). Paralogs are genes within a species that are derived from gene duplication. If a group of paralogs in a certain species has dispersed after the latest speciation event, all these genes will have the same orthology relationship with their relatives in the sister species. Thus, groups of orthologs will best represent the ancestral relationships of a collection of genes in a set of species.

Inferring orthology relationships is far from trivial, especially because orthology has been defined for the comparison between two species (Fitch 1970). It is not unusual in comparative genomics to define as orthologs those homologs that have a BLAST expectation value lower than a certain threshold (e.g., (Bansal and Meyer 2002; Fitz-Gibbon and House 1999)). Another operational definition of orthology that is often used is that of reciprocal best BLAST matches (called BeTs (Tatusov et al. 1997), BBHs (Tamames 2001), or RBMs (Clarke et al. 2002)). Although this definition will be a closer approximation of the evolutionary definition of orthology than the close homologs method, it does not give us directly a group orthology that is best suited for our study. A very suitable database of groups of orthologous genes is the manually curated COG database (Clusters of Orthologous Groups of Proteins; NCBI; see www.ncbi.nlm.nih.gov/COG (Tatusov et al. 1997)). Within each of the 3166 COGs, the proteins are assumed to have evolved from the same ancestral gene, and if present, the COG is represented by an individual protein or a group of paralogs within a certain species. We use this database, extended by von Mering et al. (2003) to a current total of 19,433 orthologous groups (OGs) in 89 completely sequenced genomes (for more information see www.bork.embl-heidelberg.de/STRING), to compare these organisms on the basis of their gene content.

Distance Measure

For each OG, a binary profile was created, indicating its presence (1) or absence (0) in the 89 genomes considered (see Figure 5). Using these profiles as a similarity measure, a matrix was made containing the distances between all species according to Equation 1 (Korbel et al. 2002).

Equation 1.

$$dist(A,B) = 1 - \frac{shared\_OGs(A,B)}{(\sqrt{2} \cdot size\_A \cdot size\_B) / (\sqrt{(size\_A^2 + size\_B^2)})}$$

As larger genomes can share more genes, we normalize the number of shared OGs by dividing by the weighted average genome size (see Equation 1 (Korbel et al. 2002)), where the genome size is defined as the number of considered OGs in the genome. Other approaches for normalization such as division by the smallest of the two genomes, or by the geometric average of the genome sizes, show an inferior fit to the relation between genome size and the number of shared genes (not shown). The distance is calculated by subtracting the resulting similarity fraction from 1 (see Equation 1).

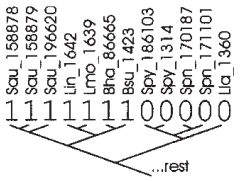

Figure 5. Example of an OG profile that shows its presence in 7 species (the OG is absent from the rest of the 89 species; not displayed). The profile covers the maximum number of subtrees in this phylogeny (i.e., 7 leaf nodes +6 internal nodes = 13). The species abbreviations are explained in the legend to Figure 7.

Iterative Removal of Phylogenetically Discordant Signals (PDSs)

The idea of the iterative method is to compare the presence/absence profile of every OG to the phylogeny to determine to what extent it can be considered discordant. Those OGs that are discordant according to a certain threshold are then removed, and a new phylogeny is inferred from the remaining profiles. This means that we need a first instance phylogeny to identify the first PDSs and start the iterations. In the standard runs, this was done by using the distance matrix calculated from all the OGs to construct a first instance neighbor- joining tree (Saitou and Nei 1987) using Neighbor (Felsenstein 1989). We also started 100 runs from randomized initial topologies.

The profile of every OG was then compared to the tree to determine to what extent its distribution was monophyletic. To do so, we counted the number of subtrees in which all the leaves contain the OG in question (i.e., all species in this partition have a 1 in the presence/absence profile; see Figure 5). The number of completely covered subtrees was used to calculate a score for how monophyletic the distribution is. For a given number of species, the score lay between the average coverage of 1000 randomly generated profiles in the first instance neighbor-joining tree (lower bound, set to 0) and the maximum number of partitions possibly covered by this number of species (upper bound, set to 1). The maximum number of tree partitions covered by a profile depends on the number of species in which the OG is present, according to Equation 2. Equation 2 is based on the number of partitions in a rooted tree, as every bipartition defines a rooted subtree in the entire phylogeny (cf. Figure 5).

Equation 2.

$$max\_covered\_partitions = 2 \cdot species - 1$$

The resulting coverage score, which can be compared between OGs present in any number of species, allows us to choose a threshold. OGs that scored below this threshold were removed from the data set, and a new distance matrix and neighbor-joining tree were computed based on the remaining profiles. For every threshold score, this procedure was iterated until convergence was established. Note that convergence to a limit cycle of phylogenies is possible as OGs are allowed to return to the data set if, in the new tree, their profile does cover sufficient branches. After each convergence, we increased the threshold in a simulated annealing-like approach (Kirkpatrick et al. 1983). In the work presented here, we chose 10 annealing steps of 0.1 each (the horizontal lines in Figure 6). Taking smaller annealing steps (e.g., 50 steps of 0.02 each) did not result in different phylogenies (not shown).

Weighting Method

As an alternative to the above method based on counting subtrees that share a gene, we employed a method that was originally developed to address the sequence weighting problem in amino acid multiple sequence alignments. The Rind program (Bruno 1996) uses a simple maximum likelihood model to estimate the frequency of characters on the tree and corrects for phylogenetic correlations. The Rind frequency gives an estimate of the number of times a character appeared de novo in evolution (Bruno 1996). If a character appears throughout a clade consisting of short branches, it is

assiged a lower frequency than a gene that appears throughout a clade of the same number of taxa but is made of long branches. If the monophyly of a clade is disrupted by taxon with an inconsistent character, this will have a smaller effect if the branch length of that taxon is longer.
As the presence/absence profiles of the OGs in all species can be seen as a multiple sequence alignment, we were able to run the Rind program on these binary sequences. Genes or columns that have a low Rind frequency, but are relatively abundant according to the raw data, are very clade specific. Thus, to get a score for the monophyly of each character, we divided the raw gene frequency by the Rind frequency. We scaled these scores so that the lowest received a weight of 0 and the highest got a weight of 1, and inferred a neighbor-joining tree as explained above, using the scores to assign weights to each OG.

Assessing Tree Quality

To determine how well the distances in the distance matrix were represented in the neighbor-joining tree, a new distance matrix was derived from the tree, by measuring the distances along the branches between all species pairs. The total difference between all the corresponding values in the two distance matrices was calculated and is expressed as a fraction of the average total distance in the trees. This gives a measure for how well the neighbor-joining tree represents the distance matrix.
We assessed the reliability of the genome tree by counting how often of the partitions occurred in 100 phylogenies constructed by resampling 100% of the OGs with replacement (bootstrapping).

Reference Trees

For reference, we used a SSU rRNA tree and an (unresolved) reference phylogeny from the NCBI taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy (Wheeler et al. 2000)). The rRNA tree is based on a database of expert aligned SSU rRNA sequences of all the species present in the current investigations (www.rna.icmb.utexas.edu (Cannone et al. 2002)). If the correct species was not available, a SSU rRNA sequence from a closely related organism was chosen; if multiple sequences per species were available, the longest and most reliable was selected. We used Clustal to construct a simple neighbor-joining tree based on this alignment (Thompson et al. 1994).

Results

The Choice for a Distance-Based Phylogeny

At first glance, the genome size effect and the concomitant parallel loss of genes should be represented by the Dollo parsimony (Farris 1977). This method is based on the idea that in evolution it is harder to gain a complex feature than to lose it, and we assert that a gene or orthologous group (OG) is such a complex feature that can only be independently gained by horizontal gene transfer. In a given phylogenetic tree, the Dollo algorithm explains the distribution of a character by allowing one origin (i.e., a change from 0 to 1) and as many reversions (1 to 0) as are necessary to explain the pattern of states seen. It then searches for the tree that minimizes the number of 1-to-0 reversions. Although this approach performs slightly better than standard parsimony (not shown), the resulting phylogeny still contains many errors including the clustering of small genomes. Likewise a maximum likelihood approach, such as implemented in MrBayes (Huelsenbeck and Ronquist 2001), in which the presence/absence of OGs was treated as the presence/absence of phenotypic characteristics, did not result in the clustering of the small parasitic genomes with their close relatives with large genomes.

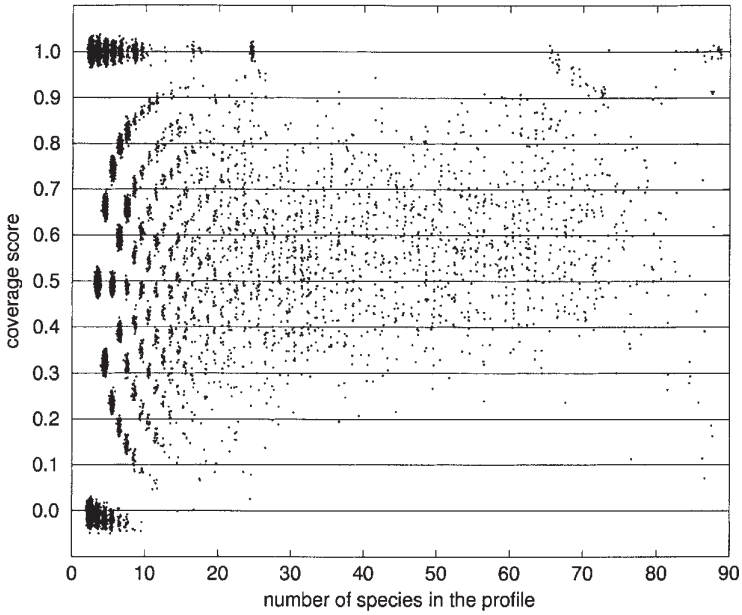

Figure 6. The coverage score of all the OGs in the first iteration neighbor-joining tree is plotted against the total number of species that contain this OG (number of species in the profile). Note that the coordinates have been scattered (by adding a random number from a normal distribution) to get better insight into the density. The horizontal lines are the simulated annealing steps going from the average of the random distribution (score 0; bottom line) to the maximum possible number of completely covered partitions (score 1).

In general, the main caveat of off-the-shelf parsimony or maximum likelihood methods is that they treat the evolution of each character independently. A model for genome evolution has to take variations in the number of genes present in a genome explicitly into account, as has been done for distance-based gene content phylogenies (Korbel et al. 2002). It is not the aim of this work to build a model but rather to develop methods to identify and filter out phylogenetic noise based on the presence/absence pattern of genes, and for that distance-based gene content phylogenies suffice.

The Signal in Gene Content

Phylogenetic Signal in Most of the OGs

Prior to iterating, we establish which OGs behave discordantly, based on the genome tree of the complete data set. This comparison already reveals how well the tree represents the data, as most of the OGs (13,375 of 19,433 = 69%; see Table 1) have a presence/absence profile that is (to a certain extent) consistent with the initial phylogeny based on all the OGs. Profiles that are present in only a few species are more likely to be either perfect or worse than random; the 31% of the OGs that had a negative coverage score contained an average of only 2.8 species. The rest of the OGs cover more subtrees than random profiles would and have a positive coverage score (see Figure 6). All these profiles are consistent with the genome tree of the complete data set (lowest simulated annealing threshold). No fewer than 5320 of the OGs (27%) are even completely in accordance with the first iteration neighbor-joining tree (they have a coverage score of 1). Many of these "perfect" OGs are present in the same few species; e.g., large groups of over 300 OGs with the same profiles occur between two species like the Ascomycota (352), the Cyanobacteria (341), the Methanosarcinales (364), the Sulfolobaceae (335), or the Xanthomonadaceae (305), but also between a three-species Mammalia–Drosophila group (579), and between the four Metazoa (890) included in this data set. All the OGs in these large groups are nonsupervised orthologous groups (NOGs) from the extended data set of von Mering et al. (2003). The other 2154 "perfect" OGs are distributed over only 90 different

profiles, among which there are profiles specific for groups such as the 16 Archaea, the 24 Archaea (16) plus Eukaryotes (8), and the 8 Alphaproteobacteria. Some of the larger groups can be seen as clusters on the line y = 1 in Figure 6.

Results from Iterations

The first instance tree (Figure 7) is already quite similar to the SSU rRNA reference phylogeny and the NCBI taxonomy (see Table 1). This confirms that gene content contains a strong phylogenetic signal (Fitz-Gibbon and House 1999; Snel et al. 1999; Tekaia et al. 1999). Throughout the iterations, this signal is shown be persistent in the evolving genome tree, and the phylogeny inferred from the restricted gene repertoire even improves. The improvements with respect to the reference trees (see Table 1; SSU rRNA, column 5; and NCBI taxonomy, column 6) are only minor, largely because the first instance tree already shows a considerable resemblance. As the threshold increases (column 2), the tree is based on a decreasing fraction of the OG profiles (column 3), which are selected to cover a maximum number of subtrees. At a certain point, the threshold becomes too high, and we start to exclude false negatives. The phylogeny then breaks down because too many OGs are removed that contain a phylogenetic signal.

Table 1. Statistics on the evolving phylogeny under a scheme that eliminates discordant OGs from the data set with an increasing stringency. Note: Similarity to the rRNA and NCBI reference trees reaches a maximum after seven simulated annealing steps. The simulated annealing threshold in the iterations is given in column 2. Note that the simulated annealing threshold scores are raised after convergence of the topology, so more topologies may be visited in a single threshold score step. The number of OG profiles used to construct each tree is given in column 3. The average scores of the OG profiles used to reconstruct the phylogeny are given in column 4. The fraction of branches shared with the SSU rRNA reference tree and in the taxonomy from NCBI is shown in columns 5 and 6 (the value for the unresolved NCBI taxonomy is higher because it contains fewer branches and will automatically share a larger fraction of its partitions). The difference between the distance matrix and the neighbor-joining tree is shown in column 7, and column 8 contains the average bootstrap value of all the partitions. The topology shifts above the boald line are shown in detail in Figure 7.

topology number	score threshold	OGs	average score	branches, rRNA	branches, NCBI	matrix vs. NJ tree	average bootstrap
0	0.0	19,433	0.494	0.628	0.818	0.237	0.881
	0.1	13,375	0.720			0.272	0.868
1	0.2	13,350	0.721	0.628	0.818	0.273	0.880
	0.3	13,152	0.729			0.278	0.880
		12,769	0.745			0.283	0.878
2	0.4	12,777	0.744	0.616	0.818	0.283	0.888
	0.5	11,737	0.780			0.302	0.871
3		9,239	0.856	0.640	0.800	0.338	0.877
		9,379	0.863			0.339	0.849
4	0.6	8,449	0.896	0.640	0.818	0.362	0.855
5		8,428	0.898	0.651	0.818	0.364	0.864
		8,468	0.897			0.361	0.827
6	0.7	7,046	0.946	0.651	0.818	0.388	0.819
7		7,074	0.945	0.651	0.818	0.396	0.829
	0.8	7,081	0.945			0.397	0.759
8		5,930	0.980	0.628	0.818	0.429	0.798
	0.9	5,975	0.981			0.430	0.627
9		5,651	0.987	0.581	0.727	0.452	0.645
		5,644	0.989			0.449	0.568
10	1.0	5,564	0.990	0.570	0.709	0.421	0.563
11		5,563	0.990	0.570	0.709	0.421	0.573
12		5,565	0.990	0.570	0.709	0.419	0.569

This breakdown is evident in Table 1: the difference between the matrix and the neighbor-joining tree increases, and after topology number 7, the overlap with the reference trees shows a sharp drop. Topology number 8 decreases the average bootstrap value of the partitions from 80 to 63%. To illustrate the types of changes that occur in the evolving tree, Figure 7 shows the shifts leading from the initial phylogeny to phylogeny contains almost 82% of the branches of topology number 7. Up to the point of this break-the (unresolved) NCBI taxonomy and just over 65% down, where

64% of the discordant OGs were ex-of the branches of the SSU rRNA tree. This result cluded, the reconstructed phylogenies change little shows that the phylogenetic signal in gene content, as and remain very close to the reference trees. The present in the first instance tree, is the dominant signal. More importantly, the shifts in the tree do not specifically affect organisms with shared phenotypic characters, e.g., parasites or hyperthermophilic species. As we do not see the effect of phenotype in the tree, such phenotypic convergence does not appear to be the cause or the result of large, systematic biases in the horizontal transfers.

Random Initializations

To investigate whether the quality of the reconstructed gene content trees throughout the iterations depended on the good first instance phylogeny, we repeated the experiments, starting from random initial topologies. The 100 random initial topologies, though completely different (they shared an average of 1% of their branches), rapidly converged. Based on the random first instance phylogenies, an average of 90% of the OGs was deleted. The second iteration phylogenies, composed of those OGs that were not discordant in the random initial trees (lowest simulated annealing threshold), already shared an average of 70% of their branches. The rapid convergence of the topology over the iterations illustrates how consistently this single phylogenetic signal is present in the gene repertoire data. To analyze the topological paths the phylogenies took after these random initializations, we looked in more detail at those trees with the highest resemblance to the reference phylogenies (the rRNA tree and the NCBI taxonomy) and to a selected topology from the standard initialization (topology 7; cf. Table 1). Of the 100 random initial topologies, a large group of 68 paths converged to one topology, which shared 97% of its branches with the standard initialization. Abundant though this topology was, it contained some improbable shifts compared to the phylogeny from the standard procedure. The position of *Halobacterium* was closer to the archaeal root, and *Thermotoga* was placed next to *Thermoanaerobacter* rather than at the root of the low-G+C Gram-positives. Seven of the paths converged exactly to the topology of the standard genome tree. The other 25 paths converged to six other phylogeny, sharing an average of 94% of the partitions with the phylogeny from the standard initialization.

A Worst-Case Scenario

An often-discussed case of "massive" horizontal gene transfer is that from the Archeae to the hyperthermophilic Bacteria *Aquifex aeolicus* and *Thermotoga maritima* (Aravind et al. 1998). We tested whether starting our iterations with an edited phylogeny, in which we grouped *A. aeolicus* and *T. maritima* at the root of the Archaea, would result in the selection of those horizontally transferred genes, and a convergence of the tree to one in which the hyperthermophilic Bacteria would cluster with the Archaea. In the first iteration the tree converged to the same tree as the one that was started with the unedited tree. This illustrates the point that there may be cases of large-scale horizontal between some species, but their signal is not strong enough to cause systematic biases in the tree based on gene content, even when biasing the selection of genes for the phylogeny toward a set involved in horizontal transfer.

Weighted Tree

The tree obtained from weighing fast-evolving positions is very similar to the rRNA phylogeny (Figure 8) and successfully improves relative to the unweighted first instance tree. The tree shares over 85% of the branches with the (unresolved) NCBI taxonomy and just over 65% of the branches with the SSU rRNA reference tree. The weighting procedure reinforces especially strongly the separation into three kingdoms, as the internal branches separating the three kingdoms have become longer. When comparing the genome phylogenies that result from the two approaches for filtering the OGs in detail, it becomes apparent that the iterative removal method has a bigger impact on the topology of the tree. As the threshold increases, there are many more topological shifts than in the weighting method. Topology number 3 looks most like the weighted tree (they share 90% of the branches), and

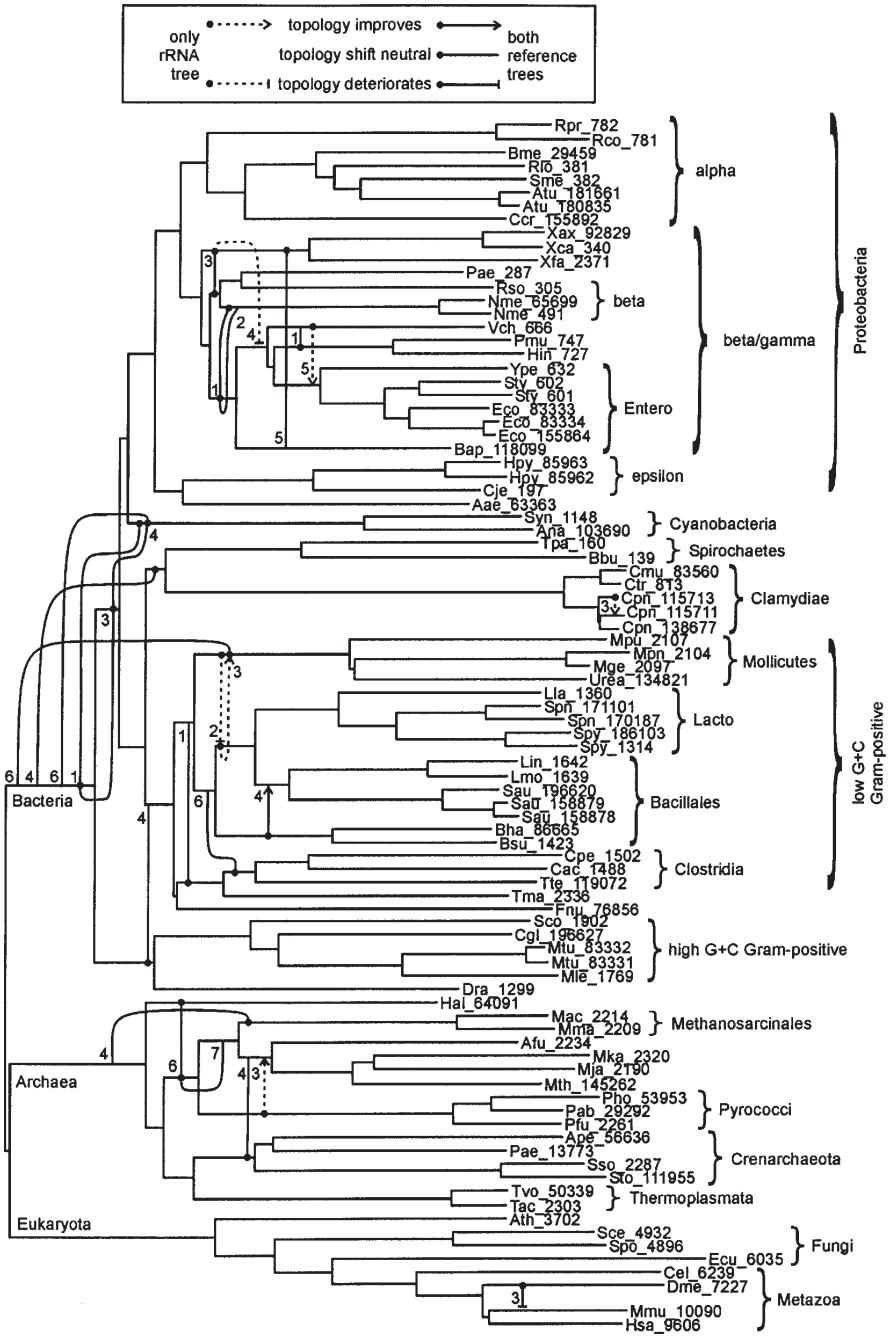

Figure 7. Initial phylogeny inferred from all the gene presence/ absence profiles. The branches that shift in the tree during the iterations are indicated. The number given is the topology number (cf. Table 1) at which the shift occurs. Note that there are no shifts that improve or deteriorate the topology relative to the NCBI taxonomy alone: because the NCBI taxonomy is not completely resolved, changes relative to this tree always also change the fraction of branches shared with the SSU rRNA tree. The phylogeny we inferred was unrooted; we chose to display the Archaea and the Eukaryota as sister taxa, as that is the most commonly accepted view (though there are many papers from the Philippe group to combat this e.g., (Philippe and Forterre 1999)). There is no legend for the branch lengths, as they are only informative in the first instance tree (in two cases they have been very slightly altered to fit the arrows indicating the shifts into the figure). The species abbreviations are the first letter of the family name and the first two letters of the species name, followed by the taxonomic identifier (in alphabetical order): Aae_63363, *Aquifex aeolicus*; Afu_2234, *Archaeoglobus*

all topologies that follow move farther away from the initial phylogeny. Nonetheless, both methods accomplish comparable improvements relative to the unfiltered tree. The advantage of the weighted tree relative to the iterated tree is that the former does not require arguably subjective criteria, like the breakdown of the bootstrap values, to determine when to stop increasing the score threshold.

Phylogenetic Implications

Shifts in the Archaea. In the archaeal phylogeny, the Crenarchaeota remain monophyletic, but they appear to be derived from the Euryarchaeota, making the Euryarchaeota a paraphyletic taxon. This is inconsistent with the rRNA tree but often found in genome trees (Wolf et al. 2002). The current approach does manage to shift Halobacterium away from its (erroneous) ancestral position, into the Euryarchaeota. Instead, the Methanosarcinales move to the archaeal root, next to Halobacterium, followed later by the Thermoplasmata. Cavalier-Smith already proposed to join the Methanosarcinales and the Halobacteria in the phylum Halomebacteria. This was based on the fact that many of the differences between the two can be attributed to the loss of ancestral proteins by the Methanobacteria, whereas many similarities in RNA polymerases, antibiotic sensitivities, and rRNAs can be found (Cavalier-Smith 1986; Cavalier-Smith 2002). Slesarev and co-workers (2002) showed that genome trees based on gene content or on conserved gene pairs group all methanogenic Archaea. Indeed, this is true for the methanogens sequenced at the time of that research, but we show here that the Methanosarcinales are not part of the otherwise strongly supported methanogenic subtree. As in the gene content tree presented by (Slesarev et al. 2002), we find the position of Archaeoglobus fulgidus to be stable at the root of the methanogens.

Hyperthermophilic Bacteria

Gene content phylogenies are especially interesting for those clades where rRNA trees might fail. The phylogenetic position of thermophilic Bacteria is such a point (Cavalier-Smith 2002). The inference that the thermophilic Bacteria are primitive, based on rRNA trees, has been doubted because this placement might be an artifact from long-branch attraction (Gribaldo and Philippe 2002) and

fulgidus; Ana_103690, *Anabaena* sp.; Ape_56636, *Aeropyrum pernix*; Ath_3702, *Arabidopsis thaliana*; Atu_181661, *Agrobacterium tumefaciens* C58/ATCC 33970 (Cereon); Atu_180835, *Agrobacterium tumefaciens* C58/ATCC 33970 (U. Washington); Bap_118099, *Buchnera aphidicola*; Bbu_139, *Borrelia burgdorferi*; Bha_86665, *Bacillus halodurans*; Bme_29459, *Brucella melitensis*; Bsu_1423, *Bacillus subtilis*; Cac_1488, *Clostridium acetobutylicum*; Ccr_155892, *Caulobacter crescentus*; Cel_6239, *Caenorhabditis elegans*; Cgl_196627, *Corynebacterium glutamicum*; Cje_197, *Campylobacter jejuni*; Cmu_83560, *Chlamydia muridarum*; Cpe_1502, *Clostridium perfringens*; Cpn_115711, *Chlamydia pneumoniae* AR37; Cpn_115713, *Chlamydia pneumoniae* CWL029; Cpn_138677, *Chlamydia pneumoniae* J138; Ctr_813, *Chlamydia trachomatis*; Dme_7227, *Drosophila melanogaster*; Dra_1299, *Deinococcus radiodurans*; Eco_155864, *Escherichia coli* O157:H7 EDL933; Eco_83333, *Escherichia coli* K-12MG1655; Eco_83334, *Escherichia coli* O157:H7 substr. RIMD 0509952; Ecu_6035, *Encephalitozoon cuniculi*; Fnu_76856, *Fusobacterium nucleatum*; Hal_64091, *Halobacterium* sp.; Hin_727, *Haemophilus influenzae*; Hpy_85962, *Helicobacter pylori* 26695; Hpy_85963, *Helicobacter pylori* J99; Hsa_9606, *Homo sapiens*; Lin_1642, *Listeria innocua*; Lla_1360, *Lactococcus lactis* subsp. lactis; Lmo_1639, *Listeria monocytogenes*; Mac_2214, *Methanosarcina acetivorans*; Mge_2097, *Mycoplasma genitalium*; Mja_2190, *Methanococcus jannaschii*; Mka_2320, *Methanopyrus kandleri*; Mle_1769, *Mycobacterium leprae*; Mma_2209, *Methanosarcina mazei*; Mmu_10090, *Mus musculus*; Mpn_2104, *Mycoplasma pneumoniae*; Mpu_2107, *Mycoplasma pulmonis*; Mth_145262, *Methanobacterium thermoautotrophicum*; Mtu_83331, *Mycobacterium tuberculosis* CDC1551; Mtu_83332, *Mycobacterium tuberculosis* H37Rv; Nme_491 *Neisseria meningitidis*; Nme_65699, *Neisseria meningitidis*; Pab_29292, *Pyrococcus abyssi*; Pae_287, *Pseudomonas aeruginosa*; Pae_13773, *Pyrobaculum aerophilum*; Pfu_2261, *Pyrococcus furiosus*; Pho_53953, *Pyrococcus horikoshii*; Pmu_747, *Pasteurella multocida*; Rco_781, *Rickettsia conorii*; Rlo_381, *Rhizobium loti*; Rme_382, *Rhizobium meliloti*; Rpr_782, *Rickettsia prowazekii*; Rso_305, *Ralstonia solanacearum*; Sau_158878, *Staphylococcus aureus* subsp. aureus Mu50; Sau_158879, *Staphylococcus aureus* subsp. aureus N315; Sau_196620, *Staphylococcus aureus* subsp. aureus MW2; Sce_4932, *Saccharomyces cerevisiae*; Sco_1902, *Streptomyces coelicolor*; Spn_170187, *Streptococcus pneumoniae* TIGR4; Spn_171101, *Streptococcus pneumoniae* R6; Spo_4896, *Schizosaccharomyces pombe*; Spy_1314, *Streptococcus pyogenes*; Spy_186103, *Streptococcus pyogenes*; Sso_2287, *Sulfolobus solfataricus*; Sto_111955, *Sulfolobus tokodaii*; Sty_601, *Salmonella typhi*; Sty_602, *Salmonella typhimurium*; Syn_1148, *Synechocystis* sp.; Tac_2303, *Thermoplasma acidophilum*; Tma_2336, *Thermotoga maritima*; Tpa_160, *Treponema pallidum*; Tte_119072, *Thermoanaerobacter tengcongensis*; Tvo_50339, *Thermoplasma volcanium*; Upa_134821, *Ureaplasma parvum*; Vch_666, *Vibrio cholerae*; Xax_92829, *Xanthomonas axonopodis*; Xca_340, *Xanthomonas campestris*; Xfa2371, *Xylella fastidiosa*; and Ype_632, *Yersinia pestis*. The strain is not specified unless more instances of the same species make this necessary.

selection for high G+C content in hyperthermophilic rRNA (Galtier and Lobry 1997). Recently it has indeed been shown that this artifact can be circumvented by considering only the slowly evolving nucleotides in the rRNA sequence. This places the hyperthermophilic Bacteria as a division whose

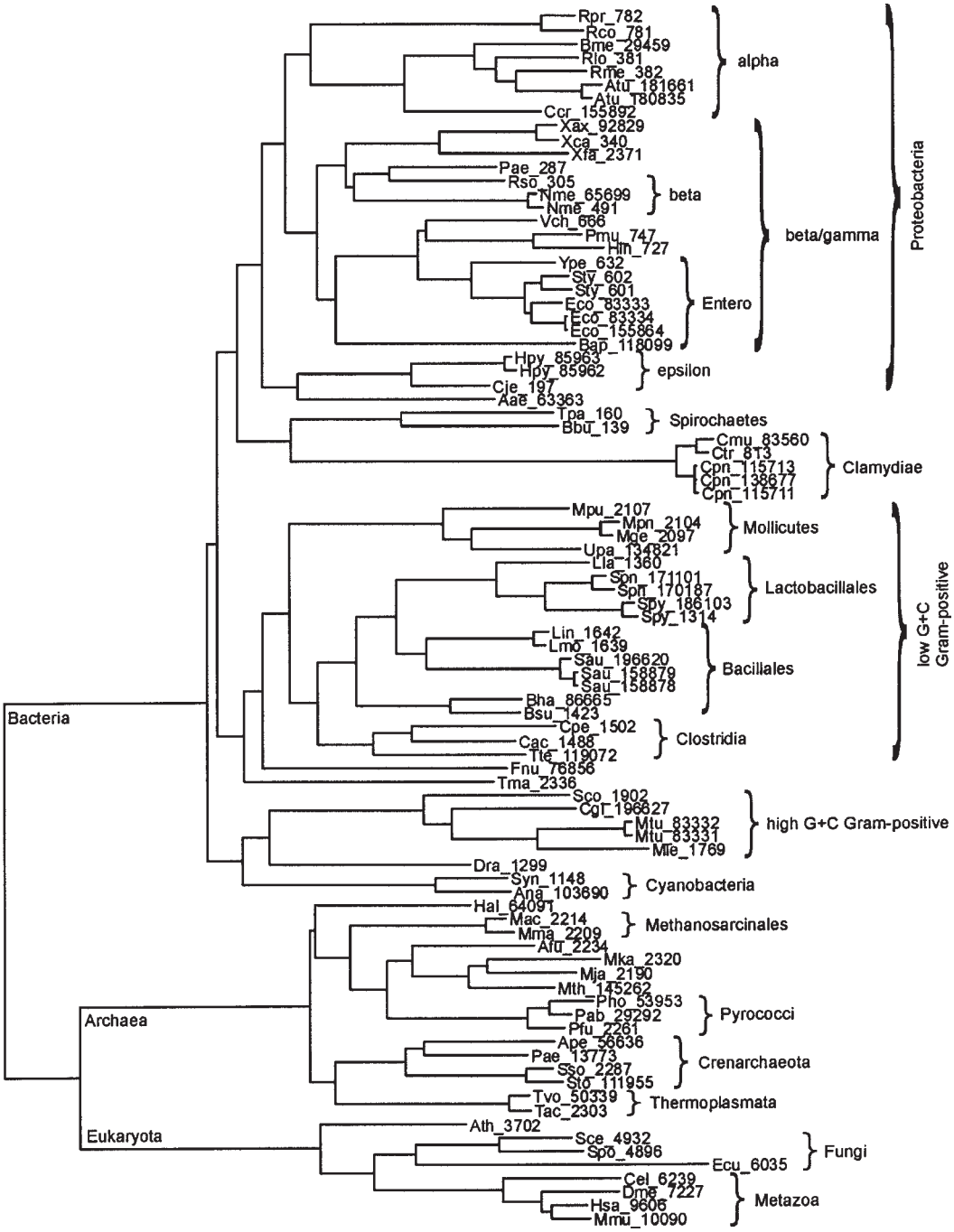

Figure 8. Phylogeny inferred from gene presence/absence profiles with weighted characters. The species abbreviations are explained in the legend to Figure 7.

relation to other divisions remains unclear (Brochier and Philippe 2002). Interestingly our results indicate a consistent (i.e., throughout the iterations) affiliation of Thermotoga with the Firmicutes and of Aquifex with the Delta- / Epsilonproteobacteria. They stay there even after removal of possible phylogenetically discordant signals, such as the abundant horizontal transfers of these species with the Archaea (Nelson et al. 1999).

