15

From sequence reads to evolutionary inferences

JAMES A. COTTON

15.1 Introduction

The history of molecular systematics can be caricatured as one of ever-increasing depth of sequence data, analysed by ever more complex models. In this respect, sequence data from whole genomes are the ultimate source of molecular markers that can act as characters for phylogenetic or population genetic analysis. While complete genomes in the strictest sense are only available for very few species, and fragmentary genome assemblies that capture the entire genome, but in many pieces, are also fairly restricted in scope beyond the prokaryotes, this is changing rapidly. More-or-less shallow genomic data, for example from EST sequencing projects, high-throughput transcriptome sequencing or some other kind of reduced-representation sequencing (see review by Davey et al. 2011) are now becoming widespread and of increasing utility in systematics and other areas of evolutionary biology. Studies using these kinds of data to reconstruct relationships between species have become known as 'phylogenomics', although the original usage of the term referred to using phylogenetic approaches to infer gene function (Eisen 1998), and the other parts of the research program proposed under this name (Eisen and Fraser 2003) have been subsumed into the broader study of comparative and evolutionary genomics. Moreover, the term 'phylogenomics' has, perhaps, become over-extended, as datasets that claim this title vary in size from as few as 11 markers (Horvath et al. 2008) or as little as 30kb of sequence data (Wiegmann et al. 2011), and in eukaryotic organisms, the 'genomes' in question are very often organelle (mitochondrial or chloroplast) genome sequences. Sequence data from whole genomes have the potential to be a rich source of molecular phylogenetic markers for any systematic question, but there are two areas in which large-scale, highly multi-locus data appears most valuable – occupying the two extremes of the range of time scales over which inference about evolutionary history is made.

Genome-scale data promise the ability to resolve ancient divergences, and in particular, fairly rapid (at least in geological terms) ancient radiations that have been difficult to reliably reconstruct with more limited molecular datasets. In this context, phylogenomic data have been applied to a wide taxonomic range of phylogenetic questions. Early usage of whole genome data were in prokaryote systematics (e.g. Daubin et al. 2002; Daubin et al. 2002). Within the eubacteria genomic data has produced results largely congruent with the previously governing paradigm derived from ribosomal RNA (rRNA) data, but our understanding of the relationship between eubacteria, archaebacteria and eukaryotes – the most ancient phylogenetic divergence of all – has been greatly altered by multilocus and genomic data (Cox et al. 2008; Williams et al. 2012). In fact, whole-genome sequencing of microbial pathogen populations has now sufficiently widespread that it has become a routine tool in understanding the epidemiology of viral and bacterial pathogens (this volume, Chapter 5), and is emerging in eukaryotic parasites (Downing et al. 2011; Manske et al. 2012). In other areas, these kind of data have radically altered our picture of animal evolution

(Edgecombe et al. 2011), as well as resolving more-or-less deep divergences within animals (Struck et al. 2011; Wiegmann et al. 2011) and plants (Timme et al. 2012; Qiu et al. 2006; Lee et al. 2011).

On a far more recent timescale, many loci are needed to reliably reconstruct the history of populations, as the random process through which different alleles are inherited means that any individual locus is a poor marker of the history of the genomic ancestry of a population (e.g. Nichols 2001). These kinds of analyses require rich datasets of multiple genomes from a single species or population, and so have to date largely been restricted to analysis of human population history, but as whole genome re-sequencing becomes increasingly accessible, scientists working on any group of organisms will be able to benefit from population genomic data to understand the phylogeography, population genetics and adaptation of non-model systems (this volume, Chapter 10).

15.2 Generating the data: choosing between sequencing technologies and targets

Sequencing technology is changing fast (Thompson and Milos 2011), and any attempt here to propose particular choices of sequencing approach are likely to be out-of-date by the time this volume is published. At the time of writing, the choice appears to be between sequencing technologies that are relatively low-throughput and expensive per nucleotide sequenced, and relatively inaccurate in terms of the reliability of the sequence produced, but that produce long sequencing reads, and technologies producing more data much more cheaply, but only capable of producing short read lengths of contiguous bases (this volume, Preface). The choice of appropriate technology will depend on budget, the specific technical details of the technologies at the time any choice is made, and (of course) availability – although sequencing is likely to be increasingly out-sourced and so a choice of technology platforms likely to be available commercially to many researchers. A more scientific consideration is whether a high-quality reference genome is available for the species being targeted, and what kind of variants are of interest. Short-read sequencing is likely to be an attractive choice for the forseeable future if a good reference is available, where single nucleotide polymorphisms are of primary interest, and if the intention is to generate whole-genome data for substantial numbers of individuals, particularly for organisms with relatively large genome sizes. If more complex, structural or insertion-deletion variants are of more interest, and if the intention is to assemble genomes rather than rely on calling variants, then longer reads may be very valuable, and the additional cost (or reduced number of samples, or lower coverage) may be a worthwhile trade-off.

Another decision to make is whether to attempt whole-genome sequencing, or just to target particular regions or subsets of the genome. Clearly genome-wide data comes at a significant cost, particularly in eukaryotes with larger genome sizes, and in these organisms much of the genome may be repetitive (at least 50% and perhaps as much as 70% of the human genome is derived from repeats; Treangen and Salzberg 2011; de Koning et al. 2011), and so may be both relatively uninteresting and difficult to work with for down-stream analyses, where repetitive genomic regions are often ignored anyway. A number of approaches have been proposed for ‘reduced representation sequencing’, targeting either particular regions of genomes, or particular partitions of the genomes. At one end of this continuum, of course, are PCR-based approaches amplifying relatively few loci, building on traditional single-locus studies in a natural way (e.g. Bybee et al. 2011). Technologies exist for 'massively parallel' PCR amplification, exploiting either highly multiplexed traditional PCR (e.g. Nguyen-Dumont et al. 2013) or microfluidic technology (e.g. Tewhey et al. 2009) to amplify tens, hundreds or up to low thousands of loci from similar numbers of samples in parallel, but these are still some way from being truly 'genomic' approaches. A PCR-free approach is to design oligonucleotide 'baits' targeting particular regions of the genome, and then using hybridization to select the regions to be sequenced (e.g. Gnirke et al. 2009). Good reviews of these approaches and others have been provided by Mamanova et al. (2010) and Turner et al (2009).

For any such 'targeted' approaches, the success of the primers/baits is likely to be lower for targets that are more distantly related to the reference genome used to design them, and – perhaps more worryingly for some applications – to do so in a biased fashion, so that conserved regions will be more easily targeted in distantly related genomes than others. This has been exploited to design approaches to target homologous regions across taxonomic groups (Smith et al. 2013; Lemmon et al. 2012), but these data will be unsuitable for some applications as more variable regions of the genome will be less accessible. These approaches are a fairly natural evolution of 'traditional' molecular systematics approaches using PCR to amplify loci of interest, adapting this to the much larger throughput of next-generation sequencing technologies. Less biased genome-wide approaches rely on different ways of sequencing short 'tags' downstream of restriction sites. Variations on this idea of genome reduction or genome filtering have been around for some time (e.g. Altshuler et al. 2000; Whitelaw et al. 2003), but recent protocols (e.g. Baird et al. 2008) have made this an increasing popular approach in evolutionary biology (see Davey et al. 2011 for a review of a number of related approaches; Davey et al. 2011). An alternative is to target exonic sequence only, by sequencing and assembling transcriptomic (RNAseq) data (Gayral et al. 2013). In some circumstances, non-sequencing based approaches to genotyping might be of interest, such as oligonucleotide arrays for detecting known SNPs or other SNP genotyping technologies (Perkel 2008), but these technologies can only ascertain known variants, that must be discovered by some other approach (i.e. by sequencing) and they are unlikely to remain competitive with sequencing-based techniques for most applications as sequencing costs drop and throughput increases. Many of these alternatives are discussed and compared at greater length in other reviews of this area (e.g. Lemmon and Lemmon 2013; McCormack et al. 2013; Godden et al. 2012).