The hypothesis of independent origins of eubacterial (hyper)thermophily finds strong support in the work of Forterre et al. (2000). They show that reverse gyrase, an enzyme that is crucial for stabilizing the DNA in hyperthermophilic organisms, in Aquifex and in Thermotoga was independently obtained by two seperate horizontal transfer events. Our iterative approach discards reverse gyrase (COG1110) as a discordant signal at a threshold score of 0.5 and in the weighting approach it is assigned a weight of 0.16. When combined with results from other genome treelike approaches and other independent evidence, the position of Aquifex with the Proteobacteria, as well as Thermotoga with the Gram-positive Bacteria, is supported.

Structurally, the outer membrane of Aquifex has been shown to contain lipopolysaccharide (Plotz et al. 2000), like the Proteobacteria, but unlike Gram-pos itive Bacteria. (Klenk et al. 1999) made phylogenetic analyses of the two largest subunits of bacterial RNA polymerases and placed Aquifex with the Proteobacteria. This position is also supported by a supertree composed of the phylogenies of hundreds of orthologous gene families (Daubin et al. 2001), gene content trees (Wolf et al. 2001), and gene order trees (Wolf et al. 2001) Analysis from rare genomic events, such as conserved insertions and deletions in several proteins, also shows that Aquifex should be placed next to the Proteobacteria (Gupta and Griffiths 2002).

The position of Thermotoga as an evolutionary neighbor to the Gram-positives is supported by the same insertions and deletions study (Gupta and Griffiths 2002). Both (Tiboni et al. 1993) and (Pesole et al. 1995) show that glutamine synthetase I trees group Thermotoga with the low-G+C Gram-positive Bacteria. (Gribaldo et al. 1999) show a deletion in the sequence of HSP70, shared by Thermotoga and the Gram-positive Bacteria, and though the phylogenies inferred from the protein sequence do not cluster these groups, this may be artifactual and the result of convergence within the hyperthermophilic sequences (Cambillau and Claverie 2000; Kreil and Ouzounis 2001; Suhre and Claverie 2003).

Problems

In the iterative method, the eukaryotic subtree is very stable. The only topological change is for the worse: Drosophila melanogaster is placed in between the mammals (see Figure 7). This results from an artifact of the definition of NOGs by von Mering et al. (2003), which unites the mammals to form a single clade, thus disallowing the formation of any NOGs shared only by these two species. The weighting method does not show this shift.

Two groups of Bacteria with exceptionally small genomes are pushed to the root during the iterations. Though we have corrected for genome size in Equation 1, the Chlamydiae/Spirochaetes group and the Mollicutes are still problematic cases, though more so in the iterative than in the weighting approach. This size effect is the result of the fact that small genomes can share only a certain maximum number of genes. This is a known problem in gene content phylogenetics (Wolf et al. 2002), and though it has been addressed (Korbel et al. 2002), a definitive solution has still not been found.

Which Genes Are Discordant?

It has been proposed that metabolic genes undergo more horizontal gene transfer than informational genes. Here we obtain detailed information on this hypothesis, by determining which types of genes were discordant, i.e., which OGs were filtered out in our procedures. To summarize this, we look at the COG functional classes (NOGs are not functionally classified). Figure 9 shows the extent to which the different COG functional categories were allowed to remain in the data set. For the weighting method, the average assigned weight of all the genes in the functional category is plotted. For the

iterative removal method, the fraction of genes that remained in the data set of topology number 7 is indicated. This topology, where the coverage score threshold of 0.8 excluded 64% of the OGs, was selected to maintain consistency with Figure 7.

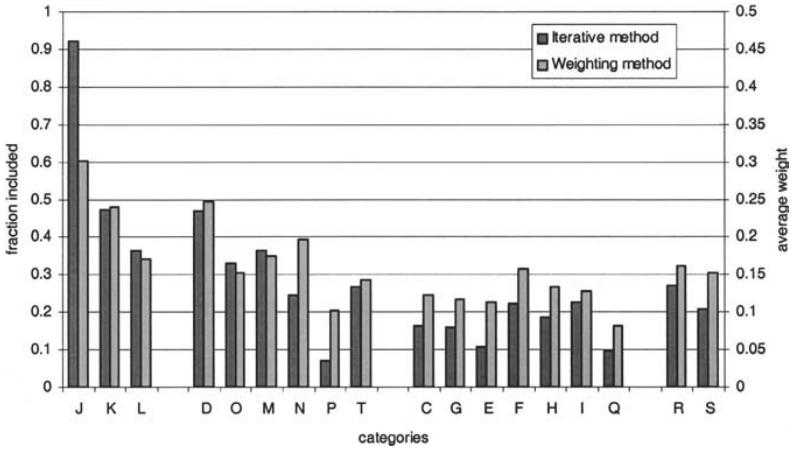

Figure 9. Functional categories of the contributing COGs for each of the two methods. The dark gray bars (left) are the fractions that were not removed in the seventh topology (coverage score, 0.8; cf. Table 1) of the iterative method. The light gray bars (right) are the average weights assigned in the weighting method. Both methods identify the same functional categories as discordant. The categories are grouped in the four main COG classes "information storage and processing" (translation, ribosomal structure, and biogenesis [J], transcription [K], and DNA replication, recombination, and repair [L]), "cellular processes" (cell division and chromosome partitioning [D], posttranslational modification, protein turnover, chaperones [O], cell envelope biogenesis, outer membrane [M], cell motility and secretion [N], inorganic ion transport and metabolism [P], and signal transduction mechanisms [T]), "metabolism" (energy production and conversion [C], carbohydrate transport and metabolism [G], amino acid transport and metabolism [E], nucleotide transport and metabolism [F], coenzyme metabolism [H], lipid metabolism [I], and secondary metabolite biosynthesis, transport, and catabolism [Q]), and "poorly characterized" (general function prediction only [R] and function unknown [S]) (Tatusov et al. 1997).

The results for both schemes investigated in this research are remarkably similar. The "translation, ribosomal structure, and biogenesis" category (J) is the least discordant; less than 8% of the COGs from this category are removed in the iterative procedure, and the average weight assigned to the COGs in this category was 0.30. Of the "inorganic ion transport and metabolism" category (P), less than 8% remained in the data set, and it can be considered the fastest-evolving category of genes with respect to gene content. The category where the lowest weights were assigned was "secondary metabolite biosynthesis, transport, and catabolism" (Q), where the COGs received an average weight of 0.08. In general, "metabolism" COGs are filtered out most in our procedure, whereas "information storage and processing" COGs are relatively stable in evolution. This supports the complexity hypothesis (Jain et al. 1999) that, generally, operational genes are transferred more readily throughout evolution than informational genes, which are more often involved in complex networks of interactions. However, our study reveals a more detailed picture: apart from the inorganic ion transport and metabolism category (P), the other "cellular processes" categories, such as "cell division and chromosome partitioning" (D) and "cell envelope biogenesis, outer membrane" (M), are intermediately discordant with the COGs from the "metabolism" and "information storage and processing" classes.

Discussion

If we correct for genome size, a very good gene content phylogeny, subject to some caveats, can already be inferred. This means that the noise, which results from processes like horizontal transfer or convergence through parallel gene loss and may confound a genome phylogeny, can be effectively

averaged out by considering genome scale data. In this initial tree, improvements can be made by reducing the impact of the noise, which is shown in the current paper using two independent approaches. This result is in contrast with Clarke et al. (2002), who did not find any improvements in their tree when filtering for discordant genes. This is probably due to the fact that the filtering scheme used by these authors is not strong enough. The topological improvement of their phylogeny may also be restrained by their choice to use one source (orthology) for the reconstruction of the tree and another, albeit related, source (sequence) for filtering. Here we show that the topology will change during the iterative removal of the noise, as well as in a scheme that selectively downweighs the noise. This is not to say that the genes designated as noise are biologically irrelevant. Genes that have a nonphylogenetic distribution often have functional significance, such as shared pathogenicity factors between *Helicobacter pylori* and *Haemophilus influenzae* (Huynen et al. 1998) or reverse gyrase in the hyperthermophiles (Forterre et al. 2000). But these qualitative, phenetic, patterns in shared gene content apparently play a quantitatively minor role relative to the phylogenetic signal and can be considered noise when constructing genome trees.

In the iterative procedure, we have shown that there is a consistent phylogenetic signal in the majority of OGs: throughout the iterations, the phylogeny shows few changes until the fraction signal over noise that is removed becomes too high. This result is also supported by the converging trajectories starting from random initial phylogenies. Being too strict in removing discordant OGs leads to a breakdown of the phylogenetic pattern, leaving too little signal for a reliable tree topology. The phylogenetic signal is thus the only detectable signal in the gene content. The rest is noise. A recent investigation of the relation between horizontal transfer and phylogenetic incongruence in gene trees revealed that, in most cases, alternate topologies represent construction artifacts rather than the accumulation of horizontal transfer events with time (Daubin et al. 2003).

In the current paper, we have implemented two methods, based on the same ideas, and both give comparable results in terms of improvements in the phylogeny and in the types of functions that are considered discordant. Improvement for the current approaches may be achieved by the implementation of a better measure for the discordance of a signal in the phylogeny, but we do not expect major changes in the results given the similarity in outcome from the two procedures. The main improvements for both the iterative and the weighting method may be expected from a better, i.e., more fine-grained, definition of orthology, which will allow more detail and thus better-defined relationships between the species.

Other improvements might come from maximum likelihood or Bayesian approaches, which can include explicit statistical models of genome evolution. Full Bayesian methods are already available for gene/ species tree reconciliation (Arvestad et al. 2003). This specific development, and that of Bayesian inference in general, opens up several lines along which gene content phylogenies can be improved. First, their model of gene content evolution can be used for the likelihood of a species phylogeny, incorporating all genome sizes and the distribution of the OGs over the species. Second, a more complicated approach could be implemented that does not treat the OGs as a binary distribution but as a gene tree. This makes it possible to directly use the methodology from Arvestad et al. (2003), but with the extension that the species tree is one of the parameters that are to be determined using the likelihood algorithm. The biggest drawbacks are expected to be the computational time needed to construct reliable gene trees for all OGs, computing the likelihood for all the trees, and the great increase in computational time needed for the Monte Carlo Markov chain to simultaneously and sufficiently sample tree space.

Acknowledgments

This work was supported in part by European Union Contract QLTR-2000-01676 and by a grant from The Netherlands Organization for Scientific Research (NWO).

Assessment of phylogenomic and orthology approaches for phylogenetic inference

Bas E. Dutilh, Vera van Noort, René T.J.M. van der Heijden,
Teun Boekhout, Berend Snel and Martijn A. Huynen
Bioinformatics (2007) 23: 815-824

Abstract

Motivation

Phylogenomics integrates the vast amount of phylogenetic information contained in complete genome sequences, and is rapidly becoming the standard for inferring reliable species phylogenies. There are however fundamental differences between the ways in which phylogenomic approaches like gene content, superalignment, superdistance and supertree integrate the phylogenetic information from separate orthologous groups. Furthermore, they all depend on the method by which the orthologous groups are initially determined. Here, we systematically compare these four phylogenomic approaches, in parallel with three approaches for large-scale orthology determination: pairwise orthology, cluster orthology and tree-based orthology.

Results

Including various phylogenetic methods, we apply a total of 54 fully automated phylogenomic procedures to the Fungi, the eukaryotic clade with the largest number of sequenced genomes, for which we retrieved a golden standard phylogeny from the literature. Phylogenomic trees based on gene content show, relative to the other methods, a bias in the tree topology that parallels convergence in life style among the species compared, indicating convergence in gene content.

Conclusions

Complete genomes are no warrant for good, or even consistent phylogenies. However, the large amounts of data in genomes enable us to carefully select the data most suitable for phylogenomic inference. In terms of performance, the superalignment approach, combined with restrictive orthology, is the most successful in recovering a fungal phylogeny that agrees with current taxonomic views, and allows us to obtain a high resolution phylogeny. We provide solid support for what has grown to be common practice in phylogenomics during its advance in recent years.

Introduction

Phylogenomics, i.e., using entire genomes to infer a species tree, has become the de facto standard for reconstructing reliable phylogenies (Ciccarelli et al. 2006; Daubin et al. 2002). Whereas phylogenetic trees, i.e., based on single gene families, may show conflict (Teichmann and Mitchison 1999) due to a variety of causes, phylogenomic trees have held the promise that they can average out these anomalies by the sheer power of genome-scale data. As it is based on the maximum genetic information, a phylogenomic tree should be the best reflection of the evolutionary history of the species, assuming this history is tree-like (Doolittle 1999b; Ge et al. 2005). Although there are discordant processes at the level of gene repertoires, such as horizontal gene transfer (Doolittle 1999b) or differences in the rates of evolution and gene loss between paralogs in different species (Daubin et al. 2003), these have been shown to add noise rather than a directional bias (Dutilh et al. 2004). However, this does not mean that phylogenomics is the end of all conflict in species trees (Jeffroy et al. 2006): there are many ways to integrate the information from the different gene families to form a single species phylogeny.

Phylogenomics

In taxonomy, the term phylogenomics indicates the construction of a phylogeny on the basis of complete genome data. We can consider this type of phylogenomics as parallel phylogenetics over all gene families, combined with a synthesis step. This step from phylogenetics to phylogenomics integrates the phylogenetic information from the different gene families to form a single species phylogeny, and can be taken at successive levels in the process. As a guide line, we classify phylogenomic methods by the level where the step from phylogenetics to phylogenomics is made (Figure 10). Here, we compare these four qualitatively different phylogenomic approaches.

For sequence-based phylogenomic methods, the first step is to make multiple alignments for every orthologous group (OG) (Delsuc et al. 2005). In the superalignment approach, the phylogenetic information is then combined by concatenating the multiple alignments to form a superalignment. Subsequently, conventional phylogenetic inference methods can be used to transform the alignment into a phylogeny. Superdistance trees continue the path of phylogenetics by first calculating distance matrices for all gene families. The phylogenomic distance between two species is then defined as the average distance between all the shared gene families (Kunin et al. 2005). Finally, the supertree approach (Bininda-Emonds 2004; Daubin et al. 2002) takes the step from phylogenetics to phylogenomics at the very end. After phylogenetic trees have been composed for all gene families, an integration step combines the multiple gene family trees to form a single phylogenomic tree.

Of the methods based on whole-genome features (Delsuc et al. 2005) we only consider gene content here, as gene order in the Fungi evolves too fast to retain a phylogenetic signal (Huynen et al. 2001). Gene content takes the step from phylogenetics to phylogenomics right after the definition of the OGs (Figure 10). Species are regarded as "bags of genes", and sequence information is only used to determine the OGs. To infer a phylogenomic tree from gene content data, a binary character matrix indicating the presence or absence of the OGs in all species can be treated in the same way as a multiple sequence alignment.

Orthology

The initial step in every phylogenomic approach is to determine which genes are to be compared between species (top row in Figure 10). We compare the performance of three types of orthology definition: pairwise orthology, cluster orthology, and tree-based orthology. The first two methods use sequence similarity scores to define orthologous groups of genes. Pairwise orthology is defined between only two species (e.g. bi-directional best hits or Inparanoid (Remm et al. 2001)), and cluster orthology (e.g. Clusters of Orthologous Groups (Tatusov et al. 1997)) is the natural extension of pairwise orthology to more than two species. Tree-based orthology comes closest to the original phylogenetic definition of orthology (Fitch 1970). Rather than using only the sequence similarity

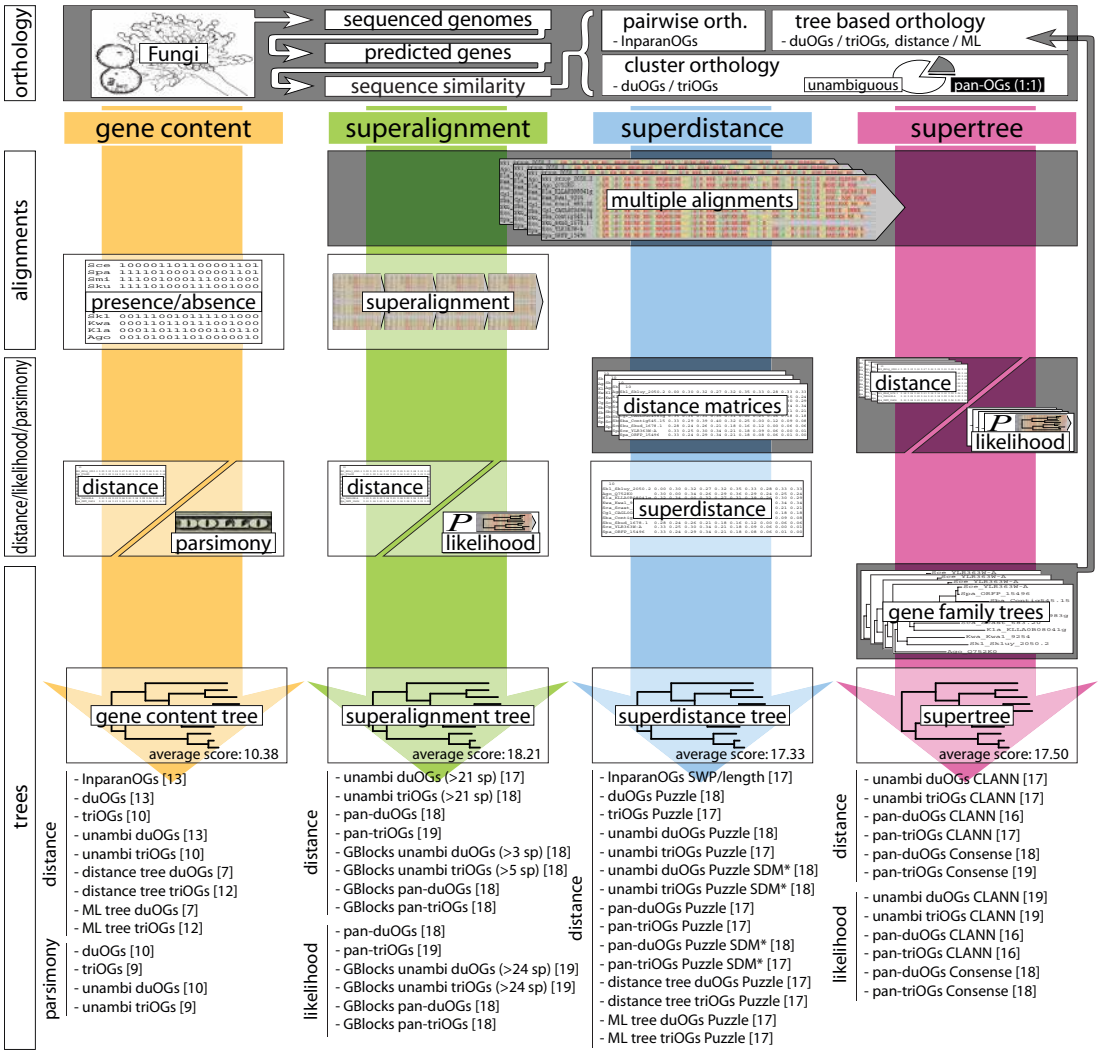

Figure 10. Making phylogenomic trees. Before starting tree inference, OGs are defined (top row). Phylogenomics follows the steps of phylogenetics, from multiple alignment through distance, likelihood or parsimony to the reconstruction of a phylogeny. Integrating separate phylogenetics for each gene family (gray boxes) to phylogenomics (white boxes) can be done at every one of these steps. This defines the phylogenomic approach: gene content (after OG definition), superalignment (after multiple alignment), superdistance (after distance calculation) or supertree (after reconstruction of gene family trees). The phylogenomic trees we reconstructed are listed at the bottom, the number between square brackets indicates the number of target nodes that the tree recovered correctly.

scores, it analyses a phylogenetic tree of a homologous group of genes to obtain orthologous relations (van der Heijden et al. submitted). Note that although tree-based orthology is an ideal approach to determine orthology at scalable levels of resolution, it needs to be operationalized: OGs have to be determined from the trees separately for each pair of species. The superalignment and supertree approaches, that consider a large set of species simultaneously, can not deal with pairwise orthology or operationalized tree-based orthology (see "Methods" and Appendix).

Fungal phylogeny

To compare the performance of phylogenomic approaches, some kind of golden standard phylogeny is imperative. We chose here to benchmark the phylogenomic methods using a phylogeny of real

species. The alternative, to work with simulated evolutionary data (Hillis et al. 1994), would require the simulation of the evolution of complete genomes for which we lack the models and parameters. Prima facie, an approach that uses a known phylogeny appears to exclude the possibility for any improvements. However, due to ambiguities in the literature our golden standard phylogeny is not completely resolved. We expect that properly derived complete genome phylogenies will allow a higher resolution both for the species analyzed here, and for other (partly) unresolved clades in future analyses.

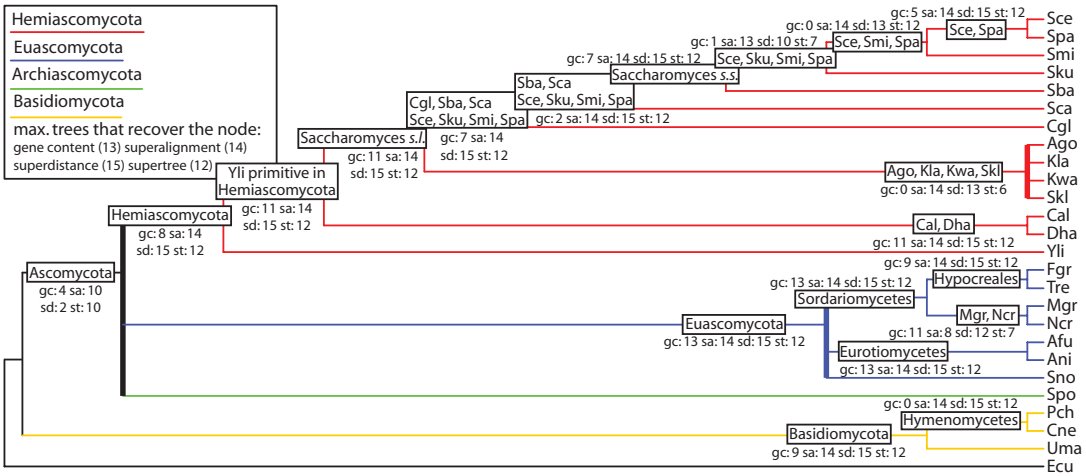

Figure 11. Target phylogeny. Labeled nodes are supported by literature. Unresolved issues are indicated by multifurcating nodes (bold lines). The numbers at every node indicate the number of the trees in each of the phylogenomic approaches that recovered this node correctly. See Table 11 for references that support this tree.

The Fungi are the eukaryotic clade with the most sequenced genomes. Saccharomyces cerevisiae has been a model organism for decades, and in this era of comparative genomics much work has focused on sequencing the genomes of more or less closely related species (Cliften et al. 2003; Dujon et al. 2004; Kellis et al. 2003). In total, 26 completely sequenced fungal genomes were available in public databases at the start of this study (September 2005): 22 Ascomycota, 3 Basidiomycota and the Microsporidium Encephalitozoon cuniculi (see Table 2 and Figure 11). We included *E. cuniculi* as an outgroup because this was the most closely related complete genome to the Fungi (Thomarat et al. 2004; Vivares et al. 2002), and Rhizopus oryzae was not available yet.

The fungal kingdom has been extensively studied by phylogeneticists. Traditional phenotypic methods (e.g. reviewed in (Guarro et al. 1999)), molecular phylogenetic analyses based on rRNA (Fell et al. 2000; Lopandic et al. 2005; Lutzoni et al. 2004; Scorzetti et al. 2002; Tehler et al. 2003) or small numbers of other proteins (Diezmann et al. 2004; James et al. 2006; Kouvelis et al. 2004; Kurtzman 2003), as well as some large scale studies (Jeffroy et al. 2006; Kuramae et al. 2006; Robbertse et al. 2006; Rokas et al. 2003; Thomarat et al. 2004) have helped resolve many of the phylogenetic relationships in the fungal kingdom. Based on the available literature (Berbee et al. 2000; Delsuc et al. 2005; Diezmann et al. 2004; Jeffroy et al. 2006; Jones et al. 2004; Kouvelis et al. 2004; Kuramae et al. 2006; Kurtzman 2003; Lopandic et al. 2005; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Robbertse et al. 2006; Tehler et al. 2003; Thomarat et al. 2004), we composed a true fungal phylogeny (Figure 11) that we use as a benchmark.

This study

Here, we compare the four phylogenomic and the three orthology approaches presented above (Figure 10) in parallel, assessing their ability to infer the 19 target nodes derived from the literature. As many different methods and algorithms exist for most of these approaches, we include several

implementations in order to buffer our findings from possible biases in the individual methods. Thus, we compose a total of 54 phylogenomic trees of the 26 complete fungal genomes, using completely automated methods.

Table 2. The organisms included in this research.

	species name	genes	reference
Ago	*Ashbya gossypii* (*Eremothecium*)	4,720	(Dietrich et al. 2004)
Afu	*Aspergillus fumigatus*	9,926	(Nierman et al. 2005)
Ani	*Aspergillus nidulans*	9,541	(Galagan et al. 2005)
Cal	*Candida albicans*	11,904	(Jones et al. 2004)
Cgl	*Candida glabrata*	5,272	(Dujon et al. 2004)
Cne	*Cryptococcus neoformans*	5,882	(Loftus et al. 2005)
Dha	*Debaryomyces hansenii*	6,896	(Dujon et al. 2004)
Ecu	*Encephalitozoon cuniculi*	1,918	(Katinka et al. 2001)
Fgr	*Fusarium graminearum*	11,640	(www.broad.mit.edu)
Kla	*Kluyveromyces lactis*	5,331	(Dujon et al. 2004)
Kwa	*Kluyveromyces waltii*	5,230	(Kellis et al. 2004)
Mgr	*Magnaporthe grisea*	11,109	(Dean et al. 2005)
Ncr	*Neurospora crassa*	10,620	(Galagan et al. 2003)
Pch	*Phanerochaete chrysosporium*	11,777	(Martinez et al. 2004)
Sba	*Saccharomyces bayanus*	4,966	(Kellis et al. 2003)
Sca	*Saccharomyces castellii*	4,690	(Cliften et al. 2003)
Sce	*Saccharomyces cerevisiae*	6,702	(Goffeau et al. 1996)
Skl	*Saccharomyces kluyveri*	2,992	(Cliften et al. 2003)
Sku	*Saccharomyces kudriavzevii*	3,813	(Cliften et al. 2003)
Smi	*Saccharomyces mikatae*	3,100	(Kellis et al. 2003)
Spa	*Saccharomyces paradoxus*	8,955	(Kellis et al. 2003)
Spo	*Schizosaccharomyces pombe*	4,990	(Wood et al. 2002)
Sno	*Stagonospora nodorum*	16,597	(www.broad.mit.edu)
Tre	*Trichoderma reesei*	9,997	(www.jgi.doe.gov)
Uma	*Ustilago maydis*	6,522	(Kamper et al. 2006)
Yli	*Yarrowia lipolytica*	6,666	(Dujon et al. 2004)

Methods

Orthology

Sequences were downloaded from the respective fungal sequencing projects (see Table 2). We compare the performance of three types of orthology definition: pairwise orthology, cluster orthology, and tree-based orthology. Using Inparanoid (Remm et al. 2001), we detected 1,025,849 pairwise "InparanOGs". For cluster orthology we used a method based on COG (Tatusov et al. 1997), yielding 8,044 triangle based "triOGs" and 10,754 pair based "duOGs". For specific purposes (see Appendix), we composed subsets of OGs without paralogs (8,722 unambiguous duOGs and 6,488 unambiguous triOGs) and OGs that occur exactly once in every species (64 pan-duOGs and 59 pan-triOGs). To compose tree-based orthology, phylogenetic trees were analyzed with LOFT (van der Heijden et al. submitted). LOFT does not impose a phylogeny on the data, but assigns orthology relations based on the species overlap between the branches of a phylogenetic tree. Because tree-based orthology yields levels of orthology, it needs to be operationalized between species pairs. We identified 858,622 distance tree-duOGs, 820,007 distance tree-triOGs, 856,363 likelihood tree-duOGs and 822,570 likelihood tree-triOGs. Further details about the orthology approaches can be found in the Appendix Orthology predictions are available at www.cmbi.ru.nl/~dutilh/phylogenomics.

Phylogenomics

Phylogenomic trees based on gene content were calculated from presence-absence profiles using either distance (Dutilh et al. 2004; Korbel et al. 2002) or parsimony (Farris 1977; Felsenstein 1989). In the distance approach, we corrected for genome size, because distantly related species with large genomes may share more genes than closer related species with small genomes (Appendix, Figure 28).

For the superalignment approach, Muscle multiple alignments (Edgar 2004b) of either unambiguous cluster OGs or pan-OGs were concatenated to form a superalignment. Unambiguous OGs that are absent from certain species were coded with question marks, and form gaps in the alignment (Philippe et al. 2004). In some superalignment trees, we analysed the effect of selecting unambiguously aligned amino acids by using GBlocks (Castresana 2000). We used either distance or maximum likelihood approaches to reconstruct the superalignment trees. The superdistance trees were calculated from superdistance matrices, based on the average distance over all OGs that are shared between the two species. We analysed the effect of correcting for rapidly evolving OGs by using SDM* (Criscuolo et al. 2006). Supertrees were composed of distance or maximum likelihood gene family trees. To integrate the different phylogenetic trees into a phylogenomic supertree, we used either the majority rule from Consense (Felsenstein 1989), or CLANN (Creevey and McInerney 2005). For further details see the Appendix, all the trees are available at www.cmbi.ru.nl/~dutilh/phylogenomics.

Scoring the reconstructed trees

To score the reconstructed phylogenomic trees, we use the target phylogeny in Figure 11. A phylogeny receives one point for each of the resolved partitions that is correctly retrieved, so a maximum of 19 points can be obtained. Note that, for example, the node "Yli primitive in Hemiascomycetes" refers to the (Ago, Cal, Cgl, Dha, Kla, Kwa, Sba, Sca, Sce, Skl, Sku, Smi, Spa) branch (see Figure 11). This means that this node can contribute a point for a certain tree, even if the Hemiascomycetes are not monophyletic in that tree, for example if *Y. lipolytica* clusters with *Sch. pombe*. In that case, however, it will not receive a point for the "Hemiascomycetes" node.