15.3 Making sense of the data

Initial quality control of next-generation sequencing data is to confirm that the molecular biology processes involved in creating and sequencing the DNA library have been successful. A first step involves confirming the yield and quality of sequencing reads, and for some technologies confirming that the sequence reads are of the desired size. This step can also involve identifying and either removing or trimming low-quality reads or reads contaminated with sequence from the adapters ligated onto the target DNA during library preparation, and detecting or (more controversially) attempting to correct biases caused by primers used in PCR of libraries, or in response to GC-content variation. A second step is to confirm that the library represents the target organism and that reads are in the expected orientation and distance apart for the intended sequencing strategy – this step is more computationally intensive and involves in some sense mapping the obtained reads (performing pairwise alignments between the reads and known sequences) against known data from the target organism or from related species. Mapping at least a random sub-sample of reads from a library against a larger set of off-target genomic data, representing possible sources of contamination such as yeast or bacterial species commonly used in molecular biology laboratories, or more specific cases such as against the host genome in the case of parasites or pathogens can also be useful in identifying contamination or other problems such as mislabeling of samples. Computationally efficient short-cuts that lose relatively little sensitivity are available to allow sequencing libraries to be routinely screened in this way (Wood and Salzberg 2014). A number of packages exist for more-or-less easy to interpret plots of base composition, sequence quality and throughput for next-generation sequencing platforms, including some packages with specific features useful for RNAseq data (DeLuca et al. 2012) and some QC work may be performed by sequencing providers if sequencing is outsourced. A comprehensive review (Zhou and Rokas 2014) includes further details of some of these steps, and reviews software available for each of these steps. An entire journal issue recently presented some detailed investigations of several QC procedures (Watson 2014), for example showing that even a poor-quality, automated assembly can be a useful substrate for quality assurance of sequencing data (Trivedi 2014).

Once high-quality sequence data have been generated, a number of different strategies are possible in moving from sequence reads to variants (Figure 1).

Fig 15.1 Reads, assembly and variation. A sequencing library is generated by first fragmenting a target genome at random, then generating sequence data, usually from both ends of each fragment (paired-end sequencing). The next step is either to identify overlapping reads within these data and merge them into a de-novo assembly, or to map the reads against a reference genome and identify variants. An assembly consists of contigs (in which sequence data is from a contiguous run of overlapping reads) and scaffolds, where contigs are joined by sequencing gaps that link paired reads, but for which no sequencing data is available (dashed line within scaffold). Red lines on the figure show positions of SNPs where the sequenced genome differs from the reference. Note also a deletion in the sequenced genome relative to the reference. This is identified either as regions where no sequence data has been assembled, or where both read and fragment coverage drop to zero when read pairs are mapped against the reference.

15.3.1 Mapping and variant calling

If genome sequence data are available for a closely related species to that being sequenced, an appropriate strategy is to *map* the reads against this genome, and then call SNP (single-nucleotide polymorphism) and copy-number variants using this. Mapping consists of taking each read and finding the position in a genome sequence to which this read is most similar, and then finding an alignment of the read to this position. This two-step approach is critical, computationally, as datasets may consist of millions or even billions of sequence reads, and alignment methods are too slow to place these reads. Two recent reviews (Fonseca et al. 2012; Li and Homer 2010) describe the variety of short read mapping algorithms and the software that implements them.

Variant calling is also a complex and fast-evolving area, as algorithms and software attempt to keep up with changes in sequencing technology and to the reality of large-scale resequencing projects directed at understanding population variation (see review by Nielsen et al. 2011; Nielsen et al. 2011). The basic task is to identify sights where single bases differ between the sequenced samples and the reference genome (a single-nucleotide polymorphism or SNP). More complex variants can be called, such as larger structural variants like large repeats or transpositions, and a great deal of computational research is ongoing into algorithmic approaches to discover these kinds of variants (Yalcin et al. 2012; Medvedev et al. 2009) . In general, discovery of large-scale variants will require long sequencing reads or long-fragment paired end sequence data. Many of the methodological developments in this area are being driven by the human genetics community, and in particular by large-scale human diversity projects such as the 1000 genomes project and UK10K (The 1000 Genomes Project Consortium 2013), but analysis of human data is made easier by some resources not available for most, if not all, other organisms – for example, the availability of large pre-existing data sets of validated SNP calls and fine-scale information about patterns and rates of recombination, both of which can be used to inform variant calling. Another product of large-scale human genome re-sequencing projects is the development of a series of file formats for different kinds of sequence and derived downstream data that, if not quite adopted standards, are convenient as they are written and read by a number of different software packages and are supported by utilities to interrogate these files and convert to and from other formats. Table 15.1 lists some of these file formats and relevant software.

Begin Table 15.1

Table 15.1
Most important file formats for genomic and sequence data, together with example software packages that write data in these formats or allow easy interrogation or manipulation of files.[1](Cock et al, 2010)· [2](Li et al, 2009), [4](Danecek et al, 2011),[5] (Gremme, Steinbiss, and Kurtz, 2013) ,[6](Quinlan and Hall, 2010). Bioinformatics toolkits for number of programming languages also provide libraries to access and manipulate these data, e.g. BioPerl (Stajich et al, 2002), Biopython (Cock et al, 2009), BioJava (Holland et al, 2008)

File formats	Data stored	Format specification or description	example software packages to write or manipulate data
Fasta	nucleotide or amino acid sequence data,	http://genetics.bwh.harvard.edu/pph/FASTA.html	many
Fastq	nucleotide sequence data with matching base quality scores. Two different encodings for base qualities exist[1].	Cock et al. 2010[1]	many
sam/bam/cram	text, binary and compressed binary files to store map positions of reads against a reference genome	Li et al. 2009[2]	SAMtools[2] NGSUtils[3] Picard (http://broadinstitute.github.io/picard)
vcf/bcf	text, binary files to store variant calls and genotypes	Danecek et al. 2011[4] http://www.1000genomes.org/wiki/analysis/variant-call-format/vcf-variant-call-format-version-42	Genome Analysis Toolkit VCFtools[4], NGSUtils[3]
gtf, gff	general feature format: annotation of sequence data, such as gene models,	http://www.sequenceontology.org/gff3.shtml	GenomeTools[5] BEDtools[6]
Bed	genome intervals – simple format for storing information about particular regions of a sequence	http://genome.ucsc.edu/FAQ/FAQformat.html	BEDtools[6] NGSUtils[3]

End Table 15.1

15.3.2 Genome Assembly

The second broad approach is to construct an assembly of each genome, and then find regions that are homologous to each other within each assembly that can then be used for phylogenetic or other comparative analyses. This is more informative when the different species/sequences being compared are more divergent, making aligning individual reads difficult. The basic principle of genome assembly is simple – by finding sequencing reads that overlap with one another, individual reads can be built up into longer and longer stretches of sequence data, representing pieces of the original genome, called *contigs*. The devil is in the detail, however – contigs can fail to be extended due to either repetitive sequences within the genome (so that there is ambiguity about how reads overlap with one another) or because of a failure to sequence a particular part of the genome.