Results

We present a systematic comparison of two important factors in phylogenomic inference: the orthology approach and the level of integration of phylogenetic information to a genomic scale. We use various implementations for each of these approaches, such as the inclusive pair-based or the more restrictive triangle-based cluster OGs; and distance, maximum likelihood or parsimony for the reconstruction of the tree (Figure 10 and Appendix). Thus, we automatically construct 54 phylogenies from the available genome data of 26 Fungi. To assess the performance of the phylogenomic methods, we compare the nodes in the reconstructed trees to the 19 resolved nodes of a partly unresolved golden standard phylogeny based on extensive literature research (Figure 11 and Appendix). All of the canonical phylogenomic methods that we tested perform remarkably well at reconstructing the known fungal phylogeny. The phylogenomic trees in the three sequence-based approaches (superalignment, superdistance and supertree) recovered at least 16 out of the 19 target nodes. This constitutes a major distinction with the gene content trees, that performed much less well: even the best methods recovered no more than 13 nodes. All the phylogenomic trees can be found in the Appendix.

Collapsing recent duplications to gain data

We included two types of cluster orthology: the inclusive pair-based "duOGs", and the more restrictive triangle based "triOGs" (see "Methods"). A subset of these cluster OGs are the unambiguous OGs, that occur no more than once in every species. Even more constrained are the pan-orthologs, that are both unambiguous and universal, occurring exactly once in every species. We detected 8,722 unambiguous duOGs, 6,488 unambiguous triOGs, 64 pan-duOGs and 59 pan-triOGs in the Fungi. This result depends on collapsing the recent duplications, as identified from the phylogenies by LOFT (van der Heijden et al. submitted), before selecting the unambiguous OGs from the cluster OGs (see Appendix). Without collapsing recent duplications, we retrieved no more than 4,421 unambiguous duOGs, 4,887 unambiguous triOGs, 13 pan-duOGs and 13 pan-triOGs. This difference (an average of 42%) illustrates the necessity to filter out species-specific gene expansions and systematic errors,

such as the diploid genome assembly of *Can. albicans* (Jones et al. 2004), to increase the number of genes that can be considered.

Orthology approaches

An orthology definition that considers a recent last common ancestor will have a higher resolution than one that considers a more ancient common ancestor. Thus, pairwise orthology and tree-based orthology should, in principle, obtain a higher resolution than cluster orthology, that includes in a single OG all gene duplications since the last common ancestor of all the species compared. However, pairwise orthology incorporates information from only two species, and may miss genes that cluster orthology and tree-based orthology can identify. We expected tree-based orthology, that includes sequence information from many different species, while allowing a high-resolution view where necessary, to combine the advantages from pairwise and cluster orthology. However, although the orthology definition does turn out to be an important factor in the quality of a phylogenomic tree, the highest scoring trees were based on either unambiguous cluster OGs (duOGs and triOGs) or pan-triOGs, rather than tree-based OGs.

It is striking that although there is a large overlap between the 64 pan-duOGs and 59 pan-triOGs (56 OGs are identical), the pan-triOGs give better trees in both the superalignment and the supertree approach. However, the choice for one of these orthology definitions is no warrant for a good phylogeny. Both the unambiguous cluster OGs and the pan-triOGs also produced relatively low-scoring trees in every phylogenomic approach (Figure 10).

Superalignment trees and supertrees can recover all target nodes

Superalignment can be considered the most successful phylogenomic approach: four of the 14 superalignment trees correctly infer all 19 target nodes (see Figure 10). The most difficult to recover as a monophyletic group are the Ascomycota (although not for the trees constructed with maximum likelihood) and the (Mgr, Ncr) node (Figure 11). In those superalignment trees that did not group *M. grisea* with *N. crassa*, neither of these species was preferentially found at the root of the Sordariomycetes.

Selecting the unambiguously aligned positions of the superalignment using GBlocks (Castresana 2000) made it computationally possible to include more unambiguous OGs (Appendix), which led the unambiguous duOGs to match the results of the unambiguous triOGs (Figure 10). However, the decrease in the number of aligned positions that GBlocks brought about in the pan-triOGs, resulted in a sub-optimal tree (Figure 10). It appears that it is not simply the selection of unambiguously aligned positions, but rather the increase in the amount of high quality data that leads to a better phylogeny. To further test this, we composed Consense supertrees from an increasing number of phylogenetic distance trees of the most restrictive OG set, the 59 pan-triOGs. Interestingly, no two single gene trees were identical, and none was identical to the target: on average, they recover only 11.5 nodes. Yet when we combine at least 30-40 phylogenetic trees to a supertree, we already recover the external golden standard (Figure 29).

Three of the 12 phylogenomic trees inferred using the supertree approach correctly recover all 19 target nodes. The Consense supertree based on phylogenetic distance trees from pan-triOGs is identical to the four highest scoring superalignment trees (Figure 12a), but differs slightly from the equally high-scoring Clann supertrees based on phylogenetic maximum likelihood trees from both duOGs and triOGs (Appendix). This is possible because of the unresolved nodes in the target phylogeny. Note that superdistance and gene content trees never retrieve all 19 target nodes.

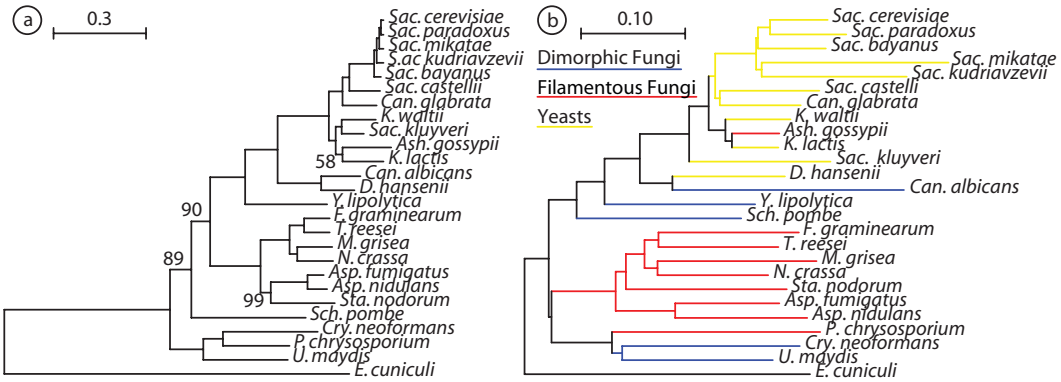

Figure 12. Phylogenomic trees. a) One of the two highest scoring fungal topologies. This topology was recovered by four superalignment trees and one supertree. A ML tree based on a superalignment of pan-triOGs, a ML tree based on a GBlocks-filtered superalignment of unambiguous duOGs (present in >24 species, 132,409 positions; this is the tree displayed, only bootstrap values <100% are indicated) or triOGs (present in >24 species), a distance tree based on a superalignment of pan-triOGs, and a Consense supertree based on phylogenetic distance trees of pan-triOGs. b) Gene content tree. Bio-NJ distance tree based on the InparanOG gene content distance between two species (see "Methods" and Supplemental Material). Like the other gene content trees, this tree indicates convergence in gene content of species with similar life styles.

Gene content trees have a phenotypic bias

Compared to the other phylogenomic methods, the gene content trees perform relatively poorly at recovering the required target nodes: on average, they only recover 10.38 nodes. Several numbers stand out in Figure 11. While almost all the other trees group the Hymenomycetes, (Sce, Smi, Spa) and (Ago, Kla, Kwa, Skl) together, none of the gene content trees recover these nodes. The distance based gene content trees also fail to retrieve the Ascomycota as a monophyletic group, although this proves to be a problem for most superdistance trees as well. Interestingly, we find that part of the explanation for these biases can be found in the lifestyle of the Fungi (Figure 12b). Although *Sch. pombe* shares relatively many genes with the Basidiomycota (Appendix and Figure 28), and might thus be expected to cluster at the root of the Ascomycota, the main dichotomy we find within the gene content tree of the Fungi is between the yeasts on the one hand, and the filamentous fungi on the other. The dimorphic fungi, *Sch. pombe*, *Y. lipolytica* and in some cases *Can. albicans* as well, are more or less placed in between these two branches. The filamentous *P. chrysosporium* is drawn closer to the filamentous Euascomycetes within the Basidiomycota, breaking up the Hymenomycetes, and leaving the dimorphic *Cry. neoformans* and *U. maydis* as the more derived Basidiomycota in most trees. The filamentous *Ash. gossypii* stays close to its relatives, *K. lactis* and *K. waltii*, but the (Ago, Kla, Kwa, Skl) branch is never intact in the gene content trees: *Sac. kluyveri* is often at the root of this cluster. This may be a remnant genome size effect, as *Sac. kluyveri* is a very incompletely sequenced genome. To investigate the effect of the small outgroup *E. cuniculi* on the position of *Sac. kluyveri*, we removed *E. cuniculi* from the data set and recomposed the Bio-NJ distance tree based on the InparanOG gene content distance (Figure 12b). The position of *Sac. kluyveri* did not alter (not shown).

This strong phenotypic effect does not explain the inability of gene conent to reproduce the target branching order in the Saccharomyces sensu stricto branch. In part, this may be explained by the fact that the genome sequences of *Sac. bayanus*, *Sac. kudriavzevii* and *Sac. mikatae* only covered 85 to 95% (Cliften et al. 2003). Another issue that may specifically hinder the correct inference of the Saccharomyces sensu stricto branching order are differential gene losses following the complete genome duplication or alloploid genome fusion in these species (Langkjaer et al. 2003; Scannell et al. 2006; Wolfe and Shields 1997). Due to the large number of redundant genes that resulted from this event, and the differential processes of gene loss that followed in the descendant lineages, a

patchwork of overlapping gene repertoires will have been the result. Although such gene losses should not be in conflict with the evolutionary signal, it may be part of the reason that the gene content approaches were confounded, resulting in the deviations from the target phylogeny within the Saccharomyces sensu stricto clade.

Suggestions for the unresolved nodes in the fungal taxonomy

The target nodes we selected from the literature were recovered in most of our phylogenomic trees (Figure 11). This high recovery rate supports our perhaps subjective golden standard phylogeny. In addition we were faced with three nodes that remained ambiguous in our review of the literature (Table 12): the internal resolution of the (Ago, Kla, Kwa, Skl) partition; the most primitive clade in the Euascomycetes; and the most primitive clade in the Ascomycota (bold lines in Figure 10). In Table 3, we have scored the support for each of the possible branching orders in these unresolved nodes over the four phylogenomic approaches. Based on our phylogenomic data, we can make some careful conclusions about the issues that remained unresolved in the fungal phylogeny thus far.

In virtually all phylogenomic trees reconstructed in the current research, *Ash. gossypii* and *K. lactis* are sister species in the (Ago, Kla, Kwa, Skl) branch. In fact the literature references that reject this hypothesis do so with low support (Diezmann et al. 2004; Kurtzman 2003), while the references that support it present well supported nodes (Jeffroy et al. 2006; Kuramae et al. 2006; Tehler et al. 2003). All the phylogenomic approaches support a clustering of *K. waltii* and *Sac. kluyveri*, except for the gene content trees. This suggests that the correct phylogeny is ((Ago, Kla), (Kwa, Skl)), as we also found in the high-scoring phylogenomic tree in Figure 12a.

Table 3. Support among the trees in each of the phylogenomic approaches for the different possible branchings in the unresolved nodes of the fungal taxonomy (see Table 12).

	Ago, Kla	Ago, Kwa	Ago, Skl	Kla, Kwa	Kla, Skl	Kwa, Skl	Ago, Kla, Kwa	Ago, Kla, Skl	Ago, Kwa, Skl	Kla, Kwa, Skl	(Sord,(Euro,Sno))	(Euro,(Sord,Sno))	(Sno,(Sord,Euro))	(Hemi,(Eu,Arch))	(Eu,(Hemi,Arch))	(Arch,(Hemi,Eu))
Gene content (13)	10	0	0	1	0	0	8	0	0	0	4	5	4	0	11	0
Superalignment (14)	14	0	0	0	0	14	0	0	0	0	14	0	0	0	0	12
Superdistance (15)	14	0	0	0	0	15	0	0	0	1	13	0	2	2	0	1
Supertree (12)	12	0	0	0	0	12	0	0	0	0	12	0	0	4	2	4

Our phylogenomic trees are also quite consistent regarding which clade should be placed at an ancestral position in the Euascomycetes (blue bold line in Figure 11). Except for two of the superdistance trees, all sequence-based trees agree that *Sta. nodorum* groups with the Eurotiomycetes, and the Sordariomycetes are ancestral (Table 3). This is largely supported by the literature (Lopandic et al. 2005; Robbertse et al. 2006; Tehler et al. 2003), while the only contradictory references contain other Pleosporales or Dothideomycetes, but not the species *Sta. nodorum* itself. Strikingly, the *Sta. nodorum* node is the single ill-supported node in a recent analysis of Ascomycota (Robbertse et al. 2006).

The solution to the third unresolved issue, that of which is the most primitive of the three Ascomycotal clades (black bold line in Figure 11), is less evident than the two above. The initial hypothesis was that *Sch. pombe* would be the first to branch off the Ascomycotal lineage (hence the name Archiascomycetes), which is also supported by most, but not all, literature references (Table 12). In all but two of the gene content trees the Euascomycetes are the most primitive Ascomycota, even though *Sch. pombe* clearly shares more genes with the Basidiomycota than do the other Ascomycota (Appendix and Figure 28). Conversely, the superalignment trees confidently provide the Archiascomycetes with this label, and the superdistance trees and the supertrees are inconclusive. As the superalignment trees have correctly recovered most of the other nodes as well, we conclude

that their placement of the Archiascomycetes as the most primitively branching ascomycotic clade is the most reliable. Thus, the topology depicted in Figure 12a is our final suggestion for the fungal phylogeny.

Concluding remarks

We have systematically compared four phylogenomic approaches in parallel with three orthology definitions that define OGs at different levels of resolution. Using various algorithms and tree building methods, we composed a total of 54 fully automated phylogenomic trees. The main dichotomy in the topologies of the reconstructed trees is that between trees reconstructed using a sequence-based method, and trees reconstructed using gene content data (Figure 13). The phylogenomic trees that best reproduced the target phylogeny can be found among the superalignment trees and the supertrees, using either unambiguous cluster OGs or pan-triOGs. However, although these approaches can yield trees that are completely consistent with the current opinions on the fungal phylogeny, they are not a guarantee for a successful phylogenomic tree. For example, the CLANN supertrees based on pan-duOGs still only retrieved 16 of the 19 target nodes.

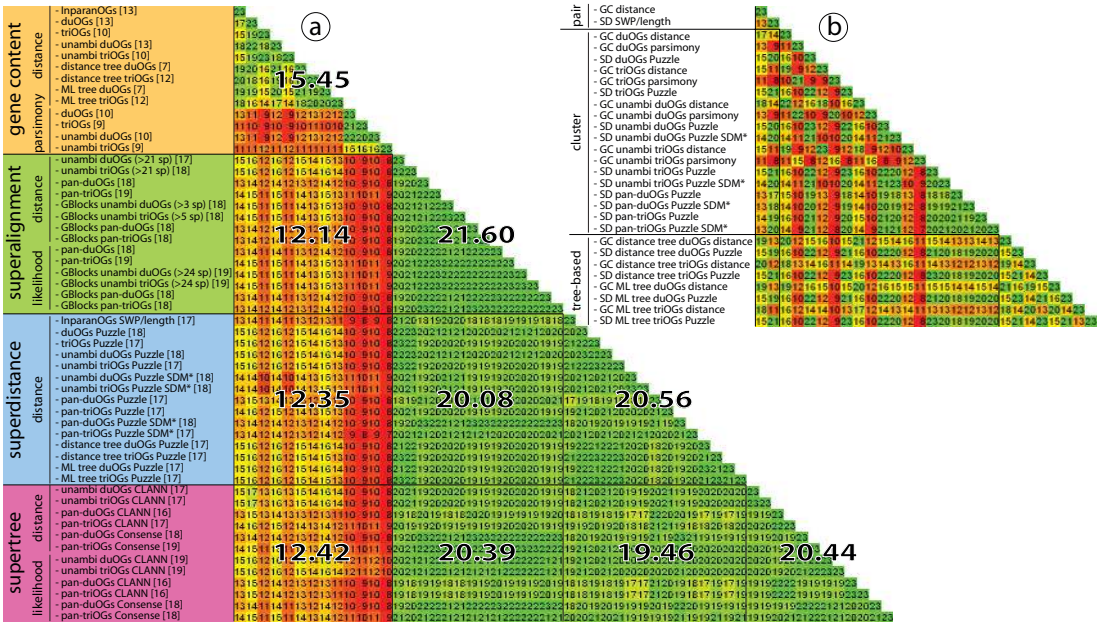

Figure 13. Similarity between the phylogenomic trees composed in this research, ordered based on a) the phylogenomic approach and b) the orthology approach. As superalignment trees and supertrees can not use pairwise or tree-based orthology, these approaches are excluded from figure b. The small numbers in the matrices are the number of partitions shared between each pair of trees. These numbers are color coded from many (green, max. 23) to few (red) shared partitions in the tree. The large numbers are the average number of shared partitions between all trees in the four main phylogenomic approaches.

Gene content trees recover relatively few of the target nodes. This is at least partly due to convergence in the gene repertoires of Fungi with comparable phenotypes: the evolutionary and phenotypic signals are combined in one tree (Snel et al. 1999). For example, we observe that the filamentous Euascomycetes and *P. chrysosporium* are drawn closer together, breaking the generally accepted topology of both the Ascomycota and the Basidiomycota (e.g. Figure 11). While prokaryotes from different lineages have previously been shown to assume convergent gene repertoires in comparable ecological niches (Zomorodipour and Andersson 1999), this is the first time (to our knowledge) that

a parallel between convergence in gene content and in phenotype has been shown in Eukaryotes, to the extent that it affects a gene content phylogeny.

This research strongly supports the fungal phylogeny as displayed in Figure 12a. The node that was recovered by the fewest phylogenomic trees is the basal position of the Archiascomycetes, represented by *Sch. pombe* here, within the Ascomycota. All other nodes are supported by many of the trees (see Figure 11 and Table 3). Although most of these branches are supported by recent literature (Table 11), this research helped provide support for those cases that were inconclusive (Table 3 and Table 12). What is striking in our phylogenetic findings is that that several of the fungal groups presented in the Genbank Taxonomy Database (Wheeler et al. 2002) should actually be adjusted. For example, Candida, Kluyveromyces, Saccharomyces and the Saccharomycetaceae remain mentioned as clades, while their members should be regrouped (see also (Diezmann et al. 2004; Kurtzman 1998; Kurtzman 2003; Lopandic et al. 2005; Prillinger et al. 2002; Tehler et al. 2003)).

Our phylogenomic trees of the Fungi reproduced many of the clades in accordance with the current taxonomic views. At least for the Fungi, we confirm a number of standard practices in the current phylogenomics field, albeit it with small differences relative to the less well-established approaches such as supertrees. A recent superalignment tree (Ciccarelli et al. 2006) has been criticised as being a "tree of one percent" of the genome (Dagan and Martin 2006). In the current study, we show that methods that are restrictive in selecting genes often create a phylogeny that is close to the golden standard. Apparently, this selection procedure is necessary to filter out the noise caused by evolutionary processes like gene duplication and gene loss, even in the absence of horizontal transfer (Andersson 2005). Complete genomes allow us to do this automatically and still retain enough genes to construct a reliable phylogeny. Our results indicate that a (1) maximum likelihood (2) superalignment tree based on (3) selected well aligned positions of (4) unambiguous cluster OGs, automatically derived at the level of resolution most suitable for the group of species considered, will yield a respectable tree. Maximum likelihood (1), because we find that distance trees may have trouble with the outgroup we used in this study; superalignment (2), because on average, this phylogenomic approach recovers the most target nodes; unambiguously aligned positions (3), because this enables the inclusion of more high quality data; and finally unambiguous cluster OGs derived at the level of the taxon of interest (4), because this ensures that you have the highest resolution possible.

Signature genes as a phylogenomic tool

Bas E. Dutilh, Berend Snel, Thijs J.G. Ettema and Martijn A. Huynen
Submitted

Abstract

Gene content has been shown to contain a strong phylogenetic signal, yet its usage is hampered by Horizontal Gene Transfer and parallel gene loss, and until now required completely sequenced genomes. Here, we introduce an approach that allows the phylogenetic signal in gene content to be applied to any set of sequences, using signature genes for phylogenetic classification. The hundreds of publicly available genomes allow us to identify signature genes for a range of taxa, and the presence of signature genes in an uncharacterized sample can help to detect its taxonomic composition. We identify 8,362 signature genes specific for 112 prokaryotic taxa. We show that these signature genes can be used to address phylogenetic questions on the basis of gene content in cases where classic gene content or sequence analyses provide an ambiguous answer, such as for *Nanoarchaeum equitans*, and even in cases where complete genomes are not available, such as for metagenomics data. The signature genes for which functional information is available reveal clade-specific processes, such as sporulation genes in Bacillaceae, and virulence-related genes, e.g. linked to the biosynthesis of phthiodiolone dimycocerosate esters in *Mycobacterium* and to alginate biosynthesis in *Pseudomonas* species.

Introduction

Gene content contains a strong phylogenetic signal (Snel et al. 1999; Tekaia et al. 1999), and has helped to clarify several taxonomic uncertainties (for review see (Snel et al. 2005)). Classic gene content is based on the fraction of genes shared between two genomes, and requires a data set of completely sequenced genomes to confirm not only the presence, but also the absence of each gene. If a complete genome cannot be obtained, gene content can still be used to address taxonomical questions by means of signature genes. In the signature gene approach, we use the wealth of completely sequenced genomes to define cores of genes for every clade. A core is the set of all genes common to (ubiquitous among) all genomes in a phylogenetically coherent group (Charlebois and Doolittle 2004). For an unidentified, even incompletely sequenced organism, its relatives can be identified by finding the overlap between its gene repertoire and these cores. Previously using this idea, we found that the number of signature genes that *Kuenenia stuttgartiensis* shares with the cores of potential sister clades supported the finding, based on a superalignment of 49 proteins, that this anaerobic ammonium oxidizing bacterium is closely related to the Chlamydiae (Strous et al. 2006).

When complete genomes are available, and when one wants to use a single method, we have shown gene content to be less suitable for phylogenomic inference than sequence similarity based approaches, at least in the Fungi (Dutilh et al. 2007). However, gene content does contain a phylogenetic signal that can be exploited if the right genes are selected (Dutilh et al. 2004). Furthermore, sequence-based approaches have to restrict themselves to sequences with a wide phylogenetic distribution. The presence or absence of genes that are stable in evolution provides independent phylogenetic evidence, that can complement sequence-based information. This information is independent from the data used in sequence similarity-based phylogenies because 1) gene content it evolves at a different level (whole genes in stead of residues), and 2) signature genes specifically exploit those genes that do not have a very wide phylogenetic distribution. This two-fold independence makes gene content a valuable complementary source of phylogenetic information to sequence similarity based approaches.

Investigations with a functional angle use signature genes as a way to characterize a taxon. Proteins that are present in all the domains of life have been used to reconstruct ancestral genomes and minimal gene-sets (Harris et al. 2003; Koonin 2003). These cores that are common to all living organisms can differ slightly in size and composition, depending on the method used to find them, but they always contain informational genes that are involved in the single process that unifies all of Life, i.e. the translation of genes into proteins. At the level of the Eukaryota, the core makes up components of the cytoskeleton, inner membranes, RNA-modification machinery and the major elements of intracellular control systems (Hartman and Fedorov 2002). For several prokaryotic clades, most of the signatures are annotated as hypothetical. However, those that can be related to a certain function reveal a sensible pattern, e.g. signature genes in Cyanobacteria are directly or indirectly involved in photosynthesis (Martin et al. 2003), and Chlamydiales specific signature proteins contain membrane proteins that are possibly involved in the interaction of the pathogen with host cells (Griffiths et al. 2006). Thus, as far as a function is known, signature genes are related to the unique and unifying features of taxa at a range of levels.

Signature genes have been identified for several taxa on an *ad hoc* basis, often using a reference genome, sequence similarity searches and manual inspection of the results (Gao et al. 2006; Griffiths et al. 2006; Kainth and Gupta 2005; Martin et al. 2003). The large variety of completely sequenced genomes that have become available in recent years, together with high quality automated cluster orthology definitions (Tatusov et al. 2000; von Mering et al. 2007b) and reliable species phylogenies (Ciccarelli et al. 2006), enable us to take a more systematic approach, and find signature genes on a large scale for many clades throughout the tree of life. To do this, we introduce a simple, phylogeny-based definition: the signature genes of a clade are those genes that occur in every daughter lineage of a

clade, but nowhere outside it (Figure 14). The most parsimonious explanation for such a distribution is that the gene originated at the root of this clade, and has an important function for the species in this clade, so that it is retained in all the descendant lineages. With a pre-defined species tree as a guide (Ciccarelli et al. 2006), we use this definition to find cores of genes for clades of different ages, at all levels in the tree. As our definition only requires that the gene is retained in at least one species per daughter of a clade, it allows for species specific losses, for example in the degenerated genomes of parasites (Fraser et al. 1995). Thus, it is broader than a definition that requires complete coverage of a clade. We introduce a coverage score that takes into account asymmetric taxon sampling to increase the reliability of thus defined signature genes.

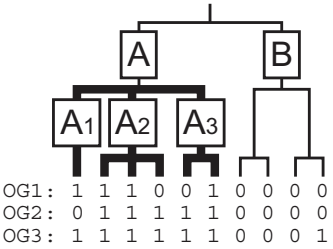

Figure 14. Definition of signature genes based on a partially unresolved phylogeny. For every species, presence (1) or absence (0) of three genes (OGs) is indicated. In this example, only OG1 is a signature for clade A, as it is present in clade A1, clade A2 and clade A3, but not in clade B. Although OG2 and OG3 are present in more species within clade A, they are not a signature for clade A because OG2 is not present in clade A1, and OG3 is present outside of clade A.

Results

Using the definition of signature genes and the method outlined above (Figure 14), we have identified 8,362 sets of signature genes (orthologous groups or OGs) for 112 clades throughout the prokaryotic tree of life (see Figure 15 and Methods) using a partly unresolved reference phylogeny (Ciccarelli et al. 2006) and a predefined set of OGs (von Mering et al. 2007b). Homologous OGs, as detected by profile-to-profile comparisons, that had largely complementary phylogenetic distributions were merged to prevent high rates of sequence evolution to lead to an overestimation of the number of signature OGs (see Methods). Subsequently signatures for a given clade were defined as those OGs that are specific for the corresponding node, and occur in every daughter lineage (Figure 14). The many signature genes we found underline the phylogenetic signal that exists in gene content. Conversely, the results justify the suspicion of clades that are completely void of signature genes. Figure 15 shows the number of signature genes identified for each branch that defines a taxon. Most taxa are confirmed by the signature genes. For example, even the Bacteroidetes / Chlorobi group, which is a difficult bacterial division to retrieve in gene content trees (see Figure 30), is supported by seven signature genes. In contrast, the controversial grouping of *Thermotoga maritima* and *Aquifex aeolicus* is not supported by any signature genes, which casts more doubt on it.

Assessing species distribution in metagenomics samples

The initial motivation for this study was to find an approach that makes use of the phylogenetic signal in gene content, but can be employed for incomplete genomes. To show that this application works, we have mapped the taxonomic distribution of signature genes identified in three metagenomics samples from the Sargasso sea (Venter et al. 2004), agricultural soil and three deep-sea "whale fall" carcasses, that have been assigned to OGs (Tringe et al. 2005). Beside the phylogenetic analyses in the papers that introduced these data sets, these environmental samples have recently been included in another phylogenetic analysis based on 31 universal marker genes (von Mering et al. 2007a), which provides insightful additional reference material to compare our signature genes approach with sequence-based approaches.

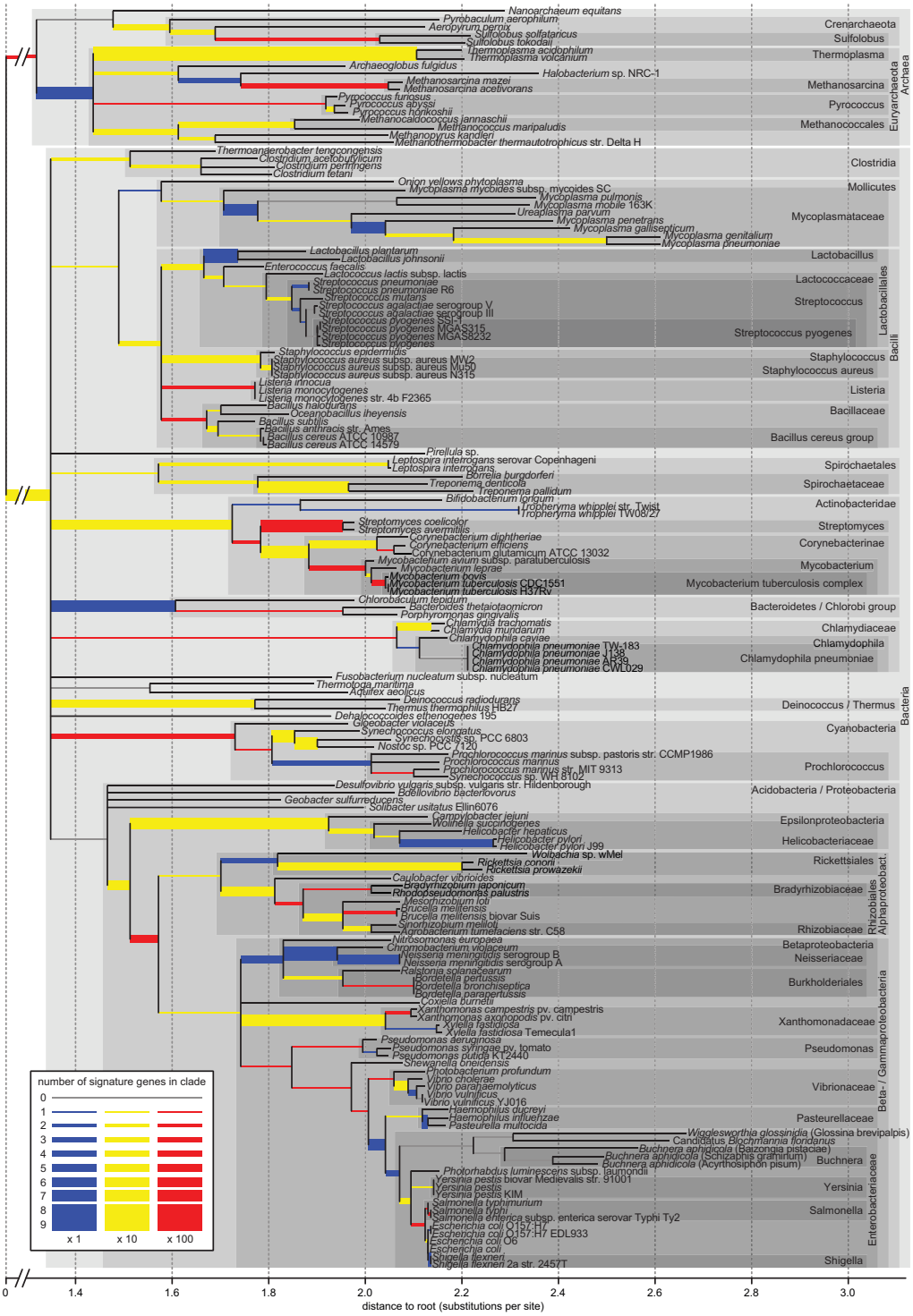

Figure 15. Amounts of signature genes identified in prokaryotic taxa. The unresolved phylogeny is based on a superalignment tree (Ciccarelli et al. 2006) where we collapsed nodes with a bootstrap value lower than 80% and removed the Eukaryota. Several node names used in this paper are indicated (to the right) with gray boxes. Branch widths and colors indicate the number of signature genes found for each node (see legend).