Different assembly algorithms take different approaches to solving the complex patterns that can occur near repeat units, but also take different approaches to identifying overlaps – with modern data, it is no longer possible to simply compare every sequencing read pair-by-pair, and align them to identify optimal overlaps, as the amount of data is so great that this would take an impractical amount of time and computer memory. To improve efficiency, reads are broken down in to words of *k* bases (k-mers). Overlap-layout-consensus assemblers use the computed k-mers to find preliminary regions of overlap between entire reads (e.g. Mullikin and Ning 2003) before aligning reads that overlap and then finding a single layout of these reads that represent genome contigs (Miller et al. 2010). Assemblers for capillary sequence data all used this approach, but this has become impossible as the sizes of sequence data sets has increased. *De Bruijn graph* assemblers use the k-mers directly, rather than the reads themselves, and construct graphs representing overlaps between these k-mers before finding contigs as paths through this graph (Compeau et al. 2011). Popular De Bruijn graph assemblers for short read data are Velvet, Abyss and SOAPdenovo (Zerbino and Birney 2008; Simpson et al. 2009; Li et al. 2010). Overlap-layout consensus approaches are likely to see something of a resurgence as improving technology leads to increasing sequence read-lengths, meaning that fewer reads are needed for each genome region, and the benefit of keeping all of the read information increases. In particular, efficient data structures allows these approaches be used for much larger data sets (Simpson and Durbin 2012; Simpson and Durbin 2010).

The completeness and contiguity of the assemblies produced by these software packages can often be improved by using specific tools to use read-pair information to join contigs together (scaffolding) external to the assembly software itself (see Hunt et al. 2014 for a review and empirical comparison), and a number of other tools designed to improve on the output of assembly software are also available (e.g. Swain et al. 2012; Boetzer and Pirovano 2012). Finally, some guidance as to how to make an appropriate choice from the ever-growing range of assembly tools available for particular data types is now available from controlled experiments in which a number of different techniques have been applied to the same datasets (Bradnam et al. 2013).

A range of technologies and techniques exist for producing improved high-quality genome assemblies (Chain et al. 2009) such as optical mapping (e.g. Latreille et al. 2007) and new, emerging sequencing technologies that generate much longer reads that may span repetitive regions and so assemble more easily (e.g. Koren et al. 2013). These new technologies currently come at far greater cost per base-pair than short-read sequencing, and have higher error rates, but do hold great promise. Generating truly complete reference genomes is still far from a trivial or cheap exercise, but is at least now becoming achievable for non-specialist groups working outside genome centres. Of course, just as reduced-representation approaches can produce useful data when mapped against a reference genome, even fragmentary and incomplete genome assemblies can be excellent sources of information about evolutionary history, although for some analyses the inability to orientate and position loci with respect to one another may be problematic – for example if genetically unlinked loci are required, and even small errors in assembly can dramatically change interpretation of some genomic features (e.g. Parkhill 2002).

15.3.3 Emerging alternatives between de novo assembly and mapping

Whereas mapping and variant calling describe two principal approaches to initial analysis of sequence data, two other approaches are worth considering and are in some ways intermediate between these two approaches. Reference-guided assembly (e.g. Schneeberger et al. 2011; Schneeberger et al. 2011) allows reads to be assembled more-or-less independently, but using some information from some related genome sequence to help guide the assembly process. A related approach might be to generate sequence contigs using a de novo assembly process, and then order and orientate these contigs using some related genome sequence (Assefa et al. 2009). A final approach is to use algorithms used in assembly to produce a multi-species or multi-individual assembly graph that allows the identification of more complex variants between the samples than is possible using standard read mapping approaches (Iqbal et al. 2012), and that can call variants between samples even without the availability of a reference genome sequence.

15.4 The phylogenomic paradigm

Whether by assembly and subsequent alignment, or by directly calling variants from reads, the end result of the above steps will be a set of sequence data for different individuals that represent distinct species or populations of some evolutionary interest. For most systematists, reconstructing organism phylogeny is likely to be one of the primary interests in using genomic data. Phylogenetic inference from many loci, or at the whole genome-scale is now sufficiently commonplace that we can characterize (possibly stereotype) a 'standard' phylogenomic analysis. While the description below is to some extent a stereotype of the phylogenomic endeavor, the steps described above, with fairly limited variation, have become routine enough to be automated in pipelines for producing phylogenomic trees directly from input sequences, and a number of packages are available that automate some or all of the required steps (e.g. Jones et al. 2011; Wu and Eisen 2008; Dunn et al. 2013; Grant and Katz 2014). The profusion of these pipelines poses a particular challenge in validating each step of analysis (this volume, Chapter 1), and many of the considerations for these steps are also relevant for other evolutionary inferences, too.

15.4.1 Identifying Orthologs

Orthologs are gene copies descended from a speciation event, rather than from a gene duplication, so two orthologs are in an important sense the ‘same’ gene in two different genomes. For our purposes, the most important implication is that the evolutionary relationship of the two gene copies will most closely reflect the evolutionary history of the species themselves. An important first step in any phylogenomic analysis is thus to identify orthologous genes across the set of species included, and in particular orthologs that are present as single gene copies in all the species, so that there is little or no ambiguity about the relationships between the different copies. A traditional, standard approach has been to find ‘bidirectional best-hits’ between genes or proteins from different genomes under some similarity measure. This approach is unsatisfactory (Dalquen and Dessimoz 2013) and a number of more sophisticated approaches have been proposed: the OrthoMCL (Li et al. 2003) algorithm has been a standard approach for some years, but more recent approaches such as OMA (Roth et al. 2009), while less-used, may be more powerful in untangling the relationships between gene copies in complex families (see e.g. Altenhoff and Dessimoz 2012 for a recent review of algorithms). Phylogenetic approaches may be more powerful, for example in identifying many more single-copy orthologs than distance-based approaches (Vilella et al. 2008; Creevey et al. 2011), but are significantly more computationally expensive. Note, however, that patterns of gene duplication and loss themselves may contain phylogenetic signal, and some approaches will be discussed later that can deal successfully with multi-copy gene families without dividing them into single-copy subtrees.

15.4.2 Multiple Sequence Alignment

Orthology analysis identifies homology at the level of whole gene copies, but an additional step is necessary to deal with the smaller-scale insertions and deletions that occur during the course of molecular evolution within a locus. This is sequence alignment. There is evidence that alignment for many loci is difficult to get right, and that accuracy is critical to both phylogenetic reconstruction (Morrison and Ellis 1997; Ogden and Rosenberg 2006) and to other downstream molecular evolution analyses (e.g. Wong et al. 2008). Fast and accurate multiple sequence alignment approaches are now available, and an enormous body of research has contributed to these developments, with even a book-length collection of reviews available (Rosenberg 2011). The recognition that phylogeny and multiple sequence alignment both depend upon each other dates back to the earliest days of computational phylogenetics (Sankoff et al. 1973) – most multiple sequence alignment algorithms use a guide tree to reduce the computationally intractable multiple sequence alignment problem to a set of pairwise alignments. A more recent advance has been the development of practical algorithms that jointly estimate phylogeny and alignment (Liu, Raghavan, et al. 2009; Wheeler and Gladstein 1994). Of particular interest are algorithms that treat both alignment and phylogeny in a statistical framework, so that probability models of both single-site substitutions and insertion-deletion processes can be used, which should both improve the accuracy of alignment and allow accurate inference of these molecular evolutionary processes (Löytynoja and Goldman 2005; Westesson et al. 2012). A number of reviews have compared different alignment algorithms, estimating their accuracy based on protein families for which extensive structural data is available to provide gold-standard alignments that do not depend heavily on primary sequence similarity (Blackshields et al. 2006; Thompson et al. 2011).

15.4.3 Cleaning and pruning data

Most analyses of phylogenomic data have taken steps to filter the data to ensure quality. These steps vary a great deal, but one important step is to clean up automated sequence alignments and a number of tools exist to identify 'conserved blocks' – regions of the alignment in which most taxa are ungapped – in which alignment accuracy is higher and aligned characters are more likely to be homologous (Castresana 2000; Capella-Gutierrez et al. 2009). While some authors have criticized these approaches as discarding too much data (Wu et al. 2012; Wu and Eisen 2008), this may be of secondary concern in genome-scale datasets. Alternative approaches have included hidden Markov model (HMM)-based (Wu et al. 2012) and other (Löytynoja and Milinkovitch 2001) approaches to identify reliable regions of alignments, and using a simple test of alignment quality (Landan and Graur 2007).

Some studies have advocated also building single-locus gene trees, and using these as a way to identify unreliable loci that should be excluded from any concatenated dataset for final analysis, for example to exclude potential lateral gene transfer events in resolving prokaryote phylogeny (Ciccarelli et al. 2006). While loci that are misleading about phylogenetic relationships are precisely those we want to avoid this approach has not been widely adopted, presumably because it appears circular – identifying surprising locus-specific phylogenies is only possible if we have some idea of the correct tree – and because it is increasingly clear that we expect a significant degree of incongruence between loci (Salichos and Rokas 2014). Other studies have suggested removing quickly evolving species, or species that are 'unstable' in the sense that gene trees do not agree on their phylogenetic placement (see Wilkinson 2006 for reference to some methods to indentify such species), where species for which many loci are missing (Lemmon et al. 2009) but see (Roure et al. 2012), or by removing genes that are evolving exceptionally quickly (Rodríguez-Ezpeleta et al. 2007; Pisani 2004), There is some empirical data suggesting that keeping only phylogenetically 'decisive' genes in the sense that they have high gene-specific bootstrap support or by some other measure may give better results in studies aiming to resolve difficult, ancient divergences (Salichos and Rokas 2014).

15.4.4 Phylogenetic analysis of concatenated data

The final stage of a standard phylogenomic analysis is to concatenate the alignments for each locus remaining after the cleaning step, to produce a single 'supermatrix' of the aligned genome data. As with any molecular phylogenetics, there is a bewildering array of choices of model and algorithmic approach to inferring the correct phylogeny for these data. Whereas distance-based methods are extremely fast and scale to large datasets well (Gascuel and Steel 2006), and very fast algorithms for maximum-parsimony inference are available (e.g. Goloboff et al. 2008), the most popular and most accurate approaches are probabilistic inference using either maximum-likelihood or Bayesian approaches. Applying these methods to large datasets has now become computationally feasible thanks to computational developments in both serial (Stamatakis 2014; Guindon and Gascuel 2003) and parallel algorithms for calculating likelihoods for sequence data on phylogenies (Flouri et al. 2015; Suchard and Rambaut 2009). One problem for analyzing very large data matrices is that bootstrapping – the standard method for assessing the robustness of inferred phylogenies – is computationally demanding. However, both fast approaches to approximate bootstrapping (Stamatakis et al. 2008) and some alternative approaches (e.g. Anisimova and Gascuel 2006) exist, and bootstrap analysis is by its very nature easily performed in parallel where computing resources are available.

15.5 Phylogenomics: the end (and beginning) of incongruence?

We might expect that genomic data where sufficiently powerful and informative that the 'right answer' would emerge from relatively simple analysis of these data without the need for much methodological consideration. This promise initially appeared to be met. The first genuinely genomic phylogenomic studies had to wait for the availability of genome-wide sequence data for several closely related organisms. Perhaps the first such study was in yeast closely related to the laboratory model *Saccharomyces cerevisiae* (Rokas et al. 2003) and appeared to demonstrate the enormous potential of such data to resolve difficult phylogenetic issues, as molecular data for 106 conserved orthologs from 8 yeast species produced a fully-resolved phylogeny with perfect (100% for every node) bootstrap support. Optimism was rather short-lived; it soon became clear that even for this small-scale problem, genomic data would not be 'ending incongruence' as an accompanying editorial suggested (Gee 2003). Re-analyses of the same alignments with other phylogenetic methods produced different phylogenies

(Phillips 2004), also often with 100% bootstrap support. This was a timely reminder that bootstrap support values are model- or method-dependent, but also highlighting that, with large amounts of data, small biases in the methods used can drive you to be very confident about incorrect results. Indeed, this early, truly genome-wide 'phylogenomic' dataset has become something of a platform for investigating these kinds of issues (e.g. Holland 2004; Phillips 2004; Taylor 2004; Fedrigo et al. 2005; Jeffroy et al. 2006; Hess and Goldman 2011; Gatesy and Baker 2005 and many others).

This challenge has persisted in other taxonomic groups. Perhaps paradoxically, the availability of very large genomic datasets has not brought an end to concerns about incongruence, but rather has brought this, and a number of other methodological issues in phylogenetics into much sharper relief. Even within animal phylogenetics, despite (or perhaps, because of) centuries of systematic effort, controversy over important relationships has long persisted despite the availability of large molecular phylogenetic datasets. One clear, if now largely resolved, example of this is the debate over the correct relationships between major animal lineages, which focused on whether such morphologically different groups as annelids and nematodes formed part of a clade known as Ecdysozoa (named after the shared presence of a moulted cuticle; Aguinaldo et al. 1997) rather than the apparently more intuitive historical grouping – dating back at least to Cuvier in 1817 – of segmented animals such as annelid worms and arthropods in an Articulata clade (Scholtz 2002) within a wider group of coelomate animals, to the exclusion of the acoelomate nematodes (Rhaesa et al. 1998). In this case, careful analyses of single or few-locus data sets supported Ecdysozoan monophyly (e.g. Aguinaldo et al. 1997; Mallatt et al. 2004), while genome-wide datasets almost universally decisively rejected this relationship (e.g. Philip et al. 2005; Blair et al. 2002). After much work, it seems that genome-scale analysis of the Ecdysozoa vs. Coelomata seems to get 'fooled' by the high rate of evolution in the nematode *Caenorhabditis elegans* – at the time the only nematode genome available (e.g. Philippe, Lartillot, et al. 2005). It took careful and sophisticated phylogenetic analyses to understand this issue ten years ago (see Telford et al. 2008 for an excellent discussion of this literature; Telford et al. 2008), but increasing taxonomic sampling of genome-scale data seems decisive in supporting an Ecdysozoa clade (e.g. Dunn et al. 2008). Similar factors may be at play in continuing uncertainty about other animal relationships, such as the most basal relationships among animal groups (e.g. Nosenko et al. 2013; Whelan et al. 2015).