In the sequence-based approaches, the soil sample was shown to contain the largest species diversity, mainly consisting of Chloroflexi and Acidobacteria (both not in our data set of complete genomes), Alphaproteobacteria, and Bacteroidetes, but also many Betaproteobacteria, Gammaproteobacteria, Gemmatimonadetes (not in our data set), Deltaproteobacteria (not a clade in the reference tree, see Figure 15) and Actinobacteria (Figure 31, Figure S2 B in Tringe et al. 2005; and Figure 34 A, Figure S1 A in von Mering et al. 2007a). In the original analysis that was based on six phylogenetic markers (16S rRNA, RecA, EF-Tu, EF-G, HSP70 and RNA polymerase B) and in the later analysis based on 31 universal marker genes, the phylotypes in the Sargasso sea were shown to be dominated by Alpha- and Gammaproteobacteria, but they were also shown to contain many Cyanobacteria, Bacteroidetes and Betaproteobacteria (Figure 33, Figure 6 in Venter et al. 2004; and Figure 34 B, Figure S1 B in von Mering et al. 2007a). Finally, the whale fall samples were primarily mapped to Bacteroidetes, Alphaproteobacteria, Epsilonproteobacteria and Gammaproteobacteria (Figure 33, Figure S4 A in Tringe et al. 2005; and Figure 34 C, Figure S1 C in von Mering et al. 2007a). As Figure 16 shows, the previously reported species distributions show a surprisingly good correspondence with the clades for which we find signature genes in these metagenomic samples, although in some cases, the precise proportions vary. Clearly, signature genes provide an independent tool that can be used to phylogenetically map unidentified, even incomplete genomes, or metagenomics data sets, allowing the exploitation of a complementary fraction of the data.

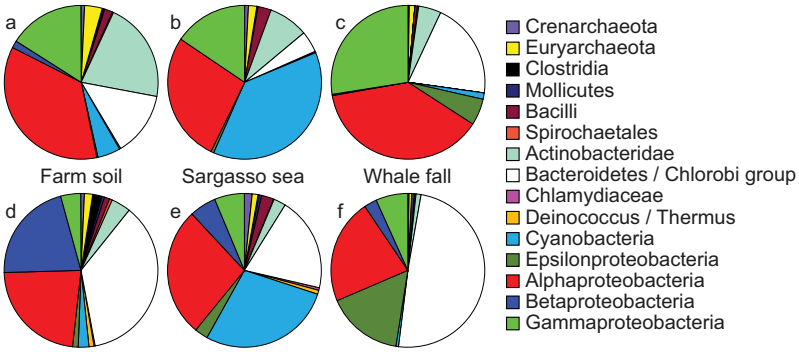

Figure 16. Fractions of signature genes present in three metagenomics data sets (Tringe et al. 2005; Venter et al. 2004). In pies a, b, and c, the fractions are the total numbers of signature genes found for each clade (including subclades); in pies d, e and f, the fractions are the percentages of the total number signature genes that exist for each clade. All three metagenomics data sets were highly dominated by bacterial signature genes (farm soil: 72%; sea: 78%; whale fall: 70%), archaeal signature genes were present in much lower percentages (farm soil: 0.05%; sea: 0.6%; whale fall: 0.1%). The phylogenetically less informative clades genes are not shown in these charts. This analysis is based on STRING 6.3 OGs as the mapping of the metagenomics data sets was only available for that version (kindly provided by C. von Mering).

Addressing taxonomic questions with signature genes

Our signature genes procedure also allows us to investigate in detail the taxonomic position of some early branching prokaryotic species, for which the phylogenetic signal in the sequences may have been lost. One by one, we removed *Aquifex aeolicus*, *Fusobacterium nucleatum*, *Halobacterium* sp., *Nanoarchaeum equitans* and *Thermotoga maritima* from the data set, and re-identified signature genes in the remaining 162 species. Table 4 shows which genes from the removed genomes were found as signature genes in the corresponding restricted data set. Thus, these signature genes can classify the removed genomes in terms of their taxonomic relatives.

Table 4. Signature genes shared by several species and potential sister clades. In some cases, no shared signature genes were found in the 1,000 randomized genome sets (e.g. o/e ratio 1/0). OGs that are linked with a hyphen were merged because they are homologous and have a non-overlapping taxon distribution (see Methods).

species	clade	o/e ratio	shared signature genes
A. aeolicus	Bacteria	60/0	60 COGs
	Acidobacteria / Proteobacteria	1/0	COG3034
	Alpha- / Beta- / Gamma- / Epsilonproteobacteria	21.13	COG3302, NOG13261, NOG09591-NOG17096
	Alpha- / Beta- / Gammaproteobacteria	2.48	COG4618, COG5611
	Helicobacteraceae (Epsilonproteobacteria)	1,000	NOG18902
	Rickettsiales (Alphaproteobacteria)	1,000	NOG07928
	Beta- / Gammaproteobacteria	500	COG4969
	Archaea	2,333	COG1423, COG1458, COG1503, COG1517, COG1730, COG2112, COG4831
	Crenarchaeota	1,000	COG4353
	Sulfolobus (Crenarchaeota)	1,000	NOG18904
	Methanosarcina (Euryarchaeota)	1,000	NOG09683
F. nucleatum	Bacteria	67,000	67 COGs
	Lactobacillales (Firmicutes)	1,000	NOG17664
	Mycoplasmataceae ex. *M. mycoides* (Firmicutes)	1,000	NOG19254-NOG36375
	Treponema (Spirochaetales)	500	NOG17678
	Alpha- / Beta- / Gamma- / Epsilonproteobacteria	20.13	COG2992, COG3713, NOG11181
	Alpha- / Beta- / Gammaproteobacteria	2.48	COG4797, NOG18514
	Pasteurellaceae ex. *H. ducreyi* (Gammaproteobacteria)	500	NOG09881
	Vibrionaceae / Pasteurellaceae / Enterobacteriaceae (Gammaproteobacteria)	5.10	COG2926
	Methanosarcina (Euryarchaeota)	1,000	NOG22419
Halobacterium sp.	Archaea	14,625	114 COGs, COG1591-NOG14885, COG3353-NOG29648, COG4023-NOG17603, NOG39364-NOG10118
	Euryarchaeota	5,000	COG1422, COG1777, COG2150, COG3390, COG1711-NOG33052
	Archaeoglobus / Methanosarcina (Euryarchaeota)	3,000	COG4749, COG4885, COG5427
	Methanosarcina (Euryarchaeota)	1,500	NOG06067, NOG17658, NOG15033
	Methanococcales / *M. kandleri* / *M. thermoautotrophicus* (Euryarchaeota)	1,000	COG3363
	Pyrococcus ex. *P. furiosus* (Euryarchaeota)	1,000	NOG24228
	Leptospira (Spirochaetaceae)	500	NOG15034
	Actinobacteridae	167	COG5282
	Mycobacterium (Actinobacteridae)	333	NOG20057
	Streptomyces (Actinobacteridae)	400	NOG36090, NOG15774
	Cyanobacteria	400	COG4250, COG5524
	Alpha- / Beta- / Gammaproteobacteria	2.57	COG3205, COG4538
	C. vibrioides / Rhizobiales (Alphaproteobacteria)	143	COG3743
N. equitans	Archaea	22,333	66 COGs, NOG21880
	Euryarchaeota	2,000	COG1311, COG1933
	Methanosarcina (Euryarchaeota)	1,000	NOG11162
	Pyrococcus (Euryarchaeota)	1,000	NOG17563
T. maritima	Bacteria	60,000	60 COGs
	Clostridia (Firmicutes)	1,000	NOG22606
	Archaea	1,200	COG1031, COG1184, COG1635, COG1992, COG3374, COG5014
	Pyrococcus (Euryarchaeota)	1,000	NOG13536
	Pyrococcus ex. *P. furiosus* (Euryarchaeota)	1,000	NOG23777

A difficult case in classic gene content trees is *Halobacterium* sp. (Dutilh et al. 2004). Due to horizontal gene transfers with the Bacteria (Kennedy et al. 2001), this euryarchaeon is often found at the root of the Archaea in gene content trees (see also Figure 30). However, our alternative application of gene content shows that many more signature genes than expected are shared with several Euryarchaeota clades (Table 4), supporting the taxonomic positioning of *Halobacterium* sp. in the Euryarchaeota.

N. equitans is a tiny thermophilic archaeal parasite that was originally assigned to a novel, anciently branching archaeal phylum on the basis of an unpolished superalignment approach (Huber et al. 2002; Waters et al. 2003). Because of the split structure of many of its genes, the position that *N. equitans* is a living fossil still receives support (Di Giulio 2006), but the argument in this paper leans heavily on the tRNA molecule, which is usually codified in a single gene, but in *N. equitans* comprises

two separate genes that are not contiguous in the genome. However, evidence for other affiliations can also be found. A BLASTP-based survey of the phylogenetic pattern of all *N. equitans* ORFs finds a strong link with the Euryarchaeota (Brochier et al. 2005), more specifically the Thermococcales. We also find that *N. equitans* clusters with the Pyrococci in a classic gene content tree (Figure 30). Conversely, in the curated superalignment phylogeny we used as a reference (Ciccarelli et al. 2006), *N. equitans* clusters with the Crenarchaeota with high bootstrap value (cf. Figure 15). However, not one signature is found for this *N. equitans* / Crenarchaeota clade (Figure 15). If we re-identify signature genes for all clades in the phylogeny after removing *N. equitans*, we find that several Euryarchaeota, among which Pyrococcus, share many more signature genes with *N. equitans* than expected, while no links to any Crenarchaeota clade are observed (Table 4). Therefore, our results support the position of *N. equitans* as a derived Euryarchaeote, possibly related to Pyrococcus (Brochier et al. 2005).
As these examples show, signature genes can complement traditional sequence based methods and classic gene content based on complete genomes in addressing taxonomic questions. Conceptually, this gene-content approach is reminiscent of the slow-fast method (Brinkmann and Philippe 1999), where slowly evolving sites in an amino-acid alignment are selected as those positions that have not mutated within pre-defined clades. These positions are the most reliable for inferring ancient relationships, as fast-evolving sites are likely to be mutationally saturated, obscuring the phylogenetic signal. Signature genes evolve slowly at the gene content level. Especially the signature genes with high coverage scores have undergone little loss or horizontal gene transfer, and are thus strong indicators of phylogenetic relatedness.

Ancient signature genes tend to be informational, recent signature genes are more operational

On the basis of the COG functional categories (Tatusov et al. 2000), those signature genes that were based on COGs could be included in an analysis of their functional repertoire. We find ancient signatures (Bacteria and Archaea) to be heavily dominated by COGs from the "Information storage and processing" category (Figure 17), in good agreement with the many previous observations that practically all the genes shared by every living organism are related to the translation of genetic material (Charlebois and Doolittle 2004; Koonin 2003). Conversely, signatures for clades that diverged later are more likely members of the "Cellular processes and signaling" and "Metabolism" categories. Despite these trends, most of the signature genes we identify remain "Poorly characterized", as has been previously observed in several single-clade analyses (Gao et al. 2006; Griffiths et al. 2006; Kainth and Gupta 2005; Martin et al. 2003). Note that on top of this majority of the 653 signature COGs, come 7,934 signature NOGs, that are also largely uninvestigated (these numbers do not add up to 8,362 because some homologous OGs were merged, see Methods).

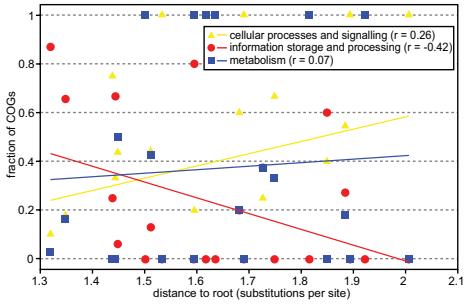

Figure 17. Functional distribution of the signature genes for clades with different ages with linear regression lines. Only signature genes with a functional annotation (Tatusov et al. 2000) are included (653 COGs).

Investigating the functions of signature genes in more detail, we found that they often carry out comparable processes in different clades (Table 5). For example, several sets of genes that are a

signature for different clades are independently related to conjugation (e.g. NOG08929, NOG10080, NOG10338, NOG12185, NOG12276, NOG13815, NOG18007, NOG21014 and NOG24221 in the Vibrionaceae / Pasteurellaceae / Enterobacteriaceae clade; or COG5442, COG5443 and COG5461 in the Rhizobiales / *C. vibrioides* clade). Similarly, independent sets of genes related to the flagella are also signatures for different clades (e.g. COG1681, COG1955, COG2874, COG3353, NOG13591 and NOG22539 in the Archaea; COG3351, COG3352 and COG3354 in the Euryarchaeota; NOG12184, NOG15376, NOG28820, NOG28897 or NOG32922 in the Salmonella / Escherichia / Shigella clade). Triggered by these observations, we decided to quantitatively compare the functional relationships of signatures in general. We used the STRING score as an indication of the intricacy of the functional relationships between genes (von Mering et al. 2007b), because STRING incorporates several sources of information, such as genomic context, high-throughput experiments, (conserved) co-expression and previous knowledge to link the functions of genes. We find that the average STRING score between all linked COGs in STRING 7.0 is 0.235. For the subset that linked two signature genes, this average score was 0.241, but if only genes were selected that were a signature for the same clade, the average STRING score increased to 0.349. Also, while signature genes constitute a minority of 26% of the almost 40 thousand OGs in STRING 7.0, we find another signature gene among their highest scoring interaction partners in 54% of the cases, and in 66% of those cases one of the highest scoring interaction partners is a signature gene for the same clade. These results confirm that the functions of signature genes are more related than the functions of genes that have a scattered distribution, especially if they are a signature for the same clade.

Table 5. Some functions found among the signature genes for prokaryotic clades.

(related to) function	clades
Conjugate transposon, pili, competence	Alpha- / Beta- / Gammaproteobacteria; Bacilli; *B. thetaiotamicron* / *P. gingivalis*; Salmonella / Escherichia / Shigella; Deinococcus / Thermus; Vibrionaceae / Pasteurellaceae / Enterobacteriaceae
DNA repair, DNA replication, DNA binding	Enterobacteriaceae; Gammaproteobacteria except Xanthomonadaceae and *C. burnetii*; Vibrionaceae / Pasteurellaceae / Enterobacteriaceae
Fatty acid synthase	Bacteria
Flagella	Archaea; Euryarchaeota; Spirochaetaceae
Peptide transport	Alpha- / Beta- / Gammaproteobacteria; Euryarchaeota; *S. oneidensis* / Vibrionaceae / Pasteurellaceae / Enterobacteriaceae
Phage related	Alpha- / Beta- / Gammaproteobacteria; Lactobacillus; *S. oneidensis* / Vibrionaceae / Pasteurellaceae / Enterobacteriaceae; Vibrionaceae / Pasteurellaceae / Enterobacteriaceae
Photosystem I / II, phycocyanin	Cyanobacteria
Protease	Bacteria; Beta- / Gammaproteobacteria
Ribosome	Archaea; Bacteria; Crenarchaeota; Gammaproteobacteria except Xanthomonadaceae and *C. burnetii*
Secretion, membrane	Alpha- / Beta- / Gammaproteobacteria; Alpha- / Beta- / Gamma- / Epsilonproteobacteria; Cyanobacteria; Deinococci; Euryarchaeota; Gammaproteobacteria except Xanthomonadaceae and *C. burnetii*; Methanococcales / *M. kandleri* / *M. thermoautotrophicus*; Pseudomonas; Vibrionaceae / Pasteurellaceae / Enterobacteriaceae
Sporulation; cell division	Bacillaceae; Bacilli
Toxins, virulence	Gammaproteobacteria except Xanthomonadaceae and *C. burnetii*; Mycobacterium except *M. avium*; Pseudomonas; Vibrionaceae

Signature genes we find for the Enterobacteriaceae (NOG06760, NOG13543 and NOG13893) and some of its parent clades within the Gammaproteobacteria (e.g. COG3006, COG3050, COG3095, COG3923 and COG4776 in the Vibrionaceae / Pasteurellaceae / Enterobacteriaceae clade; or COG3130, COG3160, COG4568 in the Gammaproteobacteria except Xanthomonadaceae and *C. burnetii* clade) have DNA related functions, suggesting that parts of DNA replication and repair mechanisms have been invented or fine tuned throughout the history of this lineage. Phage related signature proteins (e.g. COG3498, COG3499, COG3948, COG4220, COG4385, COG4540, COG5004, NOG06467, NOG14663, and NOG22103 in the Alpha / Beta / Gammaproteobacteria clade; or NOG08626, NOG09661, NOG22321, and NOG31424 in the Vibrionaceae / Pasteurellaceae / Enterobacteriaceae clade) are often found in syntenic regions on the genome (they are linked by high neighborhood scores in

STRING (von Mering et al. 2007b)). Prophages, in some cases constituting up to 10-20% of a bacterial genome, are major sources of innovation for individuals and species, and can lie in residence for very long times (Casjens 2003).
Other examples from Table 5 underline the distinctive character of a clade. Not surprisingly, many Cyanobacteria signature genes are related to photosynthesis (Martin et al. 2003), including genes involved in photosystem I and II, phycocyanin, phycoerythrin and allophycocyanin. Likewise, the Bacillaceae contain a striking number of spore-related signature genes, such as spore coat proteins and genes related to peptidoglycan biosynthesis. As we have observed in Figure 17, ancient clades like Bacteria and Archaea contain many ribosomal signature genes, although more recent clades seem to have also added specific genes to this system at a later stage. Importantly, we also identify several virulence-related signature genes, such as NOG24770 for the Mycobacterium tuberculosum clade, that has previously been identified as a signature gene in a Southern blot analysis (Rindi et al. 2001). Additionally, we identified several other signature OGs for this clade (NOG31315) and other Corynebacterinae subclades (e.g. NOG26217, NOG31900 and NOG34480), that share several STRING links to a group of genes involved in the biosynthesis of phthiodiolone dimycocerosate esters (PDIMs). PDIMs are a category of virulence-enhancing lipids, that are specific for mycobacterial pathogens. For Pseudomonas, we find a whole group of signature genes (NOG28203, NOG26205, NOG24940, NOG25112, NOG35177, NOG25420 and NOG30475) involved in alginate biosynthesis. Alginate is an extracellular polysaccharide produced by Pseudomonas strains found in the pulmonary tracts of chronically infected cystic fibrosis patients. Thus, whereas for most of the signature genes by far, the function is hypothetical or even completely unknown, the trend that we observe in the remaining cases promises a wealth of new clade specific biology awaiting discovery.

Conclusions

One of the weaknesses of classic gene content trees is that they require completely sequenced genomes (Snel et al. 1999; Tekaia et al. 1999), which may not always be available (Tringe et al. 2005). Here, we solve this problem by introducing signature genes as a novel approach to employ gene content for phylogenetic analysis. The wealth of complete genomes allows us to identify signature genes for a range of taxa, and the presence of signature genes in an unidentified sample can help to detect the taxonomic composition of the query. However, the comprehensive overview of the gene repertoires of a diversity of species has also uncovered a great plasticity in gene content, with examples of extensive gene loss (for example in parasitic genomes (Fraser et al. 1995)), and horizontal gene transfer in prokaryotes (Doolittle 1999b) as well as in Eukaryotes (Andersson 2005). Thus, a strict search for signature genes, that requires complete coverage of all genomes within the taxon, will only yield limited results (Charlebois and Doolittle 2004). To overcome this, we develop an intuitive definition that defines as signatures of a clade those genes that occur in every daughter of that clade, but complete coverage is not required. A coverage score indicates the how well the signature gene has been retained in the descendant lineages.
This study has identified a large set of 8,362 signature genes for 112 clades throughout the tree of life (Figure 15). These many signature genes underline the phylogenetic signal that exists in gene content. Based on a historical reconstruction (Figure 18), we expect that with the inclusion of more completely sequenced genomes, the number of signatures will grow, rather than shrink (Charlebois and Doolittle 2004), and the number of signature genes per taxon will remain quite stable. This is the result of on the one hand the sampling of more daughters per taxon, which increases the coverage requirement for a signature gene, and on the other hand the sampling of more species per daughter, which increases the species sampling, leading to more imperfect signatures. Theoretically, the number of signature genes may decrease due to their identification in species from other clades, or increase due to a more complete sampling of the taxon. So far, the Global Ocean Sampling project, the largest environmental sequencing project ever carried out, identifying almost 4,000 protein

families in 7.7 million sequences (Rusch et al. 2007; Yooseph et al. 2007), has hardly reduced the number of signatures for very ancient taxa (Bacteria and Archaea). Within the prokaryota, the authors find one Pfam domain that was thought to be Bacteria specific to be present in the Archaea, and four Archaea specific Pfam domains in the Bacteria (Yooseph et al. 2007). With the spring-tide of data from large-scale sequencing projects like the Global Ocean Sampling project, the trustworthiness of signature genes will increase, even if, or better, because some genes thusfar thought to be a signature have to be dropped, being discovered in other clades as well.

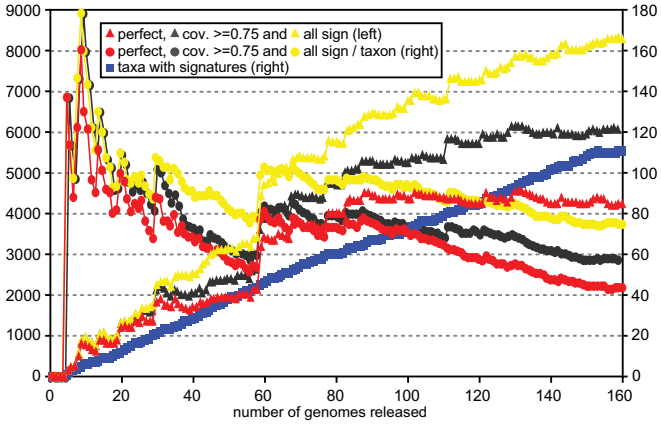

Figure 18. The number of signature genes, perfect signature genes (coverage score 1), and signature genes with a coverage score cutoff of 0.75, found with increasing numbers of completely sequenced genomes. The genomes are added one by one, in order of appearance (according to www.ncbi.nlm.nih.gov/genomes). Initially, the number of signature genes increases almost linearly with the appearance of more genomes. The 60th genome, that of S. avermitilis, completes the signature-rich Streptomyces clade (S. coelicolor was the 4th genome), and causes a great jump in the number of both perfect and normal signature genes. The average number of signature genes per clade reaches about 75 throughout the last 10-15 genomes, the perfect signatures go down to about 43.

Future work, aimed at polishing our initial approach to signature genes, may be expected to benefit from a higher resolution of the initial orthology definition. As we see in Figure 19 (Methods), a few genes that were identified as a signature for recent clades in clade-specific investigations based on sequence similarity (Gao et al. 2006; Griffiths et al. 2006; Kainth and Gupta 2005; Martin et al. 2003), were a signature for more ancient clades in our large-scale analysis (especially for the Alphaproteobacteria). If a more fine-grained orthology definition would be used, our approach would also identify those OG as a signature for the more derived clades. This point is closely related to another important issue. By their manual analysis, the cited authors could allow the sporadic presence of signature genes in unrelated species, denominating them as cases of horizontal gene transfer (HGT). In our approach, instances of HGT may be recognized as signatures at a higher depth in the phylogeny, but will become signatures with a low coverage score. For example, a Chlamydiaceae signature gene that is horizontally transferred to the betaproteobacterium *N. europaea* (NOG04874) may, depending on the phylogeny, become a low-coverage signature for the Bacteria in stead of a high-coverage signature for Chlamydiaceae. In practice we handle these situations by requiring a high coverage score. As an example, we have included the numbers for signature genes with a coverage score cutoff of 0.75 in Figure 18 and Table 6.

In many cases, the functional annotations of signature genes include properties that are unique for the taxon in question, such as virulence-related genes for Mycobacterium, Pseudomonas and Vibrionaceae (Table 5). Research aimed at elucidating lineage-specific properties for the clades included in this work will benefit from the list of uncharacterized genes, which forms a wealth of suggestions for further experimental investigations into taxon-specific processes. Concluding, signature genes are a promising tool, that can be used in a number of research areas, from taxonomic

analysis of incomplete genomes and metagenomics data to the identification of clade specific genes.

Methods

Data

The reference phylogeny we used was based on a recent superalignment phylogeny of 31 universal protein families (Ciccarelli et al. 2006), excluding all but the 163 prokaryotic species that were also present in STRING 7.0 (von Mering et al. 2007b). We excluded the Eukaryota, because due to both the large sizes of the genomes and the highly asymmetrical taxon sampling, the eukaryotic signature genes would have obscured much of the statistical and functional signal in the prokaryotic signature genes. To account for uncertainties in the Ciccarelli tree, we collapsed the nodes with a bootstrap value lower than 80%, resulting in a partly unresolved reference phylogeny (Figure 15).

The proteomes and orthology definitions were downloaded from STRING 7.0 (von Mering et al. 2007b); only COGs and NOGs present in at least two prokaryotic species were included in this study. Our concept of signature genes identifies those genes that originated at the root of a clade, and are retained in all lineages. If, for some reason, an OG has undergone accelerated evolution in a certain clade of species, these genes may be erroneously assigned to a new OG. This could cause an overestimation of the number of signature genes for the accelerated clade, or also an underestimation for the parent clade, where the OG actually originated. To avoid this, we used a highly sensitive approach to identify homology between OGs by performing profile-profile searches. We aligned the sequences of each OG using MUSCLE (Edgar 2004b). Hidden Markov models (HMMs) were created using HHmake (HHsearch 1.4 (Soding 2005)) and calibrated against a database comprising 1,250 random SCOP domain HMMs (Murzin et al. 1995). We then compared the HMM profiles all-against-all using HHsearch. For the homologous OG-pairs (query and hit aligned over >50% of their sequence; score>90), we inspected their distribution in the species tree, and if the parent clade of the OG with the narrowest distribution did not contain the OG that was more widely distributed, they were considered mergeable. We then merged the mergeable OGs using CFinder (Palla et al. 2005), at the level of communities. Remaining OGs that were not included in these communities were merged as pairs. Thus, we merged 2,958 of the 18,611 OGs, obtaining a final total of 17,323 OGs.

Table 6. Statistics of all signature genes identified, the signature genes with a coverage score cutoff of 0.75, and perfect signature genes.

	taxa with signatures	number of signatures	average coverage score
signatures	112	8,362	0.80
signatures (coverage >= 0.75)	106	6,177	0.94
perfect signatures	98	4,342	1.00

Signature genes and coverage score

Signature genes were identified automatically based on the OGs and the reference phylogeny. Signature genes for a clade are those OGs that do not occur outside the clade, and are represented by at least one copy in every one of its daughters (i.e. two for a resolved node, and more than two for an unresolved node; e.g. OG1 for clade A in Figure 14). Using this approach, we identified 8,362 signature genes for 112 of the 128 clades (Table 6, Figure 15 and Supplementary Table 1 in the original article; submitted). We found no correlation of the number of signature genes with the number of daughters ($r = 0.07$), the number of species ($r = -0.06$), the bootstrap value of the clade ($r = 0.05$) or the distance to the root ($r = -0.01$). The clade with the most signature genes was Streptomyces (796 signature genes). When we restricted our search to perfect signature genes (i.e., present in every species within the clade), we identified 4,342 signatures for 98 clades (Table 6). Because for two-species clades, the daughters in which a gene is required are single species, all their

signatures are perfect. 2,972 perfect signature genes are a signature for two-species clades, 1,370 perfect signature genes are a signature for larger clades.

To compare our results to the signature genes found in previous studies, we assigned the latter genes to OGs using STRING (von Mering et al. 2007b). 213 of 241 Actinobacteridae signatures (Gao et al. 2006), 61 of 61 signatures specific for the Alphaproteobacteria (Kainth and Gupta 2005), 174 of 205 Chlamydiales signatures (Griffiths et al. 2006) and 181 of 181 signatures identified in the Cyanobacteria (Martin et al. 2003) could be assigned to an OG (Figure 19). Many of the genes identified previously were also found as signature genes in our approach, and mostly for the same taxon or, as was already observed by the cited authors, for a sub-clade (blue in Figure 19). In a few cases we found the genes as signatures for a higher level taxon (red), because the OGs were defined at a relatively low resolution, over all species simultaneously (Tatusov et al. 2000; von Mering et al. 2007b). By manually inspecting the results, these signature genes could be identified at a higher resolution in the small-scale studies, while our approach, based on large-scale automated orthology definitions, also reported the genes in other taxa.

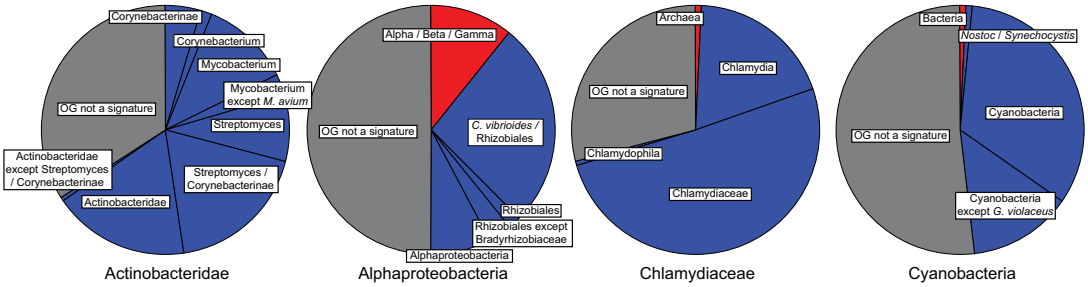

Figure 19. Identification of signature genes in four data sets obtained from the literature: 213 Actinobacteridae signature genes (Gao et al. 2006), 61 Alphaproteobacteria signature genes (Kainth and Gupta 2005); 174 Chlamydiaceae signature genes (Griffiths et al. 2006); and 181 Cyanobacteria signature genes (Martin et al. 2003). The genes were assigned to OGs using STRING (von Mering et al. 2007b). The fractions indicate how many of the genes assigned to an OG were a signature for the same clade or subclades (blue), a clade at a higher level (red) or not a signature (gray).

The coverage score is calculated as a nested coverage, a method that takes into account potential asymmetrical taxon sampling. For terminal clades, the score is equal to the coverage, i.e. the fraction of species containing the OG. For higher order clades, the score is the average of the score in its daughter clades. This is best illustrated with an example (Figure 14). The coverage score of OG1 as a signature for clade A is 0.72:

$$\frac{(1/1 + 2/3 + 1/2)}{3} = 0.72$$

Phylogenetic signal in gene repertoires

To assess whether the number of signature genes found for a clade is significant, we composed 1,000 sets of randomized genomes. Bearing in mind that the size distribution of both the OGs and the genomes is important for the identification of signature genes, we kept the number of OGs per genome identical, as well as the number of genomes in which an OG is represented. Because the phylogeny was not randomized, we could calculate the expected number of signature genes for a clade as the average for that exact same clade, with the same distribution of species sizes, over the 1,000 randomized genome sets. In these randomized data sets, we found an average of 1,667 signature genes of which only 74 had a coverage score >= 0.75, and 37 were perfect. These small numbers show the strong phylogenetic signal in the non-randomized gene repertoires.

Because our randomization procedure retained the structure of the phylogeny as well as the size distribution of the genomes, we could calculate an observed over expected ratio (o/e-ratio) for

each individual clade, based on the the number of signature genes found in the original data set and in the random gene repertoires. We observed that out of the 129 clades in the phylogeny, 103 contained more, and 24 contained less signature genes than expected (see Supplementary Table 1 in the original article; submitted). For the Chlamydophila pneumoniae clade and the Acidobacteria / Proteobacteria clade, no signature genes were found or expected based on the 1,000 randomized gene sets, for the *M. genitalium* / *M. pneumoniae* clade, 29 signature genes were found, but none expected. For the remaining 126 taxa, the average o/e-ratio was as high as 1321, which is indicative of the strong phylogenetic signal in the gene repertoires. If we applied a coverage score cutoff of 0.75, 104 clades contained more signature genes than expected, and for 41 clades, no signature genes were expected at all. Twelve clades contained less signature genes than expected, and for 13 clades no signature genes were found or expected.

A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation

Bas E. Dutilh, Martijn A. Huynen, and Berend Snel
BMC Genomics (2006) 7: 10

Abstract

Background

The massive scale of microarray derived gene expression data allows for a global view of cellular function. Thus far, comparative studies of gene expression between species have been based on the level of expression of the gene across corresponding tissues, or on the co-expression of the gene with another gene.

Results

To compare gene expression between distant species on a global scale, we introduce the "expression context". The expression context of a gene is based on the co-expression with all other genes that have unambiguous counterparts in both genomes. Employing this new measure, we show 1) that the expression context is largely conserved between orthologs, and 2) that sequence identity shows little correlation with expression context conservation after gene duplication and speciation.

Conclusions

This means that the degree of sequence identity has a limited predictive quality for differential expression context conservation between orthologs, and thus presumably also for other facets of gene function.