Another salutatory example is a large-scale 'genomic' phylogenetic 'tree of life' (Ciccarelli et al. 2006) – including bacteria, archaea and eukaryotes – which was actually based on only a small fraction (around 1%) of loci from most genomes included (Dagan and Martin 2006), as problems of orthology or lateral gene transfer, detected as unusual locus-specific gene trees, led the authors to remove the vast majority of loci. Even these few loci appear to disagree significantly in phylogenetic signal (Bapteste et al. 2007). It seems that both careful identification of the subset of genes that represent the 'core' central inheritance of eukaryotes rather than later acquisitions from e.g. endosymbiotic lateral gene transfer from the mitochondrial ancestor or other, later LGT events (Pisani et al. 2007) and careful phylogenetic modeling of the resulting genes is needed to correctly identify the relationships between eukaryotes and archaeal lineages (e.g. Cox et al. 2008; Williams et al. 2012; see Williams et al. 2014 for a recent review), a view confirmed by the discovery of a planktonic archaeal group related to the eukaryotic ancestor (Spang et al. 2015).

These results have brought an increasing focus on the particular challenges of phylogenetic inference from large-scale datasets (Philippe, DELSUC, et al. 2005; Rannala and Yang 2008; Kumar et al. 2012; Lemmon and Lemmon 2013).

15.5.1 Correct modeling of the substitution process can be critical

Standard time- and sequence-homogenous models such as the Jukes-Cantor and General Time-Reversible (GTR) models used in standard molecular phylogenetics all attempt to capture similar kinds of variation: variation in the frequency of different bases, and variation in substitution rates between pairs of bases. The most common extensions to these models attempt to model variation in the rate of evolution across alignment sites – either including a subset of 'invariant' sites that are a priori assumed not to change across the tree (this is different to sites that are not 'invariant' a priori but just happen not to show any observed change in the sample included), or by assuming some distribution (typically a discretized gamma distribution) of evolutionary rates across sites (Yang 1996b).

One important issue is for models to capture more subtle variation in the process of evolution across such very large alignments. Since the appearance of software able to fit different substitution models to different subsets of data (Yang 1996a; Nylander et al. 2004), a common approach has become to 'partition' the alignment, so that different sets of alignment positions can have different substitution models, or at least different inferred parameters of a model (figure 2). Most commonly, partitions are identified *a priori*, for example to choose a different partition for each locus (e.g. Nylander et al. 2004), or analyse first, second and third codon positions as separate partitions for protein-coding data (Shapiro 2005), and partitioning data can substantially improve how well substitution models fit the data (e.g. Hess and Goldman 2011). Other approaches are possible: standard statistical model choice criteria have been used to choose between candidate partitioning strategies (e.g. Brown and Lemmon 2007; and see Blair and Murphy 2010 for a far more extensive discussion than we have space for), and one algorithm attempts to identify a statistically optimal partitioning of the data from the (very large) possible number of schemes (Lanfear et al. 2012).

Fig 15.2 Two problems in phylogenomic inference. (a) The most basic approach is to concatenate data from multiple loci, and analyse the combined ‘supermatrix’ alignment using a single model of sequence evolution. More complex analyses might allow the rate of evolution to (b), and parameters of the substitution model (c) to vary between loci. More realistic, but more complex approach might be to allow rates of substitution (d) and even substitution model to vary between branches for each gene. The most general approach would also allow each gene to have a different topology (e) or even sample loci from overlapping but different sets of taxa. In these cases, a model of incongruence between gene tree topologies, or some non-parametric method is needed to infer a single species tree. Different Shades of grey indicate different loci and the model components (substitution matrices, trees and branch lengths) that apply to those loci, while model components in black apply across all partitions. (f) Different processes that introduce differences between the inferred phylogeny for a locus or gene family (in solid lines) and that for the species or populations they are sampled from (shown in outline).

More complex models – just as the widely used CAT model (Lartillot and Philippe 2004), for example, take a different approach – rather than assigning sites into different categories with different modes of evolution, these 'mixture models' assume that sites evolve under a combination of a small set of models, with parameters describing these models and the relative contribution of each of the set to the dynamics at that site. A related, but different approach is to separately fit a panel of possible models to the data, and then average across the various parameters inferred under these models in proportion to how well each model fits the data (Posada and Buckley 2004). Mixture models – generally fitted in a Bayesian MCMC framework to allow the many parameters to be reasonably efficiently estimated – often do an excellent job of fitting variation in amino acid composition between sites, and often do so with remarkably few different model components required, but the models discussed above do nothing to capture variation in substitution rates not driven by these compositional differences, and do not attempt to model variation in substitution processes through evolutionary time.

This kind of variation, equivalent to variation in substitution parameters across branches of a phylogenetic tree rather than across the genome, is more difficult to capture. Early attempts sought to allow variation in DNA composition (Galtier and Gouy 1998), but more sophisticated approaches allow every branch to evolve under a mixture model across branches (Pagel and Meade 2004), or across both sites and branches (Blanquart and Lartillot 2008) or allow data to infer how the process of evolution varies between a set of models (Foster 2004; Whelan 2008). A different approach is a so-called covarion model, in which sites can switch between invariant and variable states along a tree (Penny et al. 2014). The end-result of this hierarchy of increasing model complexity is the so-called general markov model (Barry and Hartigan 1987), in which every branch has a unique substitution process. This model is too parameter-rich for general use (adding each additional taxon to a phylogeny adds 24 additional parameters), but efficient inference consistent with this model is possible for quartets of taxa (Holland et al. 2012). The fully general substitution model is the most extreme example of a general trade-off between bias and variance, as the information available in the data is used to infer parameters that, while they may be necessary to correctly capture the process of molecular evolution, may be incidental to the biological questions being addressed, and there are clear limitations to this approach (Steel 2005). Systematic error caused by a lack inadequate models of evolution has long been recognized as a problem in phylogenetic inference (Swofford et al. 1996; Erixon et al. 2003) but is likely to be more so in phylogenomic data where sampling error is small (Rodríguez-Ezpeleta et al. 2007) (Kumar et al. 2012).

15.5.2 Concatenating data can mislead

The advances described above in careful alignment, orthology analysis and modeling of the substitution process help ensure correct inference of the phylogeny underlying a set of sequence data, so that for a single locus an accurate ‘gene tree’ can be obtained. In most cases, however, the aim is to infer a correct phylogeny relating a set of individuals represented by sequence data from many loci. Concatenating these loci into a large matrix and then inferring a single phylogenetic tree from these data is only a sensible thing to do if a single tree relating these genomes – often called the ‘species tree’, as the individuals are representing some taxonomic group – really exists. While the sampled genomes must be related by some evolutionary process of descent with modification, two things can happen to make searching for a single tree from many loci in this way misleading. One is that the process of evolution may not be simply bifurcating divergence, either due to processes like large-scale lateral gene transfer, allopolyploidy or introgression between populations, as in the case of endosymbiotic gene transfer mentioned above. A second problem is that evolutionary processes within the genome – gene duplication, gene loss and the coalescent process by which the different alleles present in extant populations are sampled from an ancestral population – can lead to different loci having different underlying trees, and that the best way to infer the phylogeny of the populations themselves may not be to simply join together alignments (figure 2).