Background

The two main components of the function of a gene are its molecular function (what does it do, e.g. is it a hydrolase, is it DNA binding) and its functional context (with what other elements of the cell does it collaborate). Though both aspects can only be decisively determined in *in vivo* experiments, the incredible and increasing amount of experimental information assembled in databases enables more and more accurate predictions (von Mering et al. 2005). Because of the accuracy and speed with which algorithms can identify sequence similarity, the most commonly used tool for predicting gene function is doubtlessly sequence conservation. As the sequence is the blueprint for the three-dimensional structure, and therewith the enzymatic function of a gene, this method is particularly suitable for predicting the molecular function of an unknown gene, for example in a newly sequenced species.

Predicting functional context, on the other hand, is a different story. This means inferring *in silico* in which process the gene plays a role. Whereas the molecular function is concrete, and can be described by the catalyzed chemical reaction, the functional context is more elusive and may best be described as a composition of the context (e.g. binding partners) of the encoded protein and the regulation of its expression in time and space (Werner 2002). A way to estimate the functional context is in terms of the collection of cells or tissues and biological processes or circumstances that determine when the gene is expressed. DNA microarrays measure the expression levels of many genes under the same experimental condition, and combining the information from many such experiments allows the clustering of genes based on correlations in their expression patterns (Eisen et al. 1998). If two genes are co-expressed, i.e. they have a comparable expression profile, they are assumed to have a comparable functional context, independent of what this functional context is. Using co-expression as a function prediction tool is particularly powerful when the co-expression is conserved in different organisms (Bergmann et al. 2004; Snel et al. 2004; Stuart et al. 2003; van Noort et al. 2003).

Here, we introduce a method to take the step from the comparative study of expression evolution based on the pairwise co-expression between two genes, to a definition on a global level. We present the "expression context" of a gene, based not on the expression across a range of tissues or circumstances, but on the co-expression with a range of genes. If two genes are co-expressed with the same other genes, i.e. they have a comparable co-expression profile, they thus have a comparable expression context. Not only does this allow a global view on expression evolution, but it also solves the issue of comparing gene expression between distantly related species. When studying e.g. *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (van Noort et al. 2003), one can not assign equivalent tissues like between *Homo sapiens* and *Mus musculus* (Huminiecki and Wolfe 2004). The expression context method overcomes this limitation by substituting identical tissues for orthologous genes, and levels of expression for co-expression values. In this study, we include four Eukaryote species (*C. elegans*, *Drosophila melanogaster*, *H. sapiens* and *S. cerevisiae*), for which gene co-expression data have been determined on a large scale (Stuart et al. 2003). The first issue we address in this paper is how much our new global estimate of expression context is conserved between species.

In a comparative analysis of gene properties between different species, a solid definition of orthology is critical. Current state of the art orthology methods allow for the expansion of an orthologous gene pair in one or both of the species compared. The existence of these so called in-paralogs, raises the question to what extent the expression contexts of the gene copies have diverged. Previously, we have studied genes that are duplicated in *C. elegans* relative to *S. cerevisiae* (Snel et al. 2004). We showed that the *C. elegans* orthologs of genes that in *S. cerevisiae* are reliably co-regulated with the ancestral gene, have a tendency to retain co-expression with one of the two duplicated orthologs in *C. elegans*, while the link with the other is lost (partial conservation, Figure 3 in (Snel et al. 2004)). One of the important questions this paper left us with is whether the derived gene that had retained the

ancestral regulatory context was also the least diverged at the sequence level. Therefore, the second issue addressed in the current work is the relationship between the evolution of the gene sequence and the evolution of the expression context after a gene duplication. We present an analysis between orthologous groups (after speciation), and an analysis between sibling genes (in-paralogs) within expanded orthologous groups (after gene duplication), and show that sequence and expression context tend to diverge independently.

Results and discussion

Orthology

Inparanoid is a pairwise definition of orthology that allows for species specific gene expansions (in-paralogs, (Remm et al. 2001)). In the case of this group orthology, two or more genes from one species are evolutionarily equally orthologous to one or more genes in the other species. Such a scheme is necessary if we want to study the divergence in expression context between two recent gene copies, which would not be found in, for example, a reciprocal best hit approach. On the other hand, algorithms that identify group orthology between more organisms at once would annul the resolution obtained in a pairwise definition (Koonin et al. 2004). We constructed orthology relationships separately for all species pairs, and separated the resulting orthologous groups into two categories: 1-1 orthologous groups (if both species contain a single ortholog) and X-X orthologs (if at least one of the species contains more than one ortholog). There are about twice as many 1-1 orthologs as there are X-X orthologous groups (see Table 7).

Table 7. Inparanoid pairwise orthologous groups between all species pairs for *C. elegans* (15950 genes) *D. melanogaster* (4456 genes) *H. sapiens* (12193 genes) and *S. cerevisiae* (6199 genes).

species A	species B	total OGs	1-1 OGs
C. elegans	*D. melanogaster*	2393	1907
C. elegans	*H. sapiens*	3814	2335
C. elegans	*S. cerevisiae*	2520	1516
D. melanogaster	*H. sapiens*	2739	1891
D. melanogaster	*S. cerevisiae*	1641	1193
H. sapiens	*S. cerevisiae*	2514	1580
total		15621	10422

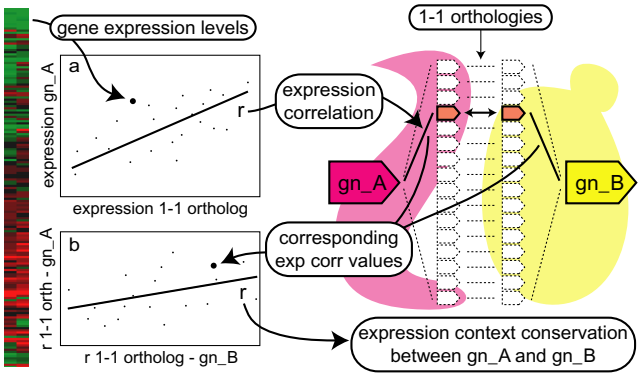

Figure 20. Method used to calculate the expression context conservation between gn_A and gn_B. Genes gn_A and gn_B are the query genes in species A and species B, respectively. First, the correlation between the expression levels of the query gene and all 1-1 orthologs over multiple microarray experiments was calculated in both species (a; uncentered correlation). The resulting expression correlation values were correlated between the two species (b; Pearson's correlation), yielding the expression context conservation between gn_A and gn_B. For an unambiguous comparison between species, we only analyze the expression correlation values of the studied genes with the 1-1 orthologs.

Expression context

The global definition of expression context introduced here is based on the expression correlations between a query gene in one species and all the members in that species of all 1-1 orthologous groups present between the two species compared (see Figure 20a). The expression context conservation is then obtained by correlating the expression correlation values of the query genes from two different species and the corresponding 1-1 orthologs in their species (see Figure 20b).
To test how meaningful this measure is, we compared the expression context conservation between different categories of orthologs and random non-orthologous gene pairs. The histograms in Figure 21 are normalized, and the data is pooled over all species comparisons. As a null model, we composed a random data set of 1000 non-orthologous gene pairs drawn from each species pair.

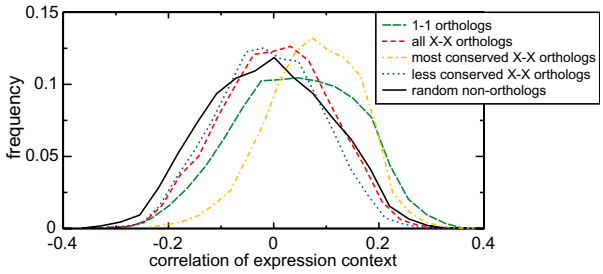

Figure 21. Expression context conservation between different classes of orthologs and random non-orthologous gene pairs. The plots are normalized histograms of the combined data from all species comparisons. For statistical comparison of the histograms see Table 8. The distributions are normally distributed (Shapiro-Wilk test, $P < 1 \cdot 10^{-4}$).

Though the distributions of the expression context conservation scores lie close to zero, we find that the expression context of both 1-1 orthologs and of X-X orthologs is significantly higher than that of random genes (see Figure 21, for P values see Table 8). This significant conservation reveals the functional and evolutionary relevance of the expression context.

Table 8. Probability that the expression context conservation scores in different classes of orthologs and random non-orthologous gene pairs were drawn from the same distribution (see histograms in Figure 21; P values, Student's t-test; the distributions are normal according to a Shapiro-Wilk test, $P < 1 \cdot 10^{-4}$). The expression context data is combined over all species comparisons: 1-1 orthologs (n = 10303) all X-X orthologs (n = 27147) most conserved X-X orthologs (n = 5180) less conserved X-X orthologs (n = 21967) random non-orthologous gene pairs (n = 6000).

	1-1 orth	most cons X-X	less cons X-X	random non-orth
all X-X orth	$6.31 \cdot 10^{-233}$	0	$1.78 \cdot 10^{-70}$	$3.55 \cdot 10^{-21}$
random non-orth	$9.66 \cdot 10^{-173}$	0	0.172	
less cons X-X	0	0		
most cons X-X	$1.38 \cdot 10^{-57}$			

Which genes have a conserved expression context?

We looked at the function of the genes with a conserved expression context using the KOG functional categories (Koonin et al. 2004). The functional categories were counted for all 1-1 orthologs assigned to a KOG (the genes were considered separately). For each functional category, the fraction of 1-1 orthologous genes with an expression context conservation score higher than zero is shown in Figure 22. We find that all "Information storage and processing" categories have a higher level of expression context conservation than all "Metabolism" categories. Within the "Cellular processes and signaling" class, which lies between the two extremes, we also find the categories with more informational genes to have a higher expression context conservation than those containing operational genes. "Nuclear structure" (Y) for example has a large fraction of genes with a highly conserved expression context, while "Cell wall/membrane/envelope biogenesis" (M) and "Extracellular structures" (W) have a low expression context conservation. These results are in accordance with other studies: the

conservation of co-expression has previously been shown to be high for genes involved in core informational cellular processes (specifically the ribosome and ribosome biogenesis (Stuart et al. 2003), as well as the GO biological process category "Metabolism", which harbors protein biosynthesis (Lefebvre et al. 2005)). Informational genes are also found to be more conserved than operational genes with respect to other properties, e.g. they have been shown to be less prone to horizontal gene transfer (Dutilh et al. 2004; Jain et al. 1999).

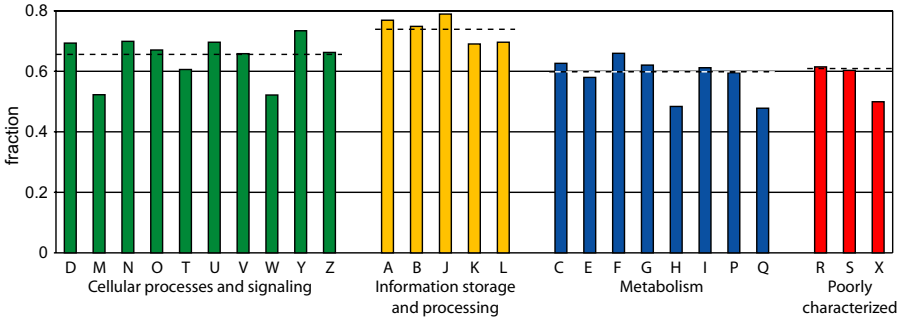

Figure 22. Functional classification of 1-1 orthologs with a conserved expression context (score higher than zero). From all species pairs, all 1-1 orthologs that could be assigned to a KOG were included. The categories are grouped in the four main KOG classes. The horizontal dashed lines are the fraction of genes with a conserved expression context for the entire class. The functional categories are (the number between brackets is the number of genes with a conserved expression context): "Cellular processes and signaling" (D: Cell cycle control, cell division, chromosome partitioning (n = 442), M: Cell wall/membrane/envelope biogenesis (n = 73), N: Cell motility (n = 23), O: Posttranslational modification, protein turnover, chaperones (n = 1330), T: Signal transduction mechanisms (n = 1151), U: Intracellular trafficking, secretion, and vesicular transport (n = 953), V: Defense mechanisms (n = 67), W: Extracellular structures (n = 111), Y: Nuclear structure (n = 96), and Z: Cytoskeleton (n = 378)), "Information storage and processing" (A: RNA processing and modification (n = 823), B: Chromatin structure and dynamics (n = 244), J: Translation, ribosomal structure and biogenesis (n = 1153), K: Transcription (n = 985), and L: Replication, recombination and repair (n = 545)), "Metabolism" (C: Energy production and conversion (n = 486), E: Amino acid transport and metabolism (n = 367), F: Nucleotide transport and metabolism (n = 205), G: Carbohydrate transport and metabolism (n = 452), H: Coenzyme transport and metabolism (n = 131), I: Lipid transport and metabolism (n = 383), P: Inorganic ion transport and metabolism (n = 228), and Q: Secondary metabolites biosynthesis, transport and catabolism (n = 71)) and "Poorly characterized" (R: General function prediction only (n = 1716), S: Function unknown (n = 912), and X: Not categorized by NCBI staff (n = 2)) (Koonin et al. 2004).

Differential expression context conservation between in-paralogs

Our previous work suggests that in an X-X orthologous group, the ancestral expression context may have been retained by one of the in-paralogs in each of the species (Snel et al. 2004), possibly because they are functionally the most conserved. We therefore sub-classify each X-X orthologous group into the gene pair that has the highest expression context conservation within this orthologous group on the one hand (we will refer to this gene pair as the "most conserved X-X orthologous gene pair"), and on the other hand the remaining, "less conserved X-X orthologs" (Figure 23).

Comparing the distribution of the expression context conservation scores in these sub-categories of orthologs with the other histograms in Figure 21 reveals that only the set of random gene pairs and the less conserved X-X orthologs do not have significantly different distributions ($P = 0.172$, Student's t-test; see Table 8). The expression context conservation in these two data sets was lowest, followed by, in order, all X-X orthologs, the 1-1 orthologs, and finally the most conserved X-X orthologs (see Figure 21). All the other pairs of distributions are highly significantly different from one another ($P \quad 3.55 \cdot 10^{-21}$, see Table 8).

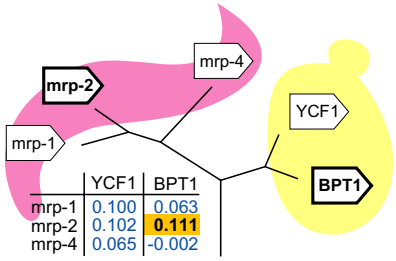

Figure 23. Example of an X-X orthologous group between *C. elegans* and *S. cerevisiae*. This X-X orthologous group (KOG0054: Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily) has three genes in *C. elegans* and two genes in *S. cerevisiae*. The expression context conservation scores are given in the table. The gene pair with the highest score is the "most conserved X-X orthologous gene pair" (yellow), the rest are the "less conserved X-X orthologs" (blue).

Correlation of sequence identity and expression context conservation between orthologous groups

To find out how the conservation of expression context (see Figure 21) is reflected in the sequence conservation, we first analyzed how the sequence divergence between orthologous groups relates to the divergence in expression context in an orthologous gene pair after speciation. To avoid having to make a potentially controversial choice on how to functionally and evolutionary interpret the multiple orthologous relationships in X-X orthologous groups (Snel et al. 2004), we only used the 1-1 orthologs for this comparison. These gene pairs originated at the speciation event, so they have all had the same amount of time to diverge. Table 9 presents the correlation coefficients between expression context conservation and sequence identity of the 1-1 orthologs for all species pairs.

Table 9. Correlation between sequence identity and expression context conservation for 1-1 orthologs between all species pairs. P is the probability that the data set is a sample drawn from a distribution with correlation coefficient zero.

species A	species B	correlation	P
C. elegans	*D. melanogaster*	0.077	$8.41 \cdot 10^{-4}$
C. elegans	*H. sapiens*	0.060	$4.49 \cdot 10^{-3}$
C. elegans	*S. cerevisiae*	0.121	$5.14 \cdot 10^{-6}$
D. melanogaster	*H. sapiens*	0.092	$6.27 \cdot 10^{-5}$
D. melanogaster	*S. cerevisiae*	0.050	$9.01 \cdot 10^{-2}$
H. sapiens	*S. cerevisiae*	0.061	$1.46 \cdot 10^{-2}$

Though the correlation coefficients are significantly positive ($P < 0.05$ for all species comparisons except DM-SC, where $P = 0.09$), they are very low (see Table 9). In this analysis of the relationship between expression context conservation and sequence identity across orthologous groups, we conclude that the evolution rate of the gene sequence does not depend on its expression context. A trend that we seem to observe is that the correlation between sequence evolution and expression context evolution reflects the predictive span of the expression data. In Figures 2 d-f of the paper by Stuart et al. (2003), the accuracy-coverage plots of *D. melanogaster* and *H. sapiens* are always lower than those of *C. elegans* and *S. cerevisiae*. In our results, we also observe the highest correlation between expression context conservation and sequence identity for the 1-1 orthologs of *S. cerevisiae* and *C. elegans*, rather than for two closer related Metazoa. Thus some of the variation in our results reflect the quality of the microarray data for function prediction.

Correlation of sequence identity and expression context conservation between orthologs after a single gene duplication

The simplest case where we can study the divergence of duplicated genes within orthologous groups is for 1-2 orthologs, where one gene duplication occurred in one of the two daughter

species since the speciation event. We carry out a straightforward analysis by counting how often the gene with the highest expression context conservation also has the highest sequence identity. Figure 24 shows the consistency of sequence evolution with expression context evolution in the 1-2 orthologous groups.

It is immediately striking how little difference there is between the observed consistent and observed inconsistent bars in Figure 24. For all species comparisons, there is no significant over-representation of consistent observations, apart for a few exceptions (CE1-HS2 orthologs (i.e. 1 ortholog in *C. elegans* and 2 orthologs in *H. sapiens*, other abbreviations are composed similarly) and HS1-SC2 orthologs; $P < 0.05$, binomial distribution). In general, all the P values are very high, so this analysis shows that for 1-2 orthologs, the expression context is not better conserved in the ortholog with the highest sequence identity.

Given the large overlap between the expression context conservation scores of the most conserved X-X orthologous gene pair and the less conserved X-X orthologs (see Figure 21), a substantial fraction of inconsistent cases is expected based on this overlap alone. We therefore examined whether

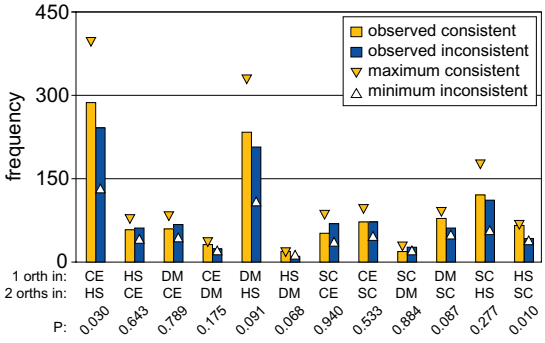

Figure 24. Consistency of sequence divergence with divergence in expression context for simple duplications. Consistency or inconsistency of sequence divergence with divergence in expression context for orthologous groups with a single gene duplication (1-2 orthologs). We display both the observed frequencies (plotted are the number of 1-2 orthologous groups; P is the probability to find at least this number of consistent observations by chance, binomial distribution) and the maximum consistent and minimum inconsistent frequencies expected (horizontal edge of the triangles), based on a completely consistent re-allocation of the expression context conservation scores from the overlapping distributions (see Methods).

the small differences between the observed consistent and inconsistent frequencies in Figure 24 resulted from this overlap. To do this, we split the expression context conservation scores of all 1-2 orthologous groups into two data sets: one containing the highest (most conserved) expression context conservation scores, the other containing the lower (less conserved) scores. We computed the expected maximum consistent and minimum inconsistent observations by drawing from these data sets consistently with the sequence conservation (see Methods). The triangles in Figure 24 show that many more consistent observations are expected if the data was initially organized consistently, even when the distributions of the most conserved and the less conserved X-X orthologs have such a large overlap.

In this analysis, we observed that the difference in sequence identity for the two duplicated genes was often small. This may in part be due to the fact that we compare evolutionarily divergent species, where the differences between in-paralogs (within species) are small relative to the differences between orthologs (between species). To be able to compare the rate of sequence evolution more accurately, we studied in detail the CE1-SC2 orthologous groups, and included the genome of *Ashbya gossypii*, a fungus closely related to *S. cerevisiae*. Where we found an AG1-SC2 orthologous group consisting of the same two *S. cerevisiae* genes as in the accompanying CE1-SC2 orthologous group, we calculated the K_a/K_s ratio between both gene pairs in the AG1-SC2 orthologous group to determine the rate of evolution for both *S. cerevisiae* genes. The ratio of nonsynonymous (K_a) to

synonymous (K_s) nucleotide substitution rates is an indicator of selective pressures on genes (Hurst 2002): a ratio higher than one indicates genes that are under positive selection pressure to change their sequence, a ratio lower than one indicates stabilizing selection. We found that the expression context was conserved for the slowest evolving *S. cerevisiae* gene in no more than 50% of the cases. These results confirm that gene sequence and expression context evolve independently after a gene duplication in 1-2 orthologous groups.

Diverged expression contexts in the two β-subunits of the Nascent polypeptide-Associated Complex in *S. cerevisiae*

As an example, we have looked in detail at a pair of in-paralogs in *S. cerevisiae* with a large difference in expression context conservation: β_1NAC (EGD1) and β_3NAC (BTT1). This example was selected because the in-paralogs in *S. cerevisiae* have an especially large difference in expression context conservation relative to *C. elegans* (for this species pair, the microarray data had the highest predictive relevance of all our species comparisons; see paragraph "Correlation of sequence identity and expression context conservation between orthologous groups" and Figures 2 d-f in (Stuart et al. 2003)). In general, one should be alert when interpreting microarray data for a particular gene. For example, its spot may not hybridize well and the level of expression, co-expression or even expression context of the gene will be correspondingly influenced. We therefore checked these two genes and found that they behave normally: the fraction of experiments where they are over- and under-expressed is comparable to that of average genes (not shown).

The β-subunit of the Nascent polypeptide-Associated Complex (βNAC) is represented by two copies in *S. cerevisiae*: β_1NAC (EGD1) and β_3NAC (BTT1) (Hu and Ronne 1994; Rospert et al. 2002). Other species have only one copy of this gene: icd-1 in *C. elegans*, bic in *D. melanogaster* and BTF3 in *H. sapiens*. Comparing the expression context of each of these three genes to the two *S. cerevisiae* genes revealed that for all species comparisons, the expression context of EGD1 was highly conserved, while the expression context of BTT1 had diverged (see Table 10). Compared to icd-1 in *C. elegans*, the expression context correlation of BTT1 was even negative. When we compare the sequence identity of the two genes with their single orthologs in the other three species in this study, we find indeed that BTT1 is more diverged than EGD1 in all cases (see Table 10), i.e. sequence divergence and expression context divergence are completely consistent.

Table 10. Sequence identity and expression context conservation of the two βNAC in-paralogs in S. cerevisiae. The β-subunit of the Nascent polypeptide-Associated Complex has two orthologs in S. cerevisiae: Enhanced Gal4 DNA binding protein 1 (EGD1, β_1NAC) and Basic Transcription factor Three 1 (BTT1, β_3NAC). The three other species in this analysis have only one ortholog: inhibitor of cell death 1 (icd-1 in *C. elegans*), bicaudal (bic in *D. melanogaster*) and Basic Transcription Factor 3 (BTF3 in *H. sapiens*).

		C. elegans icd-1	*D. melanogaster* bic	*H. sapiens* BTF3
S. cerevisiae EGD1	identity	0.385	0.350	0.375
S. cerevisiae EGD1	exp. cont.	0.302	0.203	0.199
S. cerevisiae BTT1	identity	0.300	0.305	0.340
S. cerevisiae BTT1	exp. cont.	-0.205	-0.092	0.006

The function of these two gene copies remains unclear. So far, the only difference in function found for these two genes comes from deletion experiments. Disruption of either of the *S. cerevisiae* βNAC copies yielded viable strains, that differ only in the level of GAL1 and GAL10 induction after transmission to a medium containing galactose in stead of glucose (Hu and Ronne 1994). The cross bred double negative βNAC mutant showed an increase in the expression of several genes, including the GAL genes. Hu and Ronne (1994) suggested that EGD1 and BTT1 have a redundant function, but based on the diverged expression context, it is likely that the two genes are expressed under highly divergent cellular circumstances. Given the consistent hints from the differential conservation of both the expression context and the protein sequence, we predict that EGD1 is the true ortholog of icd-1, bic and BTF3.

Correlation of sequence identity and expression context conservation within orthologous groups after multiple gene duplications

We also compared sequence conservation with expression context conservation in more expanded X-X orthologous groups, i.e. all orthologous groups with four or more genes in two species. Here, we considered sequence identity and expression context conservation consistent if they are positively correlated over all the gene pairs within an X-X orthologous group, and inconsistent when they are negatively correlated (note that carrying out this analysis on the 1-2 orthologs would give the same results as in the paragraph "Correlation of sequence identity and expression context conservation between orthologs after a single gene duplication").

Figure 25 shows that these results and the results of the analysis of simple duplications (Figure 24) are very comparable. In almost all species comparisons, there is no significant difference between the number of consistent and inconsistent observations ($P < 0.05$, binomial distribution, except CE-HS orthologs where $P = 0.018$). The predominantly inconsistent X-X orthologous groups between *D. melanogaster* and *H. sapiens* may be the result of the lower predictive relevance of the expression data in these species (as mentioned in the paragraph "Correlation of sequence identity and expression context conservation between orthologous groups").

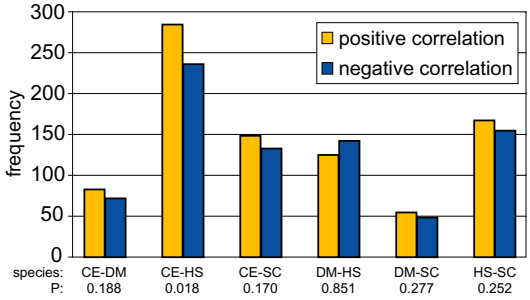

Figure 25. Consistency of sequence divergence with divergence in expression context for expanded orthologous groups. Consistency (positive correlation) or inconsistency (negative correlation) of sequence divergence with divergence in expression context for all expanded orthologous groups (X-X orthologs, except 1-2 orthologs). Plotted frequencies are the number of X-X orthologous groups with a positive and negative correlation. P is the probability to find at least this number of positively correlated observations by chance (binomial distribution).

If in both species the most conserved X-X orthologs are the only two genes with a selective constraint to maintain the ancestral function, the less conserved X-X orthologs may diverge randomly. Thus, it is possible that the negative correlation between sequence identity and expression context conservation in the whole X-X orthologous group arose by chance. For those X-X orthologous groups with a negative correlation, we therefore checked if there was one gene pair that harbored both the highest expression context conservation and the highest sequence identity. However, this was the case for only 10% of these inconsistent X-X orthologous groups, so we must conclude that their negative correlation between sequence identity and expression context conservation is not the result of one of the X-X orthologous gene pairs being conserved, and the rest of the genes diverging randomly. Rather, the conclusion is that as in 1-2 orthologs, the sequence and the expression context also evolve independently in other, more expanded X-X orthologous groups.

Conclusions

In this paper, we introduce a global definition of expression context based on gene expression data. As equivalent tissues or experiments can not be assigned between distantly related species, our method uses orthologous genes to define convertible expression contexts between species.

We represent the expression context of a query gene as the co-expression profile with a range of genes, rather than as the expression profile across corresponding experimental conditions. Though the microarrays were carried out under highly divergent conditions in the four Eukaryotes in this study (see Figure 1b in (Stuart et al. 2003)), the expression context of one gene is based on many expression correlation values, each of which in turn integrates a large collection of experiments. To test the coverage and homogeneity of the experimental data sets, we calculated the expression correlation values of all gene pairs separately over two random halves of the microarray experiments. In *D. melanogaster* (r = 0.91) and *S. cerevisiae* (r = 0.79), these scores were highly correlated (the correlation was not calculated for *C. elegans* and *H. sapiens* as these data sets were very large). Thus, we do not expect biases in the microarray experimental conditions to severely influence the correlations in expression context. Application of our method reveals that the expression context is conserved between orthologs across all species pairs, though X-X orthologs are less well conserved than 1-1 orthologs (see Figure 21). We also find that informational genes have a more conserved expression context than operational genes (see Figure 23). Taken together, these results show that the expression context presented here is a meaningful measure of the global expression context of a gene.

Using this method, we analyzed the correlation between the rates of evolution of the protein sequence and of the expression context. A correlation might be expected if the selective constraints on sequence and expression context were linked. In a comparison between all unexpanded orthologous groups, we find that this correlation is very low (see Table 9). This analysis compares genes that have branched apart at the speciation event, which means all differences in sequence conservation or expression context conservation are due to orthologous group specific evolution rates. Because of the wide range of functions carried out by the different orthologous groups, it is likely that there are also differences in the evolution rates between orthologous groups. To eliminate the possible resulting biases in the comparison between orthologous groups, we have also compared the rates of sequence and expression context evolution within orthologous groups, i.e. after one (1-2 orthologous groups) or multiple (X-X orthologous groups) gene duplication events. In these analyses, not all genes in one comparison have originated at the same time, but biases due to orthologous group specific evolution rates are absent. Still, the conclusions are the same as in the comparison between orthologous groups. For 1-2 orthologs as well as for the other X-X orthologs, the cases where sequence identity and expression context conservation were correlated were not significantly over-represented (see Figure 24 and Figure 25). The only species pair with significantly more consistent observations in both analyses was *C. elegans* and *H. sapiens*, though only the CE1-HS2 and not the HS1-CE2 orthologs were consistent. Comparing the types of microarray experiments carried out in these two species shows that there is little overlap (Stuart et al. 2003). Nonetheless, these species are almost the only pair with a significant over-representation of consistency between sequence identity and expression context conservation.

The methods employed in this research show that the expression context is conserved in orthologs between species. Sequence identity and expression context conservation are not correlated after gene duplication. Thus, annotation of different expression contexts to orthologs can not be based on sequence similarity alone.

Many of the expression correlations that compose the expression context may be irrelevant. According to the global definition of expression context introduced here, the expression correlation scores of all 1-1 orthologs in the genome add to the expression context. As few genes will possess a functional network containing all 1-1 orthologs, many co-expression values in the vector defining the expression context may be irrelevant. As an alternative, we have therefore also performed all analyses presented in this research using another method, that defined the expression context conservation as the number of overlapping orthologous groups in the top 100 co-expressed 1-1 orthologs between two genes. In other words, this method counts how many of the highly co-expressed 1-1 orthologs are shared between two genes. Qualitatively, the results found using

this alternative method were identical, indicating a robustness of the results to different definitions of expression context.

Previously, we have shown that after a gene duplication, one of the in-paralogs has a tendency to keep the ancestral regulatory interaction, while this link is lost in the other (Snel et al. 2004). We could not find evidence for such partial conservation using the global definitions of functional conservation introduced here. In other words, although reliably predicted co-regulatory links are asymmetrically conserved after gene duplication, the co-expression of in-paralogs remains similar from a global point of view. This can be explained if the divergence (which we observe studying pairwise links) indicates sub-functionalization, while the in-paralogs remain within in the same cellular process (resulting in a similar global expression context).

Methods

Data

The expression correlation of more than 326 million gene pairs over a large number of DNA microarrays in *C. elegans*, *D. melanogaster*, *H. sapiens* and *S. cerevisiae* (Stuart et al. 2003) was calculated using uncentered correlation (see Figure 20a). We used this data set as is, because it is the largest uniform collection of gene expression data available for Eukaryotes. The genomes were downloaded from Wormbase for *C. elegans* (Chen et al. 2005), Flybase for *D. melanogaster* (Drysdale et al. 2005), Refseq for *H. sapiens* (Pruitt et al. 2005) and the Saccharomyces Genome Database for *S. cerevisiae* (Christie et al. 2004). The genome of *A. gossypii* was downloaded from the Ashbya Genome Database (Dietrich et al. 2004).

Similarity and orthology

We searched the genomes for homologs using the Smith-Waterman P algorithm (Smith and Waterman 1981) on a TimeLogic DeCypher in all query-database combinations (matrix: Blosum62; e-value cutoff: 100). In the case of spurious asymmetries in the similarity search (e.g. two sequences giving different alignments depending on which was the query), the results are the average of two values, including both reciprocal experiments. Inparanoid (Remm et al. 2001) was run on the search results (default parameters; score cutoff: 50; outgroup cutoff: 50; sequence overlap cutoff: 0.5; confidence cutoff: 0.05; group overlap cutoff: 0.5; gray zone: 0). We only included genes in the orthology analysis if microarray data was available. For each pair of species, the 1-1 orthologous groups (one ortholog in each species, see Table 7) were used to define the expression context of a gene (see below and Figure 20). The rest of the orthologous groups were considered gene expansions (X-X orthologous groups, with more than one ortholog in at least one of the species). There are about twice as many 1-1 orthologs as there are X-X orthologous groups (see Table 7).

Expression context

The expression context of a gene was estimated using the co-expression values with the other genes in the genome. To be able to make an unambiguous comparison between two species, we only used the co-expression values with the 1-1 orthologs (see Figure 20b). We only included 1-1 orthologs in the list if we had co-expression data available in both species. The expression context conservation between two genes is defined as Pearson's correlation coefficient between the two vectors with co-expression values with the 1-1 orthologs.

The expected level of consistency between the sequence identity and the expression context conservation in a completely consistent set of 1-2 orthologs was calculated by separating the expression context conservation scores into two data sets. One contained the highest expression context correlation score in each 1-2 orthologous group (most conserved 1-2 orthologs, cf. Figure 23), the other contained the lower scores (less conserved 1-2 orthologs). We then randomly assigned the values from the high, most conserved data set to the 1-2 orthologous pairs with the highest

sequence identity, and the values from the low, less conserved data set to the 1-2 orthologous pairs with the lowest sequence identity, and counted the consistent cases. Thus, all orthologous groups were consistent in principle, and inconsistent observations can result only from the overlap of the distributions of the expression context conservation scores (cf. Figure 21). The numbers found (triangles in Figure 24) are thus the maximum expected number of consistent observations and the minimum expected number of inconsistent observations if the data would have been completely consistent, given the overlapping distributions.

KOG classification

The list of KOGs (euKaryotic clusters of Orthologous Groups of proteins) with assigned genes was downloaded from the COG website (Koonin et al. 2004).

K_a/K_s ratio

The K_a/K_s ratio was calculated using the kaks function of the seqinr package of the R Project for Statistical Computing (www.r-project.org). This function makes an unbiased estimate of the ratio of nonsynonymous (K_a) to synonymous (K_s) nucleotide substitution for a set of aligned sequences (Li 1993).

Authors' contributions

BED carried out the analyses, participated in the design and drafted the manuscript. MAH and BS conceived of the study, participated in the design and coordination and helped to draft the manuscript.

Acknowledgements

We thank Marc van Driel and Ludo Pagie for technical assistance.

Discussion

In this thesis, I have investigated whole genome data to learn about evolution. The many genomes that have been sequenced in recent years allow us to analyse and compare the complete genetic basis of living organisms, and the opportunities that data of this breadth present are unprecedented. In the past few years, I have been developing and using several methods to analyze aspects of evolution, that depend on the availability of complete genomes. Questions that I have been addressing include the following: what are the evolutionary relationships between species? How can we best use complete genome data to infer these relationships (chapter "Assessment of phylogenomic and orthology approaches for phylogenetic inference")? What is the influence of evolutionarily discordant processes, such as horizontal gene transfer, on the relationships between the gene contents of related organisms (chapter "The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise")? Can gene content be used to taxonomically identify incomplete genomes or environmental sequencing data (chapter "Signature genes as a phylogenomic tool")? And, comparing the functions of genes between species: does the expression context of a pair of orthologous genes correlate with their sequence similarity (chapter "A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation")?

While developing approaches to answer these questions, I came across a variety of interesting side results, but I also identified new open questions. One of the results I kept finding was the evolutionary stability of informational genes relative to metabolic genes. This has been hypothesized to reflect the more frequent occurrence of informational genes in complexes, whereas metabolic genes supposedly tend to function alone (Jain et al. 1999). Another result that jumps out of the data time and again is the consistency of the tree of life. It does not matter where you look, the evolutionary signal keeps presenting itself: not only the gene sequence, but also gene content is more alike between closely related species than between distant relatives. However, a question that is winning more and more concern is how these evolutionary relationships should be interpreted.

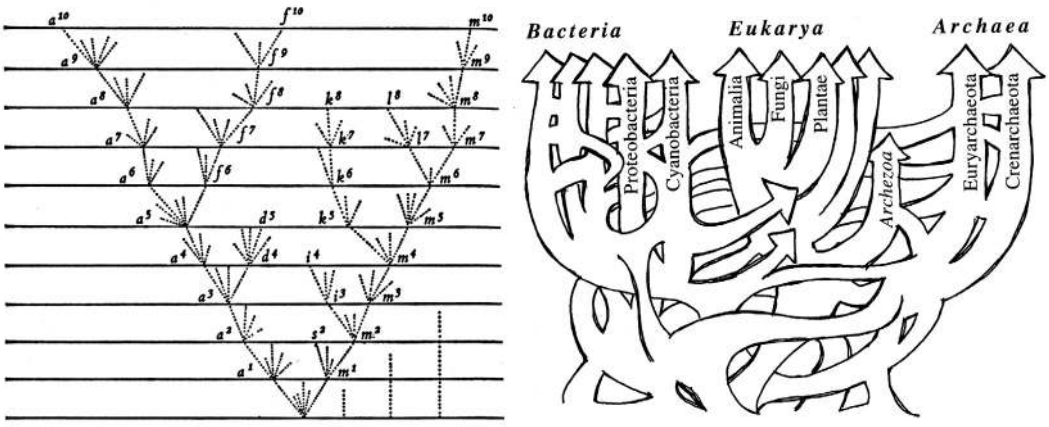

Figure 26. Left: "the affinities of all the beings of the same class have sometimes been represented by a great tree. The green and budding twigs may represent existing species; and those produced during former years may represent the long succession of extinct species" (Figure in Darwin 1859). Right: a reticulated tree, or net, which might more appropriately represent life's history (Figure 3 in Doolittle 1999a).

Since Darwin's insightful depiction of his idea about the tree-like structure of evolution (Figure 26, left; Darwin 1859), the dogma has been that evolution is by and large a vertical process. More recently, however, Doolittle pointed out that especially for prokaryotes, a tree may be insufficient to illustrate the complex evolutionary paths that led to the current-day genomes (Figure 26, right;

Doolittle 1999a). Does the diversity of trees in the phylome (i.e. the collection of all gene phylogenies) reflect noise and biases in the phylogenetic signal that have been accumulating like dust over the ages? Or are the methods for phylogenetic inference accurate enough, and do all these genes have truly different ancestries? In that case, a species tree would be an inadequate representation of the evolutionary relationships between species, and one might imagine that a cobweb might more accurately capture their interconnections (Ge et al. 2005). Personally, I could agree that a tree may fall short as a representation of the evolutionary relationships between genomes, but at the same time, describing a species as its entire genome can blur your vision. To characterize a species, I think one should look at its core, the essence of the species, and disregard confusing noise that obscures its evolutionary history (chapter "The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise"). This may also be part of the reason that gene content phylogenies, that are based on the entire genome, are less reliable than sequence similarity-based phylogenomic trees, that allow a careful selection of the meaningful sequences and characters (chapter "Assessment of phylogenomic and orthology approaches for phylogenetic inference").

Another type of data that has become available on a large scale are expression data. Gene expression *per se* can hardly be compared between species, as this requires completely harmonized experimental conditions, and it is expected to evolve very fast. However, the co-expression of two genes does contain a signal (van Noort et al. 2003), albeit a functional and not a phylogenetic signal, and this also has a lot to do with the experimental quality of the expression data. To be able to compare the functional context of genes in distantly related species, we developed the expression context, which relies on the completeness of the genome sequence, as well as on the availability of genome-wide expression experiments, which thus far are only available for a small selection of species that are distantly related.

The value and statistical significance of complete genome-based analyses increase as the number of sequenced genomes grows. But with the hundreds of genomes that are sequenced throughout all the taxa in the tree of life, there are also completely novel opportunities that present themselves. In the chapter "Signature genes as a phylogenomic tool", I have combined the gene content information from species in many taxa, and for each taxon identified signature genes that characterize that clade. As these genes remain confined to a single taxon, they evolve slowly at the gene content level, and may thus have retained better the evolutionary signal than have the remaining genes in the genome (this is in accordance with the results found in the chapter "The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise"). I show that signature genes are a useful tool for the taxonomic characterization of a sequenced sample, for example an environmental sample.

One of the questions that could still be addressed is whether the signature genes, that are evolutionarily stable at the gene content level, are also more stable at the sequence level. If that is the case, then using the signature genes we identified at a range of levels in the tree of life, we could improve the resolution and reliability of a sequence similarity-based tree. The evolutionary signal in each signature could be used to resolve the branching order in its specific clade, which I expect will yield a highly accurate phylogeny.

Furthermore, the signature approach could be expanded to include other taxon-specific properties as well. Recently, large-scale sequencing projects have entered a new era with the cheap 454 technique that yields enormous amounts of relatively short pieces of sequence (Margulies et al. 2005). As they are so short, these fragments are more difficult to assemble, but I expect that they can still contain enough signal to characterize the species composition in the sample using their taxon specificity.

References

Ahmad, S., A. Selvapandiyan, and R.K. Bhatnagar. 1999. A protein-based phylogenetic tree for gram-positive bacteria derived from hrcA, a unique heat-shock regulatory gene. Int J Syst Bacteriol 49 Pt 4: 1387-1394.

Andersson, J.O. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62: 1182-1197.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M.D. Gelpke, J. Roach, T. Oh, I.Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S.F. Smith, M.S. Clark, Y.J. Edwards, N. Doggett, A. Zharkikh, S.V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y.H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310.

Aravind, L., R.L. Tatusov, Y.I. Wolf, D.R. Walker, and E.V. Koonin. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 14: 442-444.

Arvestad, L., A.C. Berglund, J. Lagergren, and B. Sennblad. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19 Suppl 1: I7-I15.

Baldauf, S.L. and J.D. Palmer. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A 90: 11558-11562.

Bansal, A.K. and T.E. Meyer. 2002. Evolutionary analysis by whole-genome comparisons. J Bacteriol 184: 2260-2272.

Bapteste, E., Y. Boucher, J. Leigh, and W.F. Doolittle. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol 12: 406-411.

Barns, S.M., R.E. Fundyga, M.W. Jeffries, and N.R. Pace. 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci U S A 91: 1609-1613.

Berbee, M.L., D.A. Carmean, and K. Winka. 2000. Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough? Mol Phylogenet Evol 17: 337-344.

Bergmann, S., J. Ihmels, and N. Barkai. 2004. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2: E9.

Bininda-Emonds, O.R.P. 2004. The evolution of supertrees. Trends in Ecology & Evolution 19: 315-322.

Blanchette, M., T. Kunisawa, and D. Sankoff. 1996. Parametric genome rearrangement. Gene 172: GC11-17.

Boore, J.L. and W.M. Brown. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8: 668-674.

Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. 1998. Predicting function: from genes to genomes and back. J Mol Biol 283: 707-725.

Brinkmann, H. and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16: 817-825.

Brochier, C., S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol 6: R42.

Brochier, C. and H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417: 244.

Brown, J.R., C.J. Douady, M.J. Italia, W.E. Marshall, and M.J. Stanhope. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet 28: 281-285.

Bruno, W.J. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol 13: 1368-1374.

Bruno, W.J., N.D. Socci, and A.L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17: 189-197.

Cambillau, C. and J.M. Claverie. 2000. Structural and genomic correlates of hyperthermostability. J Biol Chem 275: 32383-32386.

Canback, B., S.G. Andersson, and C.G. Kurland. 2002. The global phylogeny of glycolytic enzymes. Proc Natl Acad Sci U S A 99: 6097-6102.

Canback, B., I. Tamas, and S.G. Andersson. 2004. A phylogenomic study of endosymbiotic bacteria. Mol Biol Evol 21: 1110-1122.

Cannone, J.J., S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Muller, N. Pande, Z. Shang, N. Yu, and R.R. Gutell. 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs: Correction. BMC Bioinformatics 3: 15.

Casjens, S. 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49: 277-300.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540-552.

Castresana, J., G. Feldmaier-Fuchs, S. Yokobori, N. Satoh, and S. Paabo. 1998. The mitochondrial genome of the hemichordate Balanoglossus carnosus and the evolution of deuterostome mitochondria. Genetics 150: 1115-1123.

Cavalier-Smith, T. 1986. The kingdoms of organisms. Nature 324: 416-417.

Cavalier-Smith, T. 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 52: 7-76.

Charlebois, R.L. and W.F. Doolittle. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 14: 2469-2477.

Chen, N., T.W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C.K. Chen, W.J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E.M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P.W. Sternberg, and L.D. Stein. 2005. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res 33: D383-D389.

Christie, K.R., S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J.M. Cherry. 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Res 32 Database issue: D311-314.

Ciccarelli, F.D., T. Doerks, C. von Mering, C.J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283-1287.

Clarke, G.D., R.G. Beiko, M.A. Ragan, and R.L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol 184: 2072-2080.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71-76.

Creevey, C.J. and J.O. McInerney. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21: 390-392.

Criscuolo, A., V. Berry, E.J.P. Douzery, and O. Gascuel. 2006. SDM: a fast distance-based approach for (super)tree building in phylogenomics. Syst Biol: In press.

Cunningham, C.W., H. Zhu, and D.M. Hillis. 1998. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52: 978-987.

Dagan, T. and W. Martin. 2006. The tree of one percent. Genome Biol 7: 118.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.

Darwin, C. 1859. The Origin of Species by Means of Natural Selection. Murray, London.

Daubin, V., M. Gouy, and G. Perriere. 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform Ser Workshop Genome Inform 12: 155-164.

Daubin, V., M. Gouy, and G. Perriere. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12: 1080-1090.

Daubin, V., N.A. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301: 829-832.

Dean, R.A., N.J. Talbot, D.J. Ebbole, M.L. Farman, T.K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.R. Xu, H. Pan, N.D. Read, Y.H. Lee, I. Carbone, D. Brown, Y.Y. Oh, N. Donofrio, J.S. Jeong, D.M. Soanes, S. Djonovic, E. Kolomiets, C. Rehmeyer, W. Li, M. Harding, S. Kim, M.H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L.J. Ma, R. Nicol, S. Purcell, C. Nusbaum, J.E. Galagan, and B.W. Birren. 2005. The genome sequence of the rice blast fungus Magnaporthe grisea. Nature 434: 980-986.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361-375.

Di Giulio, M. 2006. Nanoarchaeum equitans is a living fossil. J Theor Biol 242: 257-260.

Dietrich, F.S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi, R.A. Wing, A. Flavier, T.D. Gaffney, and P. Philippsen. 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304: 304-307.

Diezmann, S., C.J. Cox, G. Schonian, R.J. Vilgalys, and T.G. Mitchell. 2004. Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis. J Clin Microbiol 42: 5624-5635.

Dobzhansky, T. 1973. Nothing in Biology Makes Sense Except in the Light of Evolution. American Biology Teacher 35: 125-129.

Doolittle, W.F. 1999a. Lateral gene transfer, genome surveys, and the phylogeny of Prokaryotes. Science 286: 1443a.

Doolittle, W.F. 1999b. Phylogenetic classification and the universal tree. Science 284: 2124-2129.

Driskell, A.C., C. Ane, J.G. Burleigh, M.M. McMahon, C. O'Meara B, and M.J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306: 1172-1174.

Drysdale, R.A., M.A. Crosby, W. Gelbart, K. Campbell, D. Emmert, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhang, P. Zhou, M. Zytkovicz, M. Ashburner, A. de Grey, R. Foulger, G. Millburn, D. Sutherland, C. Yamada, T. Kaufman, K. Matthews, A. DeAngelo, R.K. Cook, D. Gilbert, J. Goodman, G. Grumbling, H. Sheth, V. Strelets, G. Rubin, M. Gibson, N. Harris, S. Lewis, S. Misra, and S.Q. Shu. 2005. FlyBase: genes and gene models. Nucleic Acids Res 33 Database Issue: D390-395.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.F. Richard, M.L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J.L. Souciet. 2004. Genome evolution in yeasts. Nature 430: 35-44.

Dutilh, B.E., C.E. Dutilh and W.H.M.M. van Laarhoven. 2001. Initiatives on Sustainable Development in the Food Sector Worldwide. Foundation for Sustainability in the Food Chain (DuVo).

Dutilh, B.E. and R.J. de Boer. 2003. Decline in excision circles requires homeostatic renewal or homeostatic death of naive T cells. J Theor Biol 224: 351-358.

Dutilh, B.E., M.A. Huynen, W.J. Bruno, and B. Snel. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58: 527-539.

Dutilh, B.E., V. van Noort, R.T. van der Heijden, T. Boekhout, B. Snel, and M.A. Huynen. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics 23:815-824.

Edgar, R.C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

Edgar, R.C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

Farris, R.J. 1977. Phylogenetic analysis under Dollo's law. Syst Zool 26: 77-88.

Fell, J.W., T. Boekhout, A. Fonseca, G. Scorzetti, and A. Statzell-Tallman. 2000. Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis. Int J Syst Evol Microbiol 50 Pt 3: 1351-1371.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19: 99-113.

Fitch, W.M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science 155: 279-284.

Fitz-Gibbon, S.T. and C.H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27: 4218-4222.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, and et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512.

Forterre, P., C. Bouthier De La Tour, H. Philippe, and M. Duguet. 2000. Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet 16: 152-154.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, R.D. Fritchman, J.F. Weidman, K.V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T.R. Utterback, D.M. Saudek, C.A. Phillips, J.M. Merrick, J.F. Tomb, B.A. Dougherty, K.F. Bott, P.C. Hu, T.S. Lucier, S.N. Peterson, H.O. Smith, C.A. Hutchison, 3rd, and J.C. Venter. 1995. The minimal gene complement of Mycoplasma genitalium. Science 270: 397-403.

Gabaldon, T. and M.A. Huynen. 2003. Reconstruction of the proto-mitochondrial metabolism. Science 301: 609.

Galagan, J.E., S.E. Calvo, K.A. Borkovich, E.U. Selker, N.D. Read, D. Jaffe, W. FitzHugh, L.J. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S.G. Wang, C.B. Nielsen, J. Butler, M. Endrizzi, D.Y. Qui, P. Ianakiev, D.B. Pedersen, M.A. Nelson, M. Werner-Washburne, C.P. Selitrennikoff, J.A. Kinsey, E.L. Braun, A. Zelter, U. Schulte, G.O. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg, A. Roy, K. Foley, J. Naylor, N. Stabge-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvysselis, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R.L. Metzenberg, D.D. Perkins, S. Kroken, C. Cogoni, G. Macino, D. Catcheside, W.X. Li, R.J. Pratt, S.A. Osmani, C.P.C. DeSouza, L. Glass, M.J. Orbach, J.A. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seller, J. Dunlap, A. Radford, R. Aramayo, D.O. Natvig, L.A. Alex, G. Mannhaupt, D.J. Ebbole, M. Freitag, I. Paulsen, M.S. Sachs, E.S. Lander, C. Nusbaum, and B. Birren. 2003. The genome sequence of the filamentous fungus Neurospora crassa. Nature 422: 859-868.

Galagan, J.E., S.E. Calvo, C. Cuomo, L.J. Ma, J.R. Wortman, S. Batzoglou, S.I. Lee, M. Basturkmen, C.C. Spevak, J. Clutterbuck, V. Kapitonov, J. Jurka, C. Scazzocchio, M. Farman, J. Butler, S. Purcell, S. Harris, G.H. Braus, O. Draht, S. Busch, C. D'Enfert, C. Bouchier, G.H. Goldman, D. Bell-Pedersen, S. Griffiths-Jones, J.H. Doonan, J. Yu, K. Vienken, A. Pain, M. Freitag, E.U. Selker, D.B. Archer, M.A. Penalva, B.R. Oakley, M. Momany, T. Tanaka, T. Kumagai, K. Asai, M. Machida, W.C. Nierman, D.W. Denning, M. Caddick, M. Hynes, M. Paoletti, R. Fischer, B. Miller, P. Dyer, M.S. Sachs, S.A. Osmani, and B.W. Birren. 2005. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature 438: 1105-1115.

Galtier, N. and J.R. Lobry. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44: 632-636.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283: 220-221.

Gao, B., R. Paramanathan, and R.S. Gupta. 2006. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. Antonie Van Leeuwenhoek 90: 69-91.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685-695.

Ge, F., L.S. Wang, and J. Kim. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol 3: e316.

Goffeau, A., B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S.G. Oliver. 1996. Life with 6000 genes. Science 274: 546, 563-567.

Gogarten, J.P., W.F. Doolittle, and J.G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19: 2226-2238.

Gogarten, J.P., H. Kibak, P. Dittrich, L. Taiz, E.J. Bowman, B.J. Bowman, M.F. Manolson, R.J. Poole, T. Date, T. Oshima, and et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci U S A 86: 6661-6665.

Goldstein, D.B. and D.D. Pollock. 1994. Least squares estimation of molecular distance--noise abatement in phylogenetic reconstruction. Theor Popul Biol 45: 219-226.

Gribaldo, S., V. Lumia, R. Creti, E.C. de Macario, A. Sanangelantoni, and P. Cammarano. 1999. Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181: 434-443.

Gribaldo, S. and H. Philippe. 2002. Ancient phylogenetic relationships. Theor Popul Biol 61: 391-408.

Griffiths, E., M.S. Ventresca, and R.S. Gupta. 2006. BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species. BMC Genomics 7: 14.

Grishin, N.V., Y.I. Wolf, and E.V. Koonin. 2000. From complete genomes to measures of substitution rate variability within and between proteins. Genome Res 10: 991-1000.

Gu, X. and H. Zhang. 2004. Genome phylogenetic analysis based on extended gene contents. Mol Biol Evol 21: 1401-1408.

Guarro, J., Gene J, and A.M. Stchigel. 1999. Developments in fungal taxonomy. Clin Microbiol Rev 12: 454-500.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

Gupta, R.S. and E. Griffiths. 2002. Critical issues in bacterial phylogeny. Theor Popul Biol 61: 423-434.

Gupta, S., N. Ferguson and R. Anderson. 1998. Chaos, persistance, and evolution of strain structure in antigenically diverse infectious agents. Science 280: 912-915.

Harris, J.K., S.T. Kelley, G.B. Spiegelman, and N.R. Pace. 2003. The genetic core of the universal ancestor. Genome Res 13: 407-412.

Hartman, H. and A. Fedorov. 2002. The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci U S A 99: 1420-1425.

Henz, S.R., D.H. Huson, A.F. Auch, K. Nieselt-Struwe, and S.C. Schuster. 2004. Whole-genome prokaryotic phylogeny. Bioinformatics.

Hillier, L.W. W. Miller E. Birney W. Warren R.C. Hardison C.P. Ponting P. Bork D.W. Burt M.A. Groenen M.E. Delany J.B. Dodgson A.T. Chinwalla P.F. Cliften S.W. Clifton K.D. Delehaunty C. Fronick R.S. Fulton T.A. Graves C. Kremitzki D. Layman V. Magrini J.D. McPherson T.L. Miner P. Minx W.E. Nash M.N. Nhan J.O. Nelson L.G. Oddy C.S. Pohl J. Randall-Maher S.M. Smith J.W. Wallis S.P. Yang M.N. Romanov C.M. Rondelli B. Paton J. Smith D. Morrice L. Daniels H.G. Tempest L. Robertson J.S. Masabanda D.K. Griffin A. Vignal V. Fillon L. Jacobbson S. Kerje L. Andersson R.P. Crooijmans J. Aerts J.J. van der Poel H. Ellegren R.B. Caldwell S.J. Hubbard D.V. Grafham A.M. Kierzek S.R. McLaren I.M. Overton H. Arakawa K.J. Beattie Y. Bezzubov P.E. Boardman J.K. Bonfield M.D. Croning R.M. Davies M.D. Francis S.J. Humphray C.E. Scott R.G. Taylor C. Tickle W.R. Brown J. Rogers J.M. Buerstedde S.A. Wilson L. Stubbs I. Ovcharenko L. Gordon S. Lucas M.M. Miller H. Inoko T. Shiina J. Kaufman J. Salomonsen K. Skjoedt G.K. Wong J. Wang B. Liu J. Yu H. Yang M. Nefedov M. Koriabine P.J. Dejong L. Goodstadt C. Webber N.J. Dickens I. Letunic M. Suyama D. Torrents C. von Mering E.M. Zdobnov K. Makova A. Nekrutenko L. Elnitski P. Eswara D.C. King S. Yang S. Tyekucheva A. Radakrishnan R.S. Harris F. Chiaromonte J. Taylor J. He M. Rijnkels S. Griffiths-Jones A. Ureta-Vidal M.M. Hoffman J. Severin S.M. Searle A.S. Law D. Speed D. Waddington Z. Cheng E. Tuzun E. Eichler Z. Bao P. Flicek D.D. Shteynberg M.R. Brent J.M. Bye E.J. Huckle S. Chatterji C. Dewey L. Pachter A. Kouranov Z. Mourelatos A.G. Hatzigeorgiou A.H. Paterson R. Ivarie M. Brandstrom E. Axelsson N. Backstrom S. Berlin M.T. Webster O. Pourquie A. Reymond C. Ucla S.E. Antonarakis M. Long J.J. Emerson E. Betran I. Dupanloup H. Kaessmann A.S. Hinrichs G. Bejerano T.S. Furey R.A. Harte B. Raney A. Siepel W.J. Kent D. Haussler E. Eyras R. Castelo J.F. Abril S. Castellano F. Camara G. Parra R. Guigo G. Bourque G. Tesler P.A. Pevzner A. Smit L.A. Fulton E.R. Mardis and R.K. Wilson. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695-716.

Hillis, D.M., J.P. Huelsenbeck, and C.W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264: 671-677.

Hu, G.Z. and H. Ronne. 1994. Yeast BTF3 protein is encoded by duplicated genes and inhibits the expression of some genes in vivo. Nucleic Acids Res 22: 2740-2743.

Huber, H., M.J. Hohn, R. Rachel, T. Fuchs, V.C. Wimmer, and K.O. Stetter. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417: 63-67.

Huelsenbeck, J.P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

Huminiecki, L. and K.H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14: 1870-1879.

Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18: 486.

Huson, D.H. and M. Steel. 2004. Phylogenetic trees based on gene content. Bioinformatics 20: 2044-2049.

Huynen, M., T. Dandekar, and P. Bork. 1998. Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett 426: 1-5.

Huynen, M.A., B. Snel, and P. Bork. 2001. Inversions and the dynamics of eukaryotic gene order. Trends Genet 17: 304-306.

Jain, R., M.C. Rivera, and J.A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96: 3801-3806.

James, T.Y., F. Kauff, C.L. Schoch, P.B. Matheny, V. Hofstetter, C.J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H.T. Lumbsch, A. Rauhut, V. Reeb, A.E. Arnold, A. Amtoft, J.E. Stajich, K. Hosaka, G.H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J.M. Curtis, J.C. Slot, Z. Wang, A.W. Wilson, A. Schussler, J.E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P.M. Letcher, M.J. Powell, J.W. Taylor, M.M. White, G.W. Griffith, D.R. Davies, R.A. Humber, J.B. Morton, J. Sugiyama, A.Y. Rossman, J.D. Rogers, D.H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R.A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R.A. Spotts, M. Serdani, P.W. Crous, K.W. Hughes, K. Matsuura, E. Langer, G. Langer, W.A. Untereiner, R. Lucking, B. Budel, D.M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D.S. Hibbett, F. Lutzoni, D.J. McLaughlin, J.W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443: 818-822.

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

Jones, T., N.A. Federspiel, H. Chibana, J. Dungan, S. Kalman, B.B. Magee, G. Newport, Y.R. Thorstenson, N. Agabian, P.T. Magee, R.W. Davis, and S. Scherer. 2004. The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A 101: 7329-7334.

Kainth, P. and R.S. Gupta. 2005. Signature proteins that are distinctive of alpha proteobacteria. BMC Genomics 6: 94.

Kamper, J., R. Kahmann, M. Bolker, L.J. Ma, T. Brefort, B.J. Saville, F. Banuett, J.W. Kronstad, S.E. Gold, O. Muller, M.H. Perlin, H.A. Wosten, R. de Vries, J. Ruiz-Herrera, C.G. Reynaga-Pena, K. Snetselaar, M. McCann, J. Perez-Martin, M. Feldbrugge, C.W. Basse, G. Steinberg, J.I. Ibeas, W. Holloman, P. Guzman, M. Farman, J.E. Stajich, R. Sentandreu, J.M. Gonzalez-Prieto, J.C. Kennell, L. Molina, J. Schirawski, A. Mendoza-Mendoza, D. Greilinger, K. Munch, N. Rossel, M. Scherer, M. Vranes, O. Ladendorf, V. Vincon, U. Fuchs, B. Sandrock, S. Meng, E.C. Ho, M.J. Cahill, K.J. Boyce, J. Klose, S.J. Klosterman, H.J. Deelstra, L. Ortiz-Castellanos, W. Li, P. Sanchez-Alonso, P.H. Schreier, I. Hauser-Hahn, M. Vaupel, E. Koopmann, G. Friedrich, H. Voss, T. Schluter, J. Margolis, D. Platt, C. Swimmer, A. Gnirke, F. Chen, V. Vysotskaia, G. Mannhaupt, U. Guldener, M. Munsterkotter, D. Haase, M. Oesterheld, H.W. Mewes, E.W. Mauceli, D. DeCaprio, C.M. Wade, J. Butler, S. Young, D.B. Jaffe, S. Calvo, C. Nusbaum, J. Galagan, and B.W. Birren. 2006. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature 444: 97-101.

Kapatral, V., I. Anderson, N. Ivanova, G. Reznik, T. Los, A. Lykidis, A. Bhattacharyya, A. Bartman, W. Gardner, G. Grechkin, L. Zhu, O. Vasieva, L. Chu, Y. Kogan, O. Chaga, E. Goltsman, A. Bernal, N. Larsen, M. D'Souza, T. Walunas, G. Pusch, R. Haselkorn, M. Fonstein, N. Kyrpides, and R. Overbeek. 2002. Genome sequence and analysis of the oral bacterium Fusobacterium nucleatum strain ATCC 25586. J Bacteriol 184: 2005-2018.

Katinka, M.D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C.P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414: 450-453.

Kellis, M., B.W. Birren, and E.S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428: 617-624.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-254.

Kennedy, S.P., W.V. Ng, S.L. Salzberg, L. Hood, and S. DasSarma. 2001. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res 11: 1641-1650.

Kirkpatrick, S., C. Gelatt, and M. Vecchi. 1983. Optimization by simulated annealing. Science 220: 671-680.

Klenk, H.P., T.D. Meier, P. Durovic, V. Schwass, F. Lottspeich, P.P. Dennis, and W. Zillig. 1999. RNA polymerase of Aquifex pyrophilus: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. J Mol Evol 48: 528-541.

Kolaczkowski, B. and J.W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984.

Koonin, E.V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1: 127-136.

Koonin, E.V., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, D.M. Krylov, K.S. Makarova, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, I.B. Rogozin, S. Smirnov, A.V. Sorokin, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5: R7.

Koonin, E.V. and A.R. Mushegian. 1996. Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr Opin Genet Dev 6: 757-762.

Korbel, J.O., B. Snel, M.A. Huynen, and P. Bork. 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet 18: 158-162.

Kouvelis, V.N., D.V. Ghikas, and M.A. Typas. 2004. The analysis of the complete mitochondrial genome of Lecanicillium muscarium (synonym Verticillium lecanii) suggests a minimum common gene organization in mtDNAs of Sordariomycetes: phylogenetic implications. Fungal Genet Biol 41: 930-940.

Kreil, D.P. and C.A. Ouzounis. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 29: 1608-1615.

Kunin, V., L. Goldovsky, N. Darzentas, and C.A. Ouzounis. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res 15: 954-959.

Kunin, V. and C.A. Ouzounis. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res 13: 1589-1594.

Kuramae, E., V. Robert, B. Snel, M. Weiss, and T. Boekhout. 2006. Phylogenomics reveal a robust fungal tree of life. FEMS Yeast Res: In press.

Kurtzman, C.P. 1998. Discussion of teleomorphic and anamorphic ascomycetous yeasts and a key to genera. In The yeasts, a taxonomic study (eds. C.P. Kurtzman and J.W. Fell), pp. 111-121. Elsevier, Amsterdam, The Netherlands.

Kurtzman, C.P. 2003. Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. FEMS Yeast Res 4: 233-245.

Langkjaer, R.B., P.F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature 421: 848-852.

Lefebvre, C., J.C. Aude, E. Glemet, and C. Neri. 2005. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. Bioinformatics 21: 1550-1558.

Lerat, E., V. Daubin, and N.A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1: E19.

Li, M., J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17: 149-154.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36: 96-99.