An alternative to combining alignments is to infer phylogenies for each locus separately before combining these distinct phylogenies. Choosing between these alternatives was a long-standing argument within molecular systematics about how to analyse data from different sources: initially fought beneath banners of 'total evidence' and 'consensus' and somewhat more recently between co-called 'supermatrix' or 'supertree' analyses (de Queiroz et al. 1995). These debates reflect a balance between concatenation to increase signal-to-noise, and so produce more precise and better supported results, and analyzing loci separately to avoid potential systematic errors from the assuming that a single tree underlies all of the loci analysed (Bapteste et al. 2007). These classic dichotomies have been somewhat resolved and replaced by a more sophisticated view of this choice as researchers have begun attempting to understand the biological processes that introduce difference between loci, and methods that attempt to directly address incongruence between loci directly have emerged.

Early methods were based on parsimony, arguing that the best estimate of a species tree is the one that minimized the number of gene duplications, gene losses or other evolutionary events that were required to explain a set of phyogenies for genes or gene families (Goodman et al. 1979; Page and Charleston 1997). More recently, statistical approaches have become more important, and a range of probabilistic models have been proposed to capture the relationship between gene tree and species trees (Szöllősi et al. 2015). Interestingly, this view of species tree inference as a hierarchy of inference, from basic sequence data at the base to species tree at the top, modeling gene and locus in-between (Szöllősi et al. 2015) is effectively realising a research program proposed as long ago as the 1970s, when sequence data from multiple species was available for just a handful of genes (Goodman et al. 1979). Seen in this context, classical methods for combining information from multiple trees such as consensus and supertree methods can be thought of as dealing with incongruence between gene or locus trees in a non-parametric way, avoiding the need to model complex processes such as coalescent lineage sorting, gene duplication and loss (Ané et al. 2007; Cotton and Wilkinson 2009 and figure 2). Such 'non-parametric' approaches might seem attractive, avoiding the need for complex methods and models to capture biological processes happening above the nucleotide level that introduce incongruence between phylogenies at different loci. The flip side is that these models allow us to exploit sequence data to learn something about these processes, for example genome evolution (e.g. Cotton and Page 2005).

In fact, the most influential and widely-used of the methods used to infer species trees from multi-locus datasets model the coalescent process of allelic 'lineage sorting' in which multiple alleles in a population persist through divergence events, leading to the phylogeny drawn for samples of these alleles from different populations not reflecting the history of the populations themselves (Maddison and Knowles 2006; Rosenberg and Nordborg 2002; Nichols 2001). Indeed, the recognition that the coalescent process can lead to strange anomalies in which most genes can have misleading topologies (Degnan and Rosenberg 2006) has lead to multispecies coalescent methods (e.g. Edwards et al. 2007; Kubatko et al. 2009; see Liu, Yu, et al. 2009 for a review) now probably being the mainstream approach for reconstructing relationships between closely related species or between populations (see this volume, Chapter 1 for further discussion). Related models allow the identification of species by estimating where barriers to gene flow between population exist (e.g. Pons et al. 2006; Choi and Hey 2011; see Carstens et al. 2013 for a recent review). Of course, the population genetics parameters that govern the coalescent process – genetic (effective) population sizes, dates of divergence between populations and others – are often of interest in themselves. Coalescent methods have long been proposed to learn about these population genetic parameters and processes from samples of molecular data (e.g. Ewens 1972), and this has become particularly important and widespread as sequence data for many individuals from populations are becoming available.

15.6 Population-level inference from genomic data

Including samples of multiple individuals from each population or species allows further insight into the population genetics of the organisms, and next generation sequencing has enabled powerful population genomics approaches to answering questions about adaptation (Pool et al. 2010), speciation (Sousa and Hey 2013), demography (Pool et al. 2010; Excoffier et al. 2013) and epidemiology (Kao et al. 2014). One approach to making use of genome-scale data is to arbitrarily 'chunk' the genome into contiguous pieces that are sufficiently widely spaced that recombination between them will be frequent and sufficiently short that recombination within them is rare. These pieces then approximate independent samples of the genealogical coalescent process. This approach is widely adopted (e.g. Gronau et al. 2011; Heled and Drummond 2010), but is far from ideal – not only does it discard much of the data, but recombination rates are sufficiently high that it is probably not possible to choose contiguous blocks of sequence that contain enough mutations to be informative while avoiding any recombination. Other approaches attempt to deal with inferring population genetic parameters for a recombining sequence, but the computational complexity of dealing correctly with this is prohibitive. A number of approximate methods to make use of genome-scale data efficiently in this way have been proposed. One class of methods (approximate Bayesian computation) uses extensive simulations to find evolutionary scenarios that are expected to produce genome data in some sense similar to observed data values (Csilléry et al. 2010; Beaumont 2010 are two recent reviews). Another approach is to fit probabilistic models to the site frequency spectrum (the distribution of allele frequencies across sites) for each population (e.g. Gutenkunst et al. 2009; Excoffier et al. 2013) – a summary statistic that captures much of the signal of selection and demographic change in a sample of sequence data. A final approach is to use a computationally convenient approximation to the coalescent process itself (McVean and Cardin 2005; Hobolth et al. 2007). These methodological developments have led to general and powerful approaches to using genome-wide data to understanding the genetic history of populations.

In fact, insight into many of these areas can be obtained even with a traditional 'molecular systematics' sample of a single individual per species or population. Some of the power that comes from sampling many individuals is present in a sample of many loci (e.g. genomic data) from single individuals, as recombination induces a more-or-less independent evolutionary history for each locus as time progresses, each subject to, and informative about, the population genetic processes that govern their joint evolution. This insight has led to methods for inferring population genetic parameters such as splitting times and rates of gene flow or migration between populations and effective population sizes of ancestral populations from samples of a single genome from each population (e.g. Rannala and Yang 2003; Hobolth et al. 2007), and to infer how the size of a population has varied through time (Li and Durbin 2011) and the history of admixture into a population (Harris and Nielsen 2013) from a single diploid genome sampled from it. A particular complication with sampling a single genome, or small number of genomes, from a population for many organisms is that phasing – inferring the sequence of each haplotype from a set of genotypes (Browning and Browning 2011) – is difficult or impossible based on such small samples, so extending these methods to use multiple samples requires integrating across the uncertainty in phase (e.g. Schiffels and Durbin 2014; Gronau et al. 2011).

A particular power of genome-wide population data is that, to a certain extent, the background pattern of variation across the genome acts as a control for many processes that affect the entire genome – such as demographic forces like population bottlenecks, or the effect of life history – and genomic loci that are outliers in some sense from this background pattern are likely to be of interest in some way. A range of methods for inferring selection from genetic, and more recently genomic, data have been proposed (see Vitti et al. 2013; Nielsen et al. 2007; Scheinfeldt and Tishkoff 2013 for recent reviews). These methods vary in what kind of signals of selection they look for – e.g. alteration in the structure of linkage disequilibrium across the genome, variation in the allele frequency spectrum from neutral expectations or between populations, variation in the rate of substitutions between lineages or between synonymous and non-synonymous sites. These different signals differ in whether they are most sensitive to recent selection within a population or older selection acting between two reproductively isolated groups and whether they are sensitive only to the results of a classical selective sweep or can pick up the more subtle signs of selection acting on multiple loci or on existing variation. These genome-wide approaches raise the prospect of taking a ‘reverse genetics’ approach to understanding ecological adaptation – rather than identifying organismal traits thought likely to be adaptive, and then going on to identify loci responsible and ultimately how natural selection has worked at the genetic level, genome-wide scans for selection can potentially produce a list of loci under strong selection, and the functional significance of these then be followed-up in the field or lab. This may lead to an understanding of adaptation potentially less biased by our preconceptions of what could be the key traits governing a particular organisms fitness in its environment.