Loftus, B.J., E. Fung, P. Roncaglia, D. Rowley, P. Amedeo, D. Bruno, J. Vamathevan, M. Miranda, I.J. Anderson, J.A. Fraser, J.E. Allen, I.E. Bosdet, M.R. Brent, R. Chiu, T.L. Doering, M.J. Donlin, C.A. D'Souza, D.S. Fox, V. Grinberg, J. Fu, M. Fukushima, B.J. Haas, J.C. Huang, G. Janbon, S.J. Jones, H.L. Koo, M.I. Krzywinski, J.K. Kwon-Chung, K.B. Lengeler, R. Maiti, M.A. Marra, R.E. Marra, C.A. Mathewson, T.G. Mitchell, M. Pertea, F.R. Riggs, S.L. Salzberg, J.E. Schein, A. Shvartsbeyn, H. Shin, M. Shumway, C.A. Specht, B.B. Suh, A. Tenney, T.R. Utterback, B.L. Wickes, J.R. Wortman, N.H. Wye, J.W. Kronstad, J.K. Lodge, J. Heitman, R.W. Davis, C.M. Fraser, and R.W. Hyman. 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science 307: 1321-1324.

Lopandic, K., O. Molnar, M. Suzuki, W. Pinsker, and H. Prillinger. 2005. Estimation of Phylogenetic relationships within the Ascomycota on the basis of 18S rDNA sequences and chemotaxonomy. Mycol Progress 4: 205-214.

Lumbsch, H.T. 2000. Phylogeny of filamentous ascomycetes. Naturwissenschaften 87: 335-342.

Lumbsch, H.T., I. Schmitt, R. Lindemuth, A. Miller, A. Mangold, F. Fernandez, and S. Huhndorf. 2005. Performance of four ribosomal DNA regions to infer higher-level phylogenetic relationships of inoperculate euascomycetes (Leotiomyceta). Mol Phylogenet Evol 34: 512-524.

Lutzoni, F., F. Kauff, C.J. Cox, D. McLaughlin, G. Celio, B. Dentinger, M. Padamsee, D. Hibbett, T.Y. James, E. Baloch, M. Grube, V. Reeb, V. Hofstetter, C. Schoch, A.E. Arnold, J. Miadlikowska, J. Spatafora, D. Johnson, S. Hambleton, M. Crockett, R. Shoemaker, S. Hambleton, M. Crockett, R. Shoemaker, G.H. Sung, R. Lucking, T. Lumbsch, K. O'Donnell, M. Binder, P. Diederich, D. Ertz, C. Gueidan, K. Hansen, R.C. Harris, K. Hosaka, Y.W. Lim, B. Matheny, H. Nishida, D. Pfister, J. Rogers, A. Rossman, I. Schmitt, H. Sipman, J. Stone, J. Sugiyama, R. Yahr, and R. Vilgalys. 2004. Assembling the fungal tree of life: Progress, classification and evolution of subcellular traits. American Journal of Botany 91: 1446-1480.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

Martin, K.A., J.L. Siefert, S. Yerrapragada, Y. Lu, T.Z. McNeill, P.A. Moreno, G.M. Weinstock, W.R. Widger, and G.E. Fox. 2003. Cyanobacterial signature genes. Photosynth Res 75: 211-221.

Martinez, D., L.F. Larrondo, N. Putnam, M.D. Gelpke, K. Huang, J. Chapman, K.G. Helfenbein, P. Ramaiya, J.C. Detter, F. Larimer, P.M. Coutinho, B. Henrissat, R. Berka, D. Cullen, and D. Rokhsar. 2004. Genome sequence

of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78. Nat Biotechnol 22: 695-700.
Medina, M. 2005. Genomes, phylogeny, and evolutionary systems biology. Proc Natl Acad Sci U S A 102 Suppl 1: 6630-6635.
Mira, A., R. Pushker, B.A. Legault, D. Moreira, and F. Rodriguez-Valera. 2004. Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. BMC Evol Biol 4: 50.
Mirkin, B.G., T.I. Fenner, M.Y. Galperin, and E.V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3: 2.
Moore, G.E. 1965. Cramming more components onto integrated circuits. Electronics 38: 114-117.
Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
Nelson, K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen, C.M. Fraser, and et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399: 323-329.
Nierman, W.C., A. Pain, M.J. Anderson, J.R. Wortman, H.S. Kim, J. Arroyo, M. Berriman, K. Abe, D.B. Archer, C. Bermejo, J. Bennett, P. Bowyer, D. Chen, M. Collins, R. Coulsen, R. Davies, P.S. Dyer, M. Farman, N. Fedorova, T.V. Feldblyum, R. Fischer, N. Fosker, A. Fraser, J.L. Garcia, M.J. Garcia, A. Goble, G.H. Goldman, K. Gomi, S. Griffith-Jones, R. Gwilliam, B. Haas, H. Haas, D. Harris, H. Horiuchi, J. Huang, S. Humphray, J. Jimenez, N. Keller, H. Khouri, K. Kitamoto, T. Kobayashi, S. Konzack, R. Kulkarni, T. Kumagai, A. Lafton, J.P. Latge, W. Li, A. Lord, C. Lu, W.H. Majoros, G.S. May, B.L. Miller, Y. Mohamoud, M. Molina, M. Monod, I. Mouyna, S. Mulligan, L. Murphy, S. O'Neil, I. Paulsen, M.A. Penalva, M. Pertea, C. Price, B.L. Pritchard, M.A. Quail, E. Rabbinowitsch, N. Rawlins, M.A. Rajandream, U. Reichard, H. Renauld, G.D. Robson, S. Rodriguez de Cordoba, J.M. Rodriguez-Pena, C.M. Ronning, S. Rutter, S.L. Salzberg, M. Sanchez, J.C. Sanchez-Ferrero, D. Saunders, K. Seeger, R. Squares, S. Squares, M. Takeuchi, F. Tekaia, G. Turner, C.R. Vazquez de Aldana, J. Weidman, O. White, J. Woodward, J.H. Yu, C. Fraser, J.E. Galagan, K. Asai, M. Machida, N. Hall, B. Barrell, and D.W. Denning. 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature 438: 1151-1156.
Novichkov, P.S., M.V. Omelchenko, M.S. Gelfand, A.A. Mironov, Y.I. Wolf, and E.V. Koonin. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 186: 6575-6585.
Ochman, H., J.G. Lawrence, and E.A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304.
Olsen, G.J., C.R. Woese, and R. Overbeek. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176: 1-6.
Otu, H.H. and K. Sayood. 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19: 2122-2130.
Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides. 2003. The ERGO genome analysis and discovery system. Nucleic Acids Res 31: 164-171.
Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435: 814-818.
Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.
Pesole, G., C. Gissi, C. Lanave, and C. Saccone. 1995. Glutamine synthetase gene evolution in bacteria. Mol Biol Evol 12: 189-197.
Philippe, H. and P. Forterre. 1999. The rooting of the universal tree of life is not reliable. J Mol Evol 49: 509-523.
Philippe, H., E.A. Snell, E. Bapteste, P. Lopez, P.W. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21: 1740-1752.
Plotz, B.M., B. Lindner, K.O. Stetter, and O. Holst. 2000. Characterization of a novel lipid A containing D-galacturonic acid that replaces phosphate residues. The structure of the lipid a of the lipopolysaccharide from the hyperthermophilic bacterium Aquifex pyrophilus. J Biol Chem 275: 11222-11228.
Prillinger, H., K. Lopandic, W. Schweigkofler, R. Deak, H.J.M. Aarts, R. Bauer, K. Sterflinger, G.F. Kraus, and A. Maraz. 2002. Phylogeny and systematics of the fungi with special reference to the Ascomycota and Basidiomycota. Fungal Allergy and Pathogenicity 81: 207-295.
Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-504.
Qi, J., H. Luo, and B. Hao. 2004a. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 32: W45-47.
Qi, J., B. Wang, and B.I. Hao. 2004b. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol 58: 1-11.
Ramani, A.K. and E.M. Marcotte. 2003. Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity. J Mol Biol 327: 273-284.
Remm, M., C.E. Storm, and E.L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041-1052.
Rindi, L., N. Lari, and C. Garzelli. 2001. Genes of Mycobacterium tuberculosis H37Rv downregulated in the attenuated strain H37Ra are restricted to M. tuberculosis complex species. New Microbiol 24: 289-294.
Robbertse, B., J.B. Reeves, C.L. Schoch, and J.W. Spatafora. 2006. A phylogenomic analysis of the Ascomycota. Fungal Genet Biol 43: 715-725.
Rokas, A., B.L. Williams, N. King, and S.B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
Rospert, S., Y. Dubaquie, and M. Gautschi. 2002. Nascent-polypeptide-associated complex. Cell Mol Life Sci 59: 1632-1639.
Rusch, D.B., A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J.E. Venter, K. Li, S. Kravitz, J.F. Heidelberg, T. Utterback, Y.H. Rogers, L.I. Falcon, V. Souza, G. Bonilla-Rosso, L.E. Eguiarte, D.M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M.R. Ferrari, R.L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J.C. Venter. 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5: e77.
Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.
Sankoff, D., G. Leduc, N. Antoine, B. Paquin, B.F. Lang, and R. Cedergren. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci U S A 89: 6575-6579.
Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341-345.
Schmidt, H.A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502-504.
Scorzetti, G., J.W. Fell, A. Fonseca, and A. Statzell-Tallman. 2002. Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions. FEMS Yeast Res 2: 495-517.
Sicheritz-Ponten, T. and S.G. Andersson. 2001. A phylogenomic approach to microbial evolution. Nucleic Acids Res 29: 545-552.
Sipiczki, M. 2000. Where does fission yeast sit on the tree of life? Genome Biol 1: REVIEWS1011.
Slesarev, A.I., K.V. Mezhevaya, K.S. Makarova, N.N. Polushin, O.V. Shcherbinina, V.V. Shakhova, G.I. Belova, L. Aravind, D.A. Natale, I.B. Rogozin, R.L. Tatusov, Y.I. Wolf, K.O. Stetter, A.G. Malykh, E.V. Koonin, and S.A. Kozyavkin. 2002. The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A 99: 4644-4649.
Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol 147: 195-197.
Snel, B., P. Bork, and M.A. Huynen. 1999. Genome phylogeny based on gene content. Nat Genet 21: 108-110.
Snel, B., P. Bork, and M.A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12: 17-25.
Snel, B., M.A. Huynen, and B.E. Dutilh. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol 59: 191-209.
Snel, B., V. van Noort, and M.A. Huynen. 2004. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. Nucleic Acids Res 32: 4725-4731.
Soding, J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951-960.
Strous, M., E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M.W. Taylor, M. Horn, H. Daims, D. Bartol-Mavel, P. Wincker, V. Barbe, N. Fonknechten, D. Vallenet, B. Segurens, C. Schenowitz-Truong, C. Medigue, A. Collingro, B. Snel, B.E. Dutilh, H.J. Op den Camp, C. van der Drift, I. Cirpus, K.T. van de Pas-Schoonen, H.R. Harhangi, L. van Niftrik, M. Schmid, J. Keltjens, J. van de Vossenberg, B. Kartal, H. Meier, D. Frishman, M.A. Huynen, H.W. Mewes, J. Weissenbach, M.S. Jetten, M. Wagner, and D. Le Paslier. 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440: 790-794.
Stuart, J.M., E. Segal, D. Koller, and S.K. Kim. 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255.
Suhre, K. and J.M. Claverie. 2003. Genomic correlates of hyperthermostability: an update. J Biol Chem.
Suyama, M. and P. Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 17: 10-13.
Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. Genome Biol 2: RESEARCH0020.
Tatusov, R.L., M.Y. Galperin, D.A. Natale, and E.V. Koonin. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33-36.
Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. Science 278: 631-637.
Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28.

Tehler, A., D.P. Little, and J.S. Farris. 2003. The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, Fungi. Mycol Res 107: 901-916.
Teichmann, S.A. and G. Mitchison. 1999. Is there a phylogenetic signal in prokaryote proteins? J Mol Evol 49: 98-107.
Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res 9: 550-557.
Thomarat, F., C.P. Vivares, and M. Gouy. 2004. Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. J Mol Evol 59: 780-791.
Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
Tiboni, O., P. Cammarano, and A.M. Sanangelantoni. 1993. Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum Pyrococcus woesei: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. J Bacteriol 175: 2961-2969.
Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. Science 308: 554-557.
Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.
van der Heijden, R.T.J.M., B. Snel, V. van Noort, and M.A. Huynen. submitted. Orthology prediction at scalable resolution through automated analysis of phylogenetic trees. BMC Bioinformatics.
van Dijk, M.A., O. Madsen, F. Catzeflis, M.J. Stanhope, W.W. de Jong, and M. Pagel. 2001. Protein sequence signatures support the African clade of mammals. Proc Natl Acad Sci U S A 98: 188-193.
van Noort, V., B. Snel, and M.A. Huynen. 2003. Predicting gene function by conserved co-expression. Trends Genet 19: 238-242.
Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, and H.O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.
Vinga, S. and J. Almeida. 2003. Alignment-free sequence comparison-a review. Bioinformatics 19: 513-523.
Vivares, C.P., M. Gouy, F. Thomarat, and G. Metenier. 2002. Functional and evolutionary analysis of a eukaryotic parasitic genome. Curr Opin Microbiol 5: 499-505.
von Mering, C., P. Hugenholtz, J. Raes, S.G. Tringe, T. Doerks, L.J. Jensen, N. Ward, and P. Bork. 2007a. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315: 1126-1130.
von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31: 258-261.
von Mering, C., L.J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. 2007b. STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358-362.
von Mering, C., L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33 Database Issue: D433-437.
Waters, E., M.J. Hohn, I. Ahel, D.E. Graham, M.D. Adams, M. Barnstead, K.Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G.G. Sutton, M. Simon, D. Soll, K.O. Stetter, J.M. Short, and M. Noordewier. 2003. The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A 100: 12984-12988.
Welch, R.A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E.L. Buckles, S.R. Liou, A. Boutin, J. Hackett, D. Stroud, G.F. Mayhew, D.J. Rose, S. Zhou, D.C. Schwartz, N.T. Perna, H.L. Mobley, M.S. Donnenberg, and F.R. Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A 99: 17020-17024.
Werner, T. 2002. Finding and decrypting of promoters contributes to the elucidation of gene function. In Silico Biol 2: 249-255.
Wheeler, D.L., C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova, and B.A. Rapp. 2000. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 28: 10-14.
Wheeler, D.L., D.M. Church, A.E. Lash, D.D. Leipe, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, T.A. Tatusova, L. Wagner, and B.A. Rapp. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30: 13-16.
Woese, C.R. 1987. Bacterial evolution. Microbiol Rev 51: 221-271.
Wolf, Y.I., I.B. Rogozin, N.V. Grishin, and E.V. Koonin. 2002. Genome trees and the tree of life. Trends Genet 18: 472-479.
Wolf, Y.I., I.B. Rogozin, N.V. Grishin, R.L. Tatusov, and E.V. Koonin. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1: 8.
Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708-713.
Wood, V. R. Gwilliam M.A. Rajandream M. Lyne R. Lyne A. Stewart J. Sgouros N. Peat J. Hayles S. Baker D. Basham S. Bowman K. Brooks D. Brown S. Brown T. Chillingworth C. Churcher M. Collins R. Connor A. Cronin P. Davis T. Feltwell A. Fraser S. Gentles A. Goble N. Hamlin D. Harris J. Hidalgo G. Hodgson S. Holroyd T. Hornsby S. Howarth E.J. Huckle S. Hunt K. Jagels K. James L. Jones M. Jones S. Leather S. McDonald J. McLean P. Mooney S. Moule K. Mungall L. Murphy D. Niblett C. Odell K. Oliver S. O'Neil D. Pearson M.A. Quail E. Rabbinowitsch K. Rutherford S. Rutter D. Saunders K. Seeger S. Sharp J. Skelton M. Simmonds R. Squares S. Squares K. Stevens K. Taylor R.G. Taylor A. Tivey S. Walsh T. Warren S. Whitehead J. Woodward G. Volckaert R. Aert J. Robben B. Grymonprez I. Weltjens E. Vanstreels M. Rieger M. Schafer S. Muller-Auer C. Gabel M. Fuchs C. Fritzc E. Holzer D. Moestl H. Hilbert K. Borzym I. Langer A. Beck H. Lehrach R. Reinhardt T.M. Pohl P. Eger W. Zimmermann H. Wedler R. Wambutt B. Purnelle A. Goffeau E. Cadieu S. Dreano S. Gloux V. Lelaure S. Mottier F. Galibert S.J. Aves Z. Xiang C. Hunt K. Moore S.M. Hurst M. Lucas M. Rochet C. Gaillardin V.A. Tallada A. Garzon G. Thode R.R. Daga L. Cruzado J. Jimenez M. Sanchez F. del Rey J. Benito A. Dominguez J.L. Revuelta S. Moreno J. Armstrong S.L. Forsburg L. Cerrutti T. Lowe W.R. McCombie I. Paulsen J. Potashkin G.V. Shpakovski D. Ussery B.G. Barrell and P. Nurse. 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415: 871-880.
Yang, S., R.F. Doolittle, and P.E. Bourne. 2005. Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A 102: 373-378.
Yooseph, S., G. Sutton, D.B. Rusch, A.L. Halpern, S.J. Williamson, K. Remington, J.A. Eisen, K.B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C.S. Miller, H. Li, S.T. Mashiyama, M.P. Joachimiak, C. van Belle, J.M. Chandonia, D.A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B.J. Raphael, V. Bafna, R. Friedman, S.E. Brenner, A. Godzik, D. Eisenberg, J.E. Dixon, S.S. Taylor, R.L. Strausberg, M. Frazier, and J.C. Venter. 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biol 5: e16.
Zhaxybayeva, O., P. Lapierre, and J.P. Gogarten. 2004. Genome mosaicism and organismal lineages. Trends Genet 20: 254-260.
Zomorodipour, A. and S.G. Andersson. 1999. Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis. FEBS Lett 452: 11-15.
Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. J Theor Biol 8: 357-366.

Appendices

Supplemental material for "Assessment of phylogenomic and orthology approaches for phylogenetic inference"

Literature support for fungal phylogeny

Table 11. Resolved topological features of the fungal taxonomy (see also Figure 11).

node description	refs that support this node	refs that contradict this node
Sce, Spa	(Delsuc et al. 2005; Jeffroy et al. 2006; Kuramae et al. 2006; Rokas et al. 2003)	(Kurtzman 2003)
Cal, Dha	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Kuramae et al. 2006; Lopandic et al. 2005; Prillinger et al. 2002)	(Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)
Hypocreales	(fungal.genome.duke.edu; Robbertse et al. 2006), NCBI taxonomy (Wheeler et al. 2002)	
Mgr, Ncr	(fungal.genome.duke.edu; Kuramae et al. 2006; Robbertse et al. 2006; Tehler et al. 2003)	
Eurotiomycetes	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Robbertse et al. 2006; Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)	
Hymenomycetes	(fungal.genome.duke.edu; James et al. 2006; Kuramae et al. 2006; Lutzoni et al. 2004; Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)	
Sce, Smi, Spa	(fungal.genome.duke.edu; Delsuc et al. 2005; Jeffroy et al. 2006; Kuramae et al. 2006; Kurtzman 2003; Rokas et al. 2003)	
Basidiomycota	(fungal.genome.duke.edu; James et al. 2006; Kuramae et al. 2006; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)	
Sce, Sku, Smi, Spa	(fungal.genome.duke.edu; Delsuc et al. 2005; Jeffroy et al. 2006; Kuramae et al. 2006; Kurtzman 2003; Rokas et al. 2003)	
Ago, Kla, Kwa, Skl	(Diezmann et al. 2004; James et al. 2006; Jeffroy et al. 2006; Kuramae et al. 2006; Tehler et al. 2003)	(Kurtzman 2003), NCBI taxonomy (Wheeler et al. 2002)
Sordariomycetes	(fungal.genome.duke.edu; James et al. 2006; Kuramae et al. 2006; Robbertse et al. 2006), NCBI taxonomy (Wheeler et al. 2002)	
Saccharomyces *sensu stricto*	(fungal.genome.duke.edu; Jeffroy et al. 2006; Kuramae et al. 2006; Kurtzman 2003; Rokas et al. 2003; Tehler et al. 2003)	
Sba, Sca, Sce, Sku, Smi, Spa	(Jeffroy et al. 2006; Kouvelis et al. 2004; Kuramae et al. 2006; Kurtzman 2003; Tehler et al. 2003)	(James et al. 2006)
Cgl, Sba, Sca, Sce, Sku, Smi, Spa	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Jeffroy et al. 2006; Kouvelis et al. 2004; Kuramae et al. 2006; Kurtzman 2003; Lopandic et al. 2005; Prillinger et al. 2002; Tehler et al. 2003)	(Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)
Euascomycetes	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Kouvelis et al. 2004; Lopandic et al. 2005; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Robbertse et al. 2006; Tehler et al. 2003; Thomarat et al. 2004), NCBI taxonomy (Wheeler et al. 2002)	
Saccharomyces *sensu lato*	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Jeffroy et al. 2006; Kuramae et al. 2006; Prillinger et al. 2002; Robbertse et al. 2006; Tehler et al. 2003)	
Yli primitive in Hemiascomycetes	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Kouvelis et al. 2004; Kuramae et al. 2006; Prillinger et al. 2002; Robbertse et al. 2006; Thomarat et al. 2004)	Saccharomyces *sensu lato* primitive in Hemiascomycetes (Tehler et al. 2003)
Hemiascomycetes	(fungal.genome.duke.edu; Diezmann et al. 2004; James et al. 2006; Kouvelis et al. 2004; Lopandic et al. 2005; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Robbertse et al. 2006; Tehler et al. 2003; Thomarat et al. 2004), NCBI taxonomy (Wheeler et al. 2002)	
Ascomycota	(fungal.genome.duke.edu; James et al. 2006; Kouvelis et al. 2004; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Robbertse et al. 2006; Tehler et al. 2003), NCBI taxonomy (Wheeler et al. 2002)	

Table 12. Unresolved issues in the fungal taxonomy (see also Figure 11).

node description	refs that support this node	refs that contradict this node
Ago, Kla	(James et al. 2006; Jeffroy et al. 2006; Kuramae et al. 2006; Tehler et al. 2003), Figure 1 in (Diezmann et al. 2004)	(Kurtzman 2003), Figure 2 in (Diezmann et al. 2004), NCBI taxonomy (Wheeler et al. 2002)
Ago, Kwa		
Ago, Skl	Figure 2 in (Diezmann et al. 2004)	(Kurtzman 2003; Tehler et al. 2003), Figure 1 in (Diezmann et al. 2004), NCBI taxonomy (Wheeler et al. 2002)
Kla, Kwa	NCBI taxonomy (Wheeler et al. 2002)	(Kurtzman 2003; Tehler et al. 2003)
Kla, Skl		
Kwa, Skl	(Jeffroy et al. 2006; Kurtzman 2003)	NCBI taxonomy (Wheeler et al. 2002)
Ago, Kla, Kwa		
Ago, Kwa, Skl		
Ago, Kwa, Skl		
Kla, Kwa, Skl		
Sordariomycetes primitive in Euascomycetes	(fungal.genome.duke.edu; Lopandic et al. 2005; Tehler et al. 2003)	(Lumbsch 2000; Lumbsch et al. 2005; Lutzoni et al. 2004; Prillinger et al. 2002)
Eurotiomycetes primitive in Euascomycetes	(Lumbsch 2000; Lumbsch et al. 2005; Lutzoni et al. 2004; Prillinger et al. 2002)	(fungal.genome.duke.edu; Lopandic et al. 2005; Tehler et al. 2003)
Sno primitive in Euascomycetes		(fungal.genome.duke.edu; Lopandic et al. 2005; Lumbsch 2000; Lumbsch et al. 2005; Lutzoni et al. 2004; Prillinger et al. 2002; Tehler et al. 2003)
Hemiascomycetes primitive in Ascomycota	(Diezmann et al. 2004; Prillinger et al. 2002; Tehler et al. 2003; Thomarat et al. 2004)	(fungal.genome.duke.edu; Berbee et al. 2000; Kouvelis et al. 2004; Kuramae et al. 2006; Lopandic et al. 2005; Lutzoni et al. 2004; Medina 2005; Sipiczki 2000; Vivares et al. 2002)
Euascomycetes primitive in Ascomycota	(Kouvelis et al. 2004; Lopandic et al. 2005)	(fungal.genome.duke.edu; Berbee et al. 2000; Diezmann et al. 2004; Kuramae et al. 2006; Lutzoni et al. 2004; Medina 2005; Prillinger et al. 2002; Sipiczki 2000; Tehler et al. 2003; Thomarat et al. 2004; Vivares et al. 2002)
Archiascomycetes primitive in Ascomycota	(fungal.genome.duke.edu; Berbee et al. 2000; Kuramae et al. 2006; Lutzoni et al. 2004; Medina 2005; Sipiczki 2000; Vivares et al. 2002)	(Diezmann et al. 2004; Kouvelis et al. 2004; Lopandic et al. 2005; Prillinger et al. 2002; Tehler et al. 2003; Thomarat et al. 2004)

Orthology approaches

All multiple alignments, orthology relations, phylogenetic trees and phylogenomic trees are available at www.cmbi.ru.nl/~dutilh/phylogenomics. The *Sac. kluyverii* gene annotations were downloaded from www.broad.mit.edu/seq/YeastDuplication/. Similarity scores between all the proteomes were computed using the Smith-Waterman P algorithm (Smith and Waterman 1981) on a TimeLogic DeCypher in all query-database combinations (matrix: Blosum62; e-value cutoff: 0.01; low-complexity filter on).

Pairwise orthology

To define pairwise orthologous groups, we used the program Inparanoid (Remm et al. 2001). We detected 1,025,849 pairwise orthologous groups, or "InparanOGs", between all 325 species pairs (score cutoff: 50; outgroup cutoff: 50; sequence overlap cutoff: 0.5; confidence cutoff: 0.05; group overlap cutoff: 0.5; gray zone: 0).

Cluster orthology

To determine cluster orthology, we used a method based on COG (Tatusov et al. 1997). The five Saccharomyces *sensu stricto* species (Sba, Sce, Sku, Smi, Spa) were first joined into one clade. Then, in-paralogs were determined within the clades (and in the remaining species): if a pair of proteins had a better Smith-Waterman P score than either of them had with any protein outside the clade. Bi-directional best hits were determined between groups of inparalogs, and triangles of bi-directional best hits were joined if they share one bi-directional best hit. Thus, we formed 8,044 triangle based cluster orthologous groups or "triOGs". Alternatively, more inclusive cluster orthologous groups were based directly on the bi-directional best hits, simply joining them if they share a gene from a

single inparalogous group. This approach yielded 10,754 pair based cluster orthologous groups or "duOGs".

Unambiguous orthologous groups are a sub-group of the cluster orthologous groups that have at most one representative in any species. Because a diploid genome assembly (Jones et al. 2004) or recent, species specific duplications in general may greatly reduce the number of unambiguous orthologous groups identified (cf. Figure 27), we collapsed all recent duplications identified by LOFT (van der Heijden et al. submitted) in distance tree-based orthology (see below), retaining only the gene with the shortest branch length to the root. Thus, we obtained 8,722 unambiguous duOGs and 6,488 unambiguous triOGs. Among these unambiguous cluster orthologous groups, the pan-orthologs are those that are present in every one of the species considered. We found 64 pan-duOGs and 59 pan-triOGs.

We chose *E. cuniculi* as an outgroup because it is the species most closely related to the Fungi that has a completely sequenced genome (Thomarat et al. 2004; Vivares et al. 2002). As an intracellular parasite, *E. cuniculi* has a degenerated genome (Katinka et al. 2001). Thus, this choice will likely limit the number of pan-orthologs identified. Indeed there are 143 pan-duOGs and 140 pan-triOGs without *E. cuniculi*. 422 unambiguous duOGs and 412 unambiguous triOGs contain *E. cuniculi*.

Tree-based orthology

We built tree-based orthology from phylogenetic trees that were constructed using both distance and maximum likelihood. To do this, we first aligned the duOGs and the triOGs using Muscle 3.52 (Edgar 2004b) with default parameters (one exceptionally large duOG was aligned without refinement, using the -maxiters 2 option). Then, for the phylogenetic distance trees, pairwise protein distances were calculated with Tree-Puzzle 5.2 (Schmidt et al. 2002) (approximate parameter estimates; parameter estimation uses neighbor-joining tree; JTT model of substitution; estimate amino acid frequencies from data set; 4 gamma categories; alpha = 1.00 (weak rate heterogeneity); to calculate protein distances in orthologous groups with less than four sequences, we added each sequence twice). We used Bio-NJ (Gascuel 1997) to reconstruct distance trees.

For the phylogenetic maximum likelihood trees, we used PhyML (JTT model of substitution; estimated proportion of invariable sites; 4 substitution rate categories; gamma fixed with alpha = 1.00 (Guindon and Gascuel 2003)). Two exceptionally large duOGs (831 sequences aligned over 2,664 positions and 537 sequences aligned over 2,379 positions) had to be discarded from further analysis, as we could not reconstruct a maximum likelihood tree for them.

Subsequently, the phylogenetic trees were analyzed with LOFT (van der Heijden et al. submitted), using the autoroot option. LOFT does not impose a phylogeny on the data, but assigns orthology relations in a tree based on species overlap between branches. The results were analyzed for each species pair, resulting in a total of 858,622 distance tree-duOGs, 820,007 distance tree-triOGs, 856,363 likelihood tree-duOGs and 822,570 likelihood tree-triOGs over the 325 species pairs. There are more InparanOGs than tree-based orthologous groups because the latter are based on phylogenetic trees with at least three species, whereas the InparanOGs are defined between species pairs, and thus have a larger coverage.

Note that although tree-based orthology is in principle an ideal approach to determine levels of orthology, it needs to be operationalized. Assume that the phylogeny in Figure 27 reflects the evolutionary history of a gene family. In this gene family, *Sac. cerevisiae* is represented by Sce_YOR285W and Sce_YOR286W, and *D. hansenii* is represented by Dha_DEHA0G24948g. Inspection of the tree shows us that in orthologous group 1.2, Dha_DEHA0G24948g is orthologous to Sce_YOR286W, but not to Sce_YOR285W, which is in orthologous group 1.1. The two *Sac. cerevisiae* genes are paralogs that duplicated before the ancestor of the Saccharomycetaceae. However, the duplication took place after *Y. lipolytica* branched off, as Yli_YALI0F29667g is in orthologous group 1, which is at a higher level of orthology. Thus, Yli_YALI0F29667g is equally orthologous to both the *Sac. cerevisiae* genes. The other *Y. lipolytica* gene, Yli_YALI0B01650g, is not orthologous to any *Sac. cerevisiae* gene:

according to this phylogeny, orthologous group 2 was likely lost after *Y. lipolytica* branched off in the Hemiascomycetes. As these examples show, the operationalized tree-based orthologous groups look like pairwise orthologous groups. However, tree-based orthology is more accurate, as it is based on a phylogeny, while similarity-based orthology methods such as Inparanoid compose orthologous groups directly from the similarity scores (Remm et al. 2001).

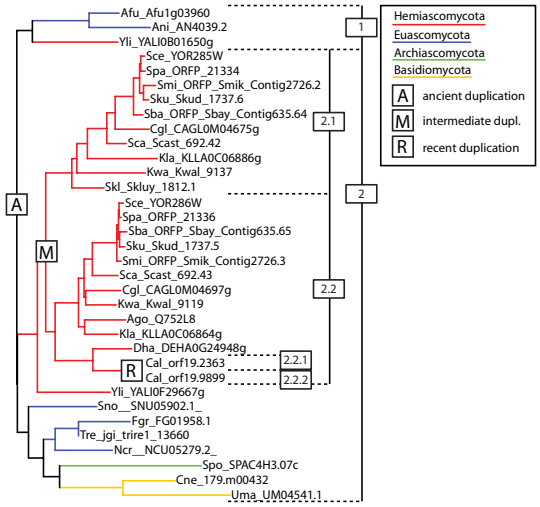

Figure 27. Tree-based orthology. LOFT analyzed the phylogenetic distance tree of this triOG to identify the duplication nodes (van der Heijden et al. submitted). In the history of this gene family, there have been three duplications: an ancient duplication (A), an intermediate duplication (M) and a recent duplication (R; note that the recent duplication identified in Can. albicans was likely the result of its diploid genome sequence (Jones et al. 2004)). The remaining nodes are speciation nodes. The levels of orthology as identified by LOFT are shown in the boxes to the right of the tree.