15.7 Conclusion

High-throughput ‘next generation’ sequencing data enables systematists to rapidly generate large multi-locus datasets with unprecented ease and at increasingly low cost. The cost of generating molecular phylogenetic data is lower than ever, but the difficultly of handling these data is greater than for traditional molecular data, and care is needed in analysis. Luckily, an extensive and growing ecosystem of software is now available for handling sequence data, and for interpreting genome-scale data in a phylogenetic context. It is tempting to think of ‘phylogenomics’ as super-sized traditional molecular phylogenetics, and this chapter describes what is currently a standard approach to analyzing phylogenomic data. However, massively multilocus datasets pose some unique challenges for phylogenetic inference – including some methodological challenges introduced by the simple scale of the data, but also more conceptual issues around how to best make use of multilocus data, and most excitingly, what we can learn from multilocus data that is not possible from individual loci. Particularly exciting is that the differences between loci can reflect evolutionary processes acting on both the genomic and population levels.

chapter-references

References

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S. et al. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387, 489–93.

Altenhoff, A. M., and Dessimoz, C. (2012). Inferring Orthology and Paralogy. In *Methods in Molecular Biology*, 855, 259–79.

Altshuler, D., Pollara, V. J., Cowles, C. R., et al. (2000). An SNP Map of the Human Genome Generated by Reduced Representation Shotgun Sequencing. *Nature*, 407, 513–16.

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian Estimation of Concordance Among Gene Trees. *Molecular Biology and Evolution*, 24, 412–26.

Anisimova, M., and Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: a Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55, 539–52.

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., and Berriman, M. (2009). ABACAS: Algorithm-Based Automatic Contiguation of Assembled Sequences. *Bioinformatics*, 25, 1968–69.

Baird, N. A., Etter, P. D., Atwood, T. S., et al. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Edited by Justin C Fay. *PLoS ONE*, 3, e3376.

Bapteste, E., Susko, E., Leigh, J., et al. (2007). Alternative Methods for Concatenation of Core Genes Indicate a Lack of Resolution in Deep Nodes of the Prokaryotic Phylogeny. *Molecular Biology and Evolution*, 25, 83–91.

Barry, D., and Hartigan, J. A. (1987). Asynchronous Distance Between Homologous DNA Sequences. *Biometrics*, 43, 261–76.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41, 379–406.

Blackshields, G., Wallace, I. M., Larkin, M., and Higgins, D. G. (2006). Analysis and Comparison of Benchmarks for Multiple Sequence Alignment. *In Silico Biology*, 6, 0030.

Blair, C., and Murphy, R. W. (2010). Recent Trends in Molecular Phylogenetic Analysis: Where to Next? *Journal of Heredity*, 102, 130–38.

Blair, J. E., Ikeo, K., Gojobori, T., and Hedges, S. B. (2002). The Evolutionary Position of Nematodes. *BMC Evolutionary Biology*, 2, 7.

Blanquart, S., and Lartillot, N. (2008). A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Molecular Biology and Evolution*, 25, 842–58.

Boetzer, M., and Pirovano, W. (2012). Toward Almost Closed Genomes with GapFiller. *Genome Biology*, 13, R56.

Bradnam, K. R., Fass, J. N., Alexandrov, A., et al. (2013). Assemblathon 2: Evaluating De Novo Methods of Genome Assembly in Three Vertebrate Species. *GigaScience*, 2, 10.

Brown, J. M., and Lemmon, A. R. (2007). The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics. *Systematic Biology*, 56, 643–55.

Browning, S. R., and Browning, B. L. (2011). Haplotype Phasing: Existing Methods and New Developments. *Nature Reviews Genetics*, 12, 703–14.

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., et al. (2011). Targeted Amplicon Sequencing (TAS): a Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. *Genome Biology and Evolution*, 3, 1312–23.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics*, 25, 1972–73.

Carstens, B. C., Pelletier, T. A., Reid, N. M., and Satler, J. D. (2013). How to Fail at Species Delimitation. *Molecular Ecology*, 22, 4369–83.

Castresana, J. (2000). Selection of Conserved Blocks From Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17, 540–552.

Chain, P. S. G., Grafham, D. V., Fulton, R. S., et al. (2009). Genomics. Genome Project Standards in a New Era of Sequencing. *Science*, 326, 236–37.

Choi, S. C., and Hey, J. (2011). Joint Inference of Population Assignment and Demographic History. *Genetics*, 540–552. 189, 561–77.

Ciccarelli, F., Doerks, T., Mering, von, C., et al. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311, 1283–87.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to Apply De Bruijn Graphs to Genome Assembly. *Nature Biotechnology*, 29, 987–91.

Cotton, J. A., and Page, R. D. M. (2005). Rates and Patterns of Gene Duplication and Loss in the Human Genome. *Proceedings of the Royal Society B: Biological Sciences*, 272, 277–83.

Cotton, J. A., and Wilkinson, M. (2009). Supertrees Join the Mainstream of Phylogenetics. *Trends in Ecology and Evolution*, 24, 1–3.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. (2008). The Archaebacterial Origin of Eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 20356–61.

Creevey, C. J., Muller, J., Doerks, T., et al. (2011). Identifying Single Copy Orthologs in Metazoa. *PLoS Computational Biology*, 7, e1002269.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in Practice. *Trends in Ecology and Evolution*, 25, 410–18.

Dagan, T., and Martin, W. (2006). The Tree of One Percent. *Genome Biology*, 7, 118.

Dalquen, D. A., and Dessimoz, C. (2013). Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades Such as Plants and Animals. *Genome Biology and Evolution*, 5, 1800–1806.

Daubin, V., Gouy, M., and Perrière, G. (2002). A Phylogenomic Approach to Bacterial Phylogeny: Evidence of a Core of Genes Sharing a Common History. *Genome Research*, 12, 1080–90.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., et al. (2011). Genome-Wide Genetic Marker Discovery and Genotyping Using Next-Generation Sequencing. *Nature Reviews Genetics*, 12, 499–510.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. Edited by Gregory P Copenhaver. *PLoS Genetics*, 7, e1002384.

de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate Versus Combined Analysis of Phylogenetic Evidence. *Annual Review of Ecology and Systematics*, 26, 657–681.

Degnan, J. H., and Rosenberg, N. A. (2006). Discordance of Species Trees with Their Most Likely Gene Trees. *PLoS Genetics*, 2, e68.

DeLuca, D. S., Levin, J. Z., Sivachenko, A., et al. (2012). RNA-SeQC: RNA-Seq Metrics for Quality Control and Process Optimization. *Bioinformatics*, 28, 1530–32.

Downing, T., Imamura, H., Decuypere, S., et al. (2011). Whole Genome Sequencing of Multiple Leishmania Donovani Clinical Isolates Provides Insights Into Population Structure and Mechanisms of Drug Resistance. *Genome Research*, 21, 2143–56.

Dunn, C. W., Hejnol, A., Matus, D. Q., et al. (2008). Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life. *Nature*, 452, 745–49.

Dunn, C. W., Howison, M., and Zapata, F. (2013). Agalma: an Automated Phylogenomics Workflow. *BMC Bioinformatics*, 14, 330.

Edgecombe, G. D., Giribet, G., Dunn, C. W., et al. (2011). Higher-Level Metazoan Relationships: Recent Progress and Remaining Questions. *Organisms Diversity and Evolution*. 11, 151–172.

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-Resolution Species Trees Without Concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 5936–41.

Eisen, J. A. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, 8, 163–67.

Eisen, J. A., and Fraser, C. M. (2003). Phylogenomics: Intersection of Evolution and Genomics. *Science*, 300, 1706–7.

Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003). Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. *Systematic Biology*, 52, 665–73.

Ewens, W. J. (1972). The Sampling Theory of Selectively Neutral Alleles. *Theoretical Population Biology*, 3, 87–112.

Excoffier, L., Dupanloup, I., Huerta-SÃ nchez, E., Sousa, V. C., and Foll, M. (2013). Robust Demographic Inference From Genomic and SNP Data. Edited by Joshua M Akey. *PLoS Genetics*, 9, e1003905.

Fedrigo, O., Naylor, G., and Collins, T. (2005). Choosing the Best Genes for the Job: the Case for Stationary Genes in Genome-Scale Phylogenetics. *Systematic Biology*, 54, 493–500.

Flouri, T., Izquierdo-Carrasco, F., Darriba, D., et al. (2015). The Phylogenetic Likelihood Library. *Systematic Biology*, 64, 356–62.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for Mapping High-Throughput Sequencing Data. *Bioinformatics*, 28, 3169–77.

Foster, P. G. (2004). Modeling Compositional Heterogeneity. *Systematic Biology*, 53, 485–95.

Galtier, N., and Gouy, M. (1998). Inferring Pattern and Process: Maximum-Likelihood Implementation of a Nonhomogeneous Model of DNA Sequence Evolution for Phylogenetic Analysis. *Molecular Biology and Evolution*, 15, 871–79.

Gascuel, O., and Steel, M. (2006). Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23, 1997–2000.

Gatesy, J., and Baker, R. (2005). Hidden Likelihood Support in Genomic Data: Can Forty-Five Wrongs Make a Right? *Systematic Biology*, 54, 483–92.

Gayral, P., Melo-Ferreira, J., Glémin, S., et al. (2013). Reference-Free Population Genomics From Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. *PLoS Genetics*, 9, e1003457.

Gee, H. (2003). Evolution: Ending Incongruence. *Nature* 425, 782.

Gnirke, A., Melnikov, A., Maguire, J., et al. (2009). Solution Hybrid Selection with Ultra-Long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nature Biotechnology*, 27, 182–89.

Godden, G. T., Jordon-Thaden, I. E., and Chamala, S. (2012). Making Next-Generation Sequencing Work for You: Approaches and Practical Considerations for Marker Development and Phylogenetics. *Plant Ecology and Diversity*, 5, 427–450.

Goloboff, P. A., Farris, J. S., and Nixon, K. C. (2008). TNT, a Free Program for Phylogenetic Analysis. *Cladistics*, 24, 774–86.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the Gene Lineage Into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms From Globin Sequences. *Systematic Zoology*, 28, 132–63.

Grant, J. R., and Katz, L. A. (2014). Building a Phylogenomic Pipeline for the Eukaryotic Tree of Life – Addressing Deep Phylogenies with Genome-Scale Data. *PLoS Currents Tree of Life*. 2014 Apr 2. Edition 1.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian Inference of Ancient Human Demography From Individual Genome Sequences. *Nature Genetics*, 43, 1031–1034.

Guindon, S., and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52, 696–704.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations From Multidimensional SNP Frequency Data. *PLoS Genetics*, 5, e1000695.

Harris, K., and Nielsen, R. (2013). Inferring Demographic History From a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, 9, e1003521.

Heled, J., and Drummond, A. J. (2010). Bayesian Inference of Species Trees From Multilocus Data. *Molecular Biology and Evolution*, 27, 570–80.

Hess, J., and Goldman, N. (2011). Addressing Inter-Gene Heterogeneity in Maximum Likelihood Phylogenomic Analysis: Yeasts Revisited. *PLoS ONE*, 6, e22783.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. (2007). Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred From a Coalescent Hidden Markov Model. *PLoS Genetics*, 3, e7.

Holland, B. R. (2004). Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny. *Molecular Biology and Evolution*, 21, 1459–61.

Holland, B. R., Jarvis, P. D., and Sumner, J. G. (2012). Low-Parameter Phylogenetic Inference Under the General Markov Model. *Systematic Biology*, 62, 78–92.

Horvath, J. E., Weisrock, D. W., Embry, S. L., et al. (2008). Development and Application of a Phylogenomic Toolkit: Resolving the Evolutionary History of Madagascar's Lemurs. *Genome Research*, 18, 489–99.

Hunt, M., Newbold, C., Berriman, M., and Otto, T. D. (2014). A Comprehensive Evaluation of Assembly Scaffolding Tools. *Genome Biology*, 15, R42.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De Novo Assembly and Genotyping of Variants Using Colored De Bruijn Graphs. *Nature Genetics*, 44, 226–32.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the Beginning of Incongruence? *Trends in Genetics*, 22, 225–31.

Jones, M. O., Koutsovoulos, G. D., and Blaxter, M. L. (2011). iPhy: an Integrated Phylogenetic Workbench for Supermatrix Analyses. *BMC Bioinformatics*, 12, 30.

Kao, R. R., Haydon, D. T., Lycett, S. J., and Murcia, P. R. (2014). Supersize Me: How Whole-Genomesequencing and Big Data Aretransforming Epidemiology. *Trends in Microbiology*, 22, 282–291.

Koren, S., Harhay, G. P., Smith, T. P., et al. (2013). Reducing Assembly Complexity of Microbial Genomes with Single-Molecule Sequencing. *Genome Biology*, 14, R101.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: Species Tree Estimation Using Maximum Likelihood for Gene Trees Under Coalescence. *Bioinformatics*, 25, 971–73.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2012). Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution* 29, 457–72.

Landan, G., and Graur, D. (2007). Heads or Tails: a Simple Reliability Check for Multiple Sequence Alignments. *Molecular Biology and Evolution*, 24, 1380–83.

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution*, 29, 1695–1701.

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

Latreille, P., Norton, S., Goldman, B. S., et al. (2007). Optical Mapping as a Routine Tool for Bacterial Genome Sequence Finishing. *BMC Genomics*, 8, 321.

Lee, E. K., Cibrian-Jaramillo, A., Kolokotronis, S.-O., et al. (2011). A Functional Phylogenomic View of the Seed Plants. *PLoS Genetics*, 7, e1002411.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., and Lemmon, E. M. (2009). The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Systematic Biology*, 58, 130–45.

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biolog, y* 61, 727–44.

Lemmon, E. M., and Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44, 99–121.

Li, H., and Durbin, R. (2011). Inference of Human Population History From Individual Whole