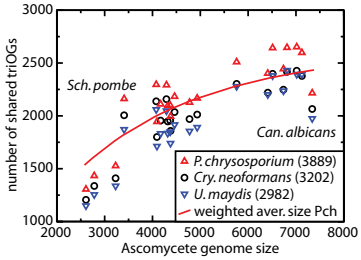

Figure 28. Large genomes share relatively many genes. For each of the three Basidiomycota included in this research, we have plotted the number of triOGs shared with all Ascomycota. The genome size (defined as the number of genes in any triOG) is indicated between brackets for the Basidiomycota, and on the x-axis for the Ascomycota. The drawn line is $y = c \cdot (\sqrt{2} \cdot x \cdot 3889) / (\sqrt{(x^2 + 3889^2)})$ (see Equation 3), where $c = 0.5$ (this scales the line into the figure).

Genome size effect in shared gene content

Because larger genomes can share more genes, we normalize the number of shared OGs in the gene content approach by dividing by the weighted average genome size (Dutilh et al. 2004; Korbel et al. 2002). Even though there are no extraordinarily reduced genomes within the Fungi, the number of shared OGs between the Basidiomycota and the Ascomycota increases and saturates with increasing ascomycotal genome size (Figure 28). In this respect, fungal genomes seem to behave like bacterial genomes (Snel et al. 1999). There are two genomes in Figure 28 that stand out. The first is *Sch. pombe*, which lies above the general curve, sharing a relatively large number

of triOGs with the Basidiomycota. This suggests that *Sch. pombe* belongs at an ancestral position in the Ascomycota. The second organism is *Can. albicans*, which shares a strikingly low number of triOGs with the Basidiomycota. Considering the diploid genome of this organism (Jones et al. 2004), it might be better to halve the estimate of the *Can. albicans* genome size (on the *x*-axis) from 7,340 to about 3,670 genes that are in a triOG. Indeed, this operation would place *Can. albicans* neatly in line with the other Ascomycota.

Phylogenomic approaches

All multiple alignments, orthology relations, phylogenetic trees and phylogenomic trees are available at www.cmbi.ru.nl/~dutilh/phylogenomics.

Gene content

Transforming gene content information into a phylogeny can, in principle, be done using distance, parsimony or likelihood for tree reconstruction. A distance measure that corrects for differences in genome size has been shown to yield good results (Dutilh et al. 2004; Korbel et al. 2002), although for simulated data, it has been reported to be outperformed by likelihood distance and Dollo parsimony (Huson and Steel 2004). Although some maximum likelihood distance methods have been published (Gu and Zhang 2004; Huson and Steel 2004), pure maximum likelihood thus far requires too much computer time, and a heuristic algorithm such as PhyML (Guindon and Gascuel 2003) has not yet been developed for character data with asymmetric conversion rates. This is necessary if we want to account for the fact that it is harder to acquire an orthologous group than to lose it. However, Dollo parsimony has been especially developed for presence/absence data of complex characters such as genes (Farris 1977; Felsenstein 1989). Under the Dollo principle, a gene family can be obtained only once, but lost several times in evolution. Although this scenario may be less likely in prokaryotes, where genes may be transmitted between species through horizontal gene transfer, this simplified model of evolution will be readily applicable to Fungi, where horizontal transfer is rare (Andersson 2005). Thus, we use Dollo parsimony rather than maximum likelihood or a likelihood distance as an alternative to the distance measure to analyze gene content.

Gene content trees were calculated from presence-absence profiles using distance and parsimony. For the distance method, the evolutionary distance between each species pair is calculated according to Equation 3 (Korbel et al. 2002), where shared_OGs(A,B) is the number of orthologous groups shared between species A and species B; size_A and size_B are the genome sizes of species A and species B, respectively.

Equation 3. Pairwise distance based on gene content.

$$dist\,(A, B) = 1 - \frac{shared\_OGs\,(A, B)}{(\sqrt{2} \cdot size\_A \cdot size\_B) / (\sqrt{(size\_A^2 + size\_B^2)})}$$

As distantly related species with large genomes may share more genes than closer related species with small genomes (Figure 28), the number of shared orthologous groups was corrected for genome size by dividing by the geometric mean of the genome sizes. For each type of orthology, the genome size was calculated separately as the total number of genes with orthologs in any other species. We used Bio-NJ (Gascuel 1997) to transform the distance matrix into a gene content tree.

For Dollo parsimony (Farris 1977), we used the Dollop program implemented in the Phylip package (Felsenstein 1989) on the presence/absence matrices of the cluster orthologous groups (duOGs and triOGs).

Superalignment

For the superalignment approach, multiple alignments obtained using Muscle (Edgar 2004b) (see the paragraph "Orthology" above) of either the unambiguous cluster orthologous groups or the pan-orthologous groups were concatenated to form a superalignment (the superalignment has also been called a supermatrix (Bininda-Emonds 2004; Delsuc et al. 2005), but we refrain from using this

term to avoid confusion with the superdistance approach). Unambiguous orthologous groups that are absent from certain species were coded with question marks, and form gaps in the alignment (Philippe et al. 2004). For the distance approach, we used Tree-Puzzle (Schmidt et al. 2002) to obtain maximum likelihood distances between all the supersequences (see the paragraph "Orthology" above). Subsequently, we used Bio-NJ (Gascuel 1997) to obtain a distance based superalignment tree. For the maximum likelihood approach, PhyML (Guindon and Gascuel 2003) calculated a maximum likelihood based superalignment tree (for parameters see paragraph "Orthology" above).

Because these phylogenetic inference programs are relatively computer intensive, we had to somewhat restrict the length of the superalignment. Instead of superalignments of the complete set of unambiguous orthologous groups, that reached a length of up to 5,783,459 characters for the unambiguous duOGs, we used a set intermediate between the unambiguous orthologous groups and the pan-orthologs. Superalignments were composed for those gene families that were present in more than a certain number of species, depending on what the phylogenetic inference program could handle. To calculate the distances between the supersequences with Puzzle, the maximum superalignment was that of unambiguous orthologous groups present in more than 21 species, reaching a length of 878,973 positions for the unambiguous duOGs and 910,270 positions for the unambiguous triOGs. Because of computer limitations, PhyML could not be used for superalignments longer than those composed only of the pan-orthologs (i.e. unambiguous orthologs present in more than 25 species). These superalignments had a total length of 33,731 positions for the pan-duOGs and 28,925 positions for the pan-triOGs.

We selected blocks of unambiguously aligned amino acids using GBlocks (Castresana 2000) under default parameters, and were thus able to include more unambiguous orthologous groups in the superalignments. After filtering the alignments with GBlocks, Tree-Puzzle could handle superalignments from unambiguous duOGs present in more than three species (1,099,070 conserved positions), and from unambiguous triOGs present in more than five species (1,017,559 conserved positions). PhyML could infer phylogenomic trees from concatenated alignments of unambiguous orthologous groups present in more than 24 species (132,409 conserved positions for the unambiguous duOGs and 132,527 conserved positions for the unambiguous triOGs). The GBlocks filtered superalignments of pan-orthologs were 11,625 and 11,006 conserved positions long for the pan-duOGs and the pan-triOGs, respectively.

Superdistance

Distance matrices contain an evolutionary distance for all species pairs. On a phylogenomic scale, the distance between a species pair in a superdistance matrix is the average over all orthologous groups that are shared between the two species. These evolutionary distances were calculated using various methods. For InparanOGs, similarities between all orthologous gene pairs between two species were first calculated as the Smith-Waterman P score divided by the length of the match. Then, the similarity scores were averaged over all gene pairs within an orthologous group (alternatively, the smallest or largest similarity was taken, but this yielded identical superdistance trees). Finally, the similarity score between a species pair is the average of the similarities of all orthologous groups. These scores were normalized between zero and one: for each species pair, the average of the two scores of the species with themselves was set to one, and the scores between the species were scaled down. The resulting normalized similarity scores were subtracted from one to yield the distance. We used Bio-NJ (Gascuel 1997) to infer the superdistance trees based on pairwise orthology.

For the other orthology definitions, the maximum likelihood distances between all orthologous genes for two species were calculated using Tree-Puzzle (Schmidt et al. 2002) with the parameters detailed in the paragraph "Orthology" above. The distance for one orthologous group was the average of all the distances between its gene pairs, and the distance between a species pair was the average of all the distances between the orthologous groups they share. We used Bio-NJ (Gascuel 1997) to transform the superdistance matrices into superdistance trees.

If certain orthologous groups evolve very rapidly, it is likely that they will only be identified as orthologs in closely related species. As the distances for such orthologous groups will be relatively large, they will increase the distance between the closely related species, pushing them apart unnecessarily. To correct for this, we used SDM* (Criscuolo et al. 2006) to normalize the superdistance matrices of the unambiguous duOGs and triOGs, and of the pan-duOGs and pan-triOGs. Subsequently, the normalized matrices were treated the same way as the original matrices.

Supertree

Supertrees were composed of gene family trees, which were calculated from the duOGs and triOGs using distance or maximum likelihood. For the phylogenetic distance trees, we used Tree-Puzzle to obtain maximum likelihood distances (Schmidt et al. 2002) and BioNJ to reconstruct distance trees (Gascuel 1997); for the phylogenetic maximum likelihood trees, we used PhyML (Guindon and Gascuel 2003) (see also the paragraph "Orthology" above). To integrate the different phylogenetic trees into a phylogenomic supertree, we used two algorithms. The first was the majority rule from Consense 3.64 (Felsenstein 1989), which requires pan-orthologous groups. As an alternative, we used CLANN (Creevey and McInerney 2005), which requires unambiguous orthologous groups (parameters: dfit optimality criterion; heuristic search algorithm = SPR; nsteps=5; maxswaps=1,000,000; nreps=10; weighting scheme = comparisons; starting trees = top 10 random trees chosen from 10,000 random samples).

CLANN and Consense differ in two ways. Firstly, Consense requires that the input trees all contain exactly the same species, while CLANN can handle trees with different sets of species. This is because CLANN is specifically developed to handle trees from many different studies, and hence with different species compositions. For our specific purpose, this allows CLANN to compute a supertree from a much larger set of OGs, namely from all unambigious OGs and not only the the pan-OGs. The second difference lies in how the programs compute a supertree from the set of input trees. Consense heuristically composes the supertree by simply counting for each partition how often it occurs in each tree. This is a debatable definition, because the tree editing distance (and thus the supertree) between trees can be expressed not only in terms of how many partitions they share (which is implicitly used by Consense), but also by other measures such as cutting and pruning distances. CLANN explicitly and more systematically searches for the average supertree, scanning for the tree that is closest to all trees in the input data using a balanced tree distance method that combines partition sharing and "grafting and pruning".

To test what effect the increase in the amount of high quality data had on the supertree, we composed Consense supertrees from an increasing number of phylogenetic distance trees of the most restrictive set of OGs, the pan-triOGs. Figure 29 shows that although the individual phylogenetic trees were all different (high standard deviation on the left), by combining the data, the supertrees converge toward the external golden standard, reaching the golden standard phylogeny from a minimum of 30-40 combined phylogenetic trees. Thus, even though we only use 59 pan-triOGs, the Consense supertree is well into the flat range of the curve, showing that this relatively small amount of data (on average, 0.7% of the genome) is sufficient for recovering the target phylogeny.

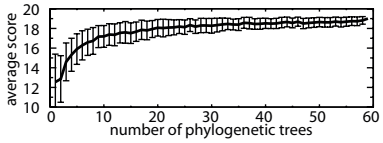

Figure 29. Including more high quality data leads to a better phylogeny. Supertrees (Consense) were composed from an increasing number of phylogenetic distance trees of pan-triOGs. The average number of target nodes recovered correctly increases, and finally the phylogenomic supertree composed from all 59 pan-triOGs recovers all 19 target nodes correctly. Average ± standard deviation of 100 random sub-samplings of the pan-triOGs.

Supplemental figures for "Signature genes as a phylogenomic tool"

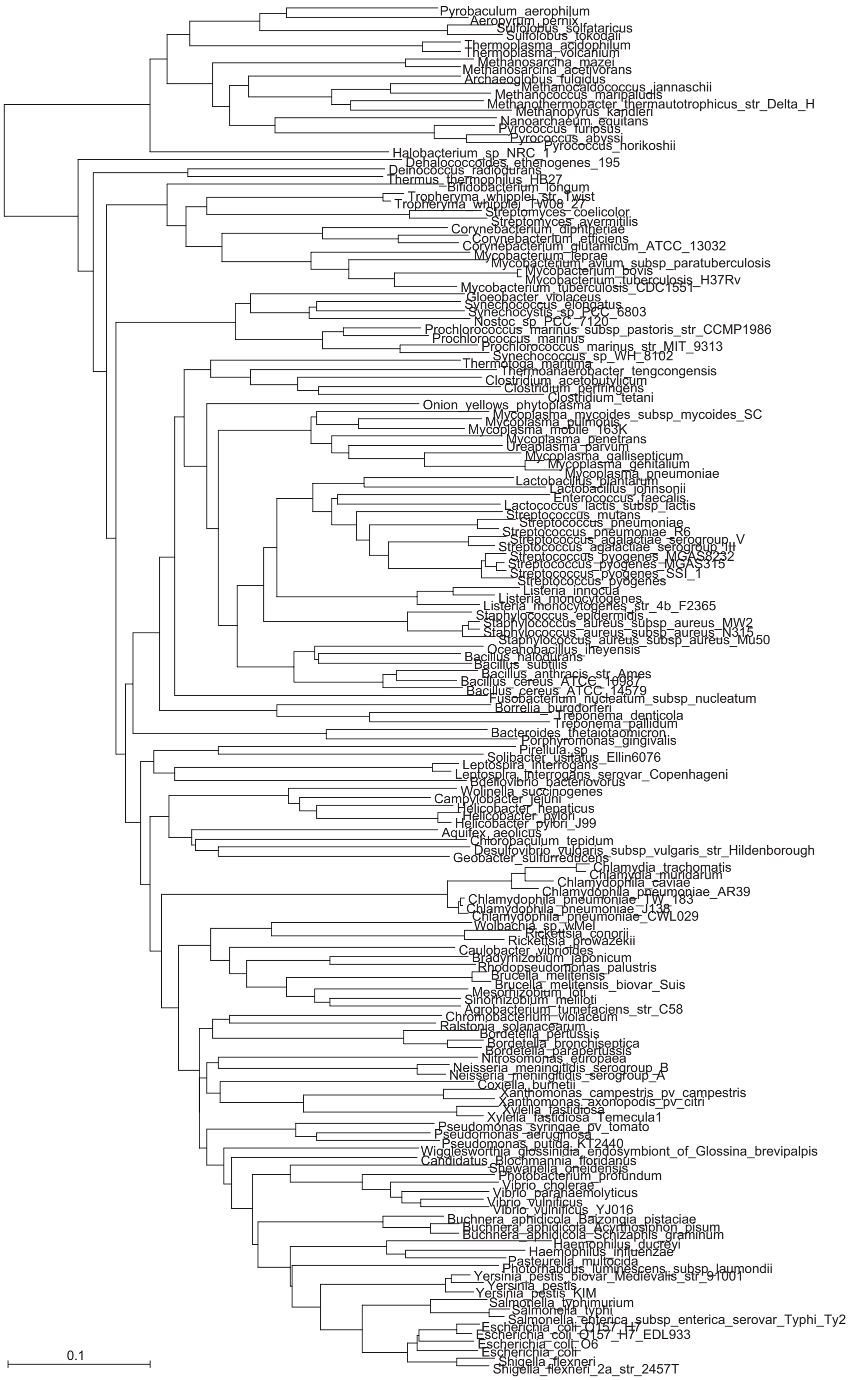

Figure 30. Gene content tree of completely sequenced prokaryotes. The tree was composed using the SHOT algorithm (Korbel et al. 2002) to calculate distances between genomes, and BioNJ (Gascuel 1997) to make a neighbor-joining tree.

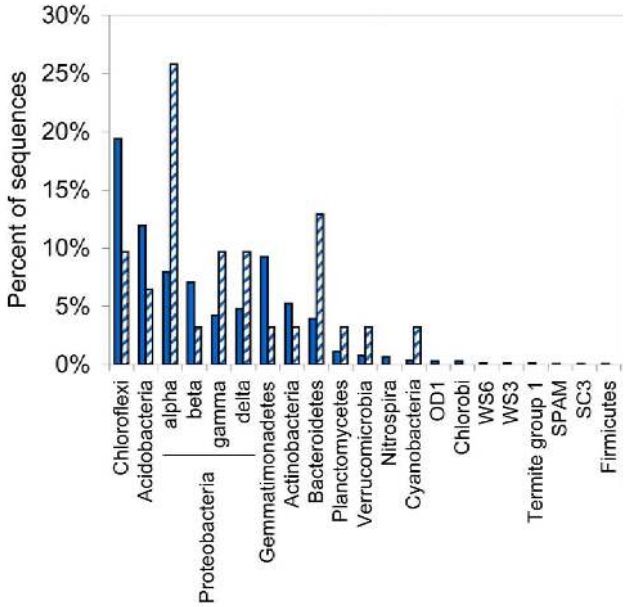

Figure 31. rRNA analysis of soil. Phylogenetic distribution of soil 16S rRNA sequences from PCR clone library (solid) and genomic library (hatched). Figure S2B from (Tringe et al. 2005).

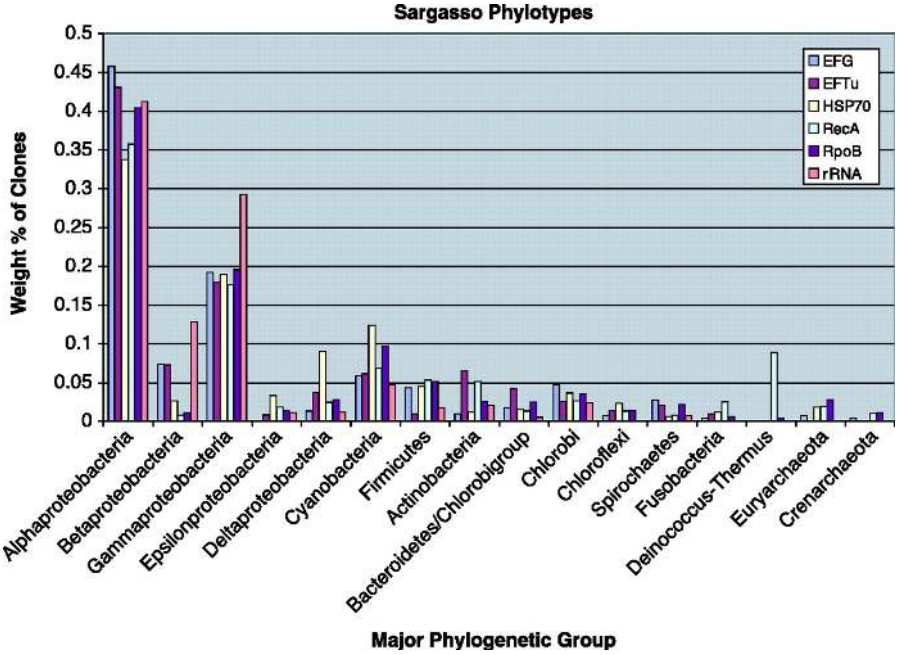

Figure 32. Phylogenetic diversity of Sargasso Sea sequences using multiple phylogenetic markers. The relative contribution of organisms from different major phylogenetic groups (phylotypes) was measured using multiple phylogenetic markers that have been used previously in phylogenetic studies of prokaryotes: 16S rRNA, RecA, EF-Tu, EF-G, HSP70, and RNA polymerase B (RpoB). The relative proportion of different phylotypes for each sequence (weighted by the depth of coverage of the contigs from which those sequences came) is shown. Figure 6 from (Venter et al. 2004).

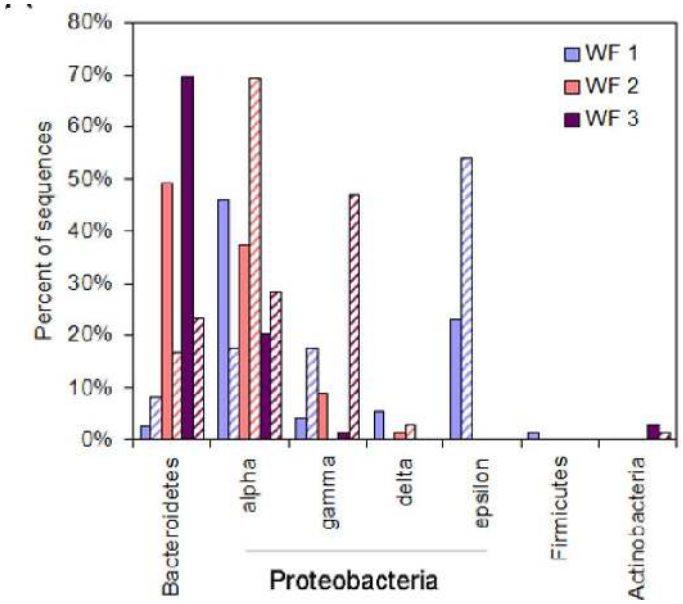

Figure 33. Rank-abundance curves for whale fall bacterial 16S sequences. Assignment of 16S rRNA sequences to bacterial phyla for both PCR clone libraries (solid bars) and genomic libraries (hatched bars). WF 1, Santa Cruz bone; WF 2, Santa Cruz microbial mat; WF 3, Antarctic bone. Figure S4A from (Tringe et al. 2005).

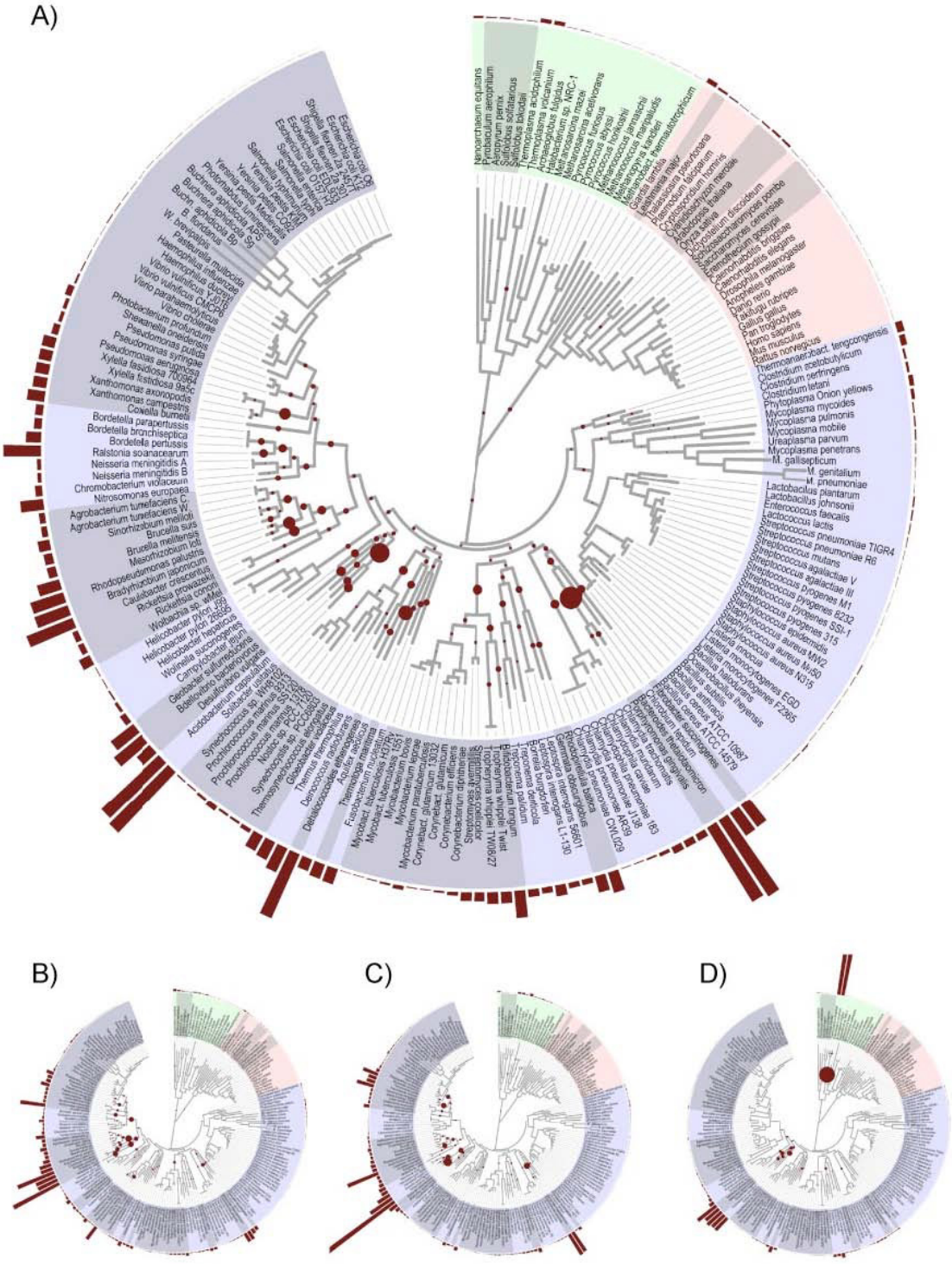

Figure 34. Phylogenetic distribution of communities, separately for each environment: A) agricultural soil; B) surface ocean water; C) deep sea whale bone; and D) acidic mine drainage (not part of the current analyses). Figure S1 from (von Mering et al. 2007a).

Publications

For updates see www.cmbi.ru.nl/~dutilh/publications.

Bas E. Dutilh, Berend Snel, Thijs J.G. Ettema and Martijn A. Huynen (2007), "Signature genes as a phylogenomic tool", submitted.

Philip R. Kensche, Vera van Noort, **Bas E. Dutilh** and Martijn A. Huynen (2007), "Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution", *Journal of the Royal Society Interface*.

Nicole A Datson, Maarten C Morsink, Srebrena Atanasova, Victor W Armstrong, Hans Zischler, Christina Schlumbohm, **Bas E Dutilh**, Martijn A Huynen, Brigitte Waegele, Andreas Ruepp, E. Ronald de Kloet and Eberhard Fuchs (2007), "Development of the first marmoset-specific DNA microarray (EUMAMA): a new genetic tool for large-scale expression profiling in a non-human primate", *BMC Genomics* **8**: 190.

Bas E. Dutilh, Vera van Noort, René T.J.M. van der Heijden, Teun Boekhout, Berend Snel and Martijn A. Huynen (2007), "Assessment of phylogenomic and orthology approaches for phylogenetic inference", *Bioinformatics* **23**: 815-824.

Marc Strous, Eric Pelletier, Sophie Mangenot, Thomas Rattei, Angelika Lehner, Michael W. Taylor, Matthias Horn, Holger Daims, Delphine Bartol-Mavel, Patrick Wincker, Valérie Barbe, Nuria Fonknechten, David Vallenet, Béatrice Segurens, Chantal Schenowitz-Truong, Claudine Médigue, Astrid Collingro, Berend Snel, **Bas E. Dutilh**, Huub J. M. Op den Camp, Chris van der Drift, Irina Cirpus, Katinka T. van de Pas-Schoonen, Harry R. Harhangi, Laura van Niftrik, Markus Schmid, Jan Keltjens, Jack van de Vossenberg, Boran Kartal, Harald Meier, Dmitrij Frishman, Martijn A. Huynen, Hans-Werner Mewes, Jean Weissenbach, Mike S. M. Jetten, Michael Wagner and Denis Le Paslier (2006), "Deciphering the evolution and metabolism of an anammox bacterium from a community genome", *Nature* **440**: 790-794. F1000 Exceptional.

Guenola Ricard, Neil R. McEwan, **Bas E. Dutilh**, Jean-Pierre Jouany, Didier Macheboeuf, Makoto Mitsumori, Freda M. McIntosh, Tadeusz Michalowski, Takafumi Nagamine, Nancy Nelson, Charles J. Newbold, Eli Nsabimana, Akio Takenaka, Nadine A. Thomas, Kazunari Ushida, Johannes H.P. Hackstein and Martijn A. Huynen (2006), "Horizontal Gene Transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic carbohydrates rich environment", *BMC Genomics* **7**: 22. F1000 Recommended.

Bas E. Dutilh, Martijn A. Huynen and Berend Snel (2006), "A global definition of expression context is conserved between orthologs but does not correlate with sequence conservation", *BMC Genomics* **7**: 10. BMC highly accessed.

Berend Snel, Martijn A. Huynen and **Bas E. Dutilh** (2005), "Genome trees and the nature of genome evolution", *Annual Reviews in Microbiology* **59**: 191-209.

Bas E. Dutilh, Martijn A. Huynen, William J. Bruno and Berend Snel (2004), "The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise", *Journal of Molecular Evolution* **58**: 527-539. F1000 Must Read.

Bas E. Dutilh and Rob J. de Boer (2003), "Decline in excision circles requires homeostatic renewal or homeostatic death of naive T cells", *Journal of Theoretical Biology* **224**: 351-358.

Resume

Bastiaan Elie Dutilh (Bas) was born in Utrecht, The Netherlands, on May 2nd 1976. He studied biology at Utrecht University from 1994 to 1999. During his first internship (Utrecht University, 1996) with dr. Ad Borstlap, he studied the transport of the species specific carbohydrates sorbitol and mannitol across the cell membrane of *Plantago major* (broadleaf plantain) and *Apium graveolens* (celery) phloem cells, respectively, using radioactive labeling. During this work, Bas showed that the control solution without membrane vesicles could take up as much radioactivity as the solution supposedly containing vesicles, implying that it is pretty difficult to make membrane vesicles from vascular tissue.

His second internship (Australian National University, Canberra, 1997) with prof. David Day and prof. Hans Lambers was aimed at determining the activity and genetic sequence of the alternative oxidase (AOX) in *Lycopersicon esculentum* (tomato), which involved working with substances like myxothiazol and ethidium bromide. Part of the gene was sequenced, showing homology to *Aox* genes in other plants, and AOX activity was found in several of the *Lycopersicon* tissues, especially in green fruit after a seven day cold treatment.

In his third and last internship (Utrecht University, 1998) with prof. Paulien Hogeweg, Bas finally entered the safe harbour of theoretical biology. He used a cellular automata model to study the structuring of the genotypes in a population of viruses evolving under the selection pressure of the host immune system. Pinpointing what was overlooked in a recent Science paper (Gupta et al. 1998), he showed that while the viruses optimize their infection rate by structuring into a set of minimally overlapping genotypes, the spatial structure of a host population relieves the viral load by favouring different sets of viral strains. He warns that excessive long-distance travelling allows the viruses to escape from this constraint, resulting in an increase in the number of infections.

Writing his master thesis "Gene networks from microarray data - analysis of data from microarray experiments, the state of the art in gene network reconstruction" (Utrecht University, 1999) under the supervision of prof. Paulien Hogeweg, was Bas' first venture into the area of "static bioinformatics". The thesis has been cited internationally, and has been recommended by educators as a valuable essay providing a comprehensive overview of the state of the art in the reconstruction of gene networks from large scale expression data (unpublished, available at www-binf.bio.uu.nl/~dutilh/gene-networks).

After Bas graduated, dr. Rob de Boer gave him his first employment in science, hiring him as a junior researcher in theoretical immunology, paid by the Sanquin Blood Foundation (Utrecht University, 2000). Using a mathematical model to describe the dynamics of small, unreplicated circles of DNA in naive immune cells (T cell receptor rearrangement excision circles or TRECs), he challenged existing dogmas by showing that, given the decreasing function of the aging thymus, these dynamics can only be explained by assuming homeostasis of naive T cells (Dutilh and De Boer 2003).

He then continued working as a junior researcher with dr. Chris Dutilh (Dutilh BOSA, 2001), among other things investigating which initiatives exist around the world to monitor and improve sustainable development in the food industry. This led to an overview that was published online by the Foundation for a Sustainable Food Chain DuVo (Dutilh et al. 2001).

In 2002, Bas started working with prof. Martijn Huynen at the Center for Molecular and Biomolecular Informatics (CMBI) and the Nijmegen Center for Molecular Life Sciences (NCMLS). As a junior researcher, he used a vast bioinformatics toolbox to study the carbohydrate metabolism of three reductive *Pyrococcus* species, and spent a while trying to predict interacting genes by finding conserved sharing of dimeric motifs in a range of genomes. Other research areas that were explored during the years spent at the CMBI are described in this book.

Currently, Bas is still working with prof. Martijn Huynen, collecting the available genomic data on the skin development protein P63, and assembling and analyzing the genomes in a methanotrophic microbial community, a project resulting from a Horizon grant obtained with dr. Marc Strous.

Acknowledgements

Many thanks to everyone, within and outside the CMBI, who helped to make this possible! Personal acknowledgements are provided on request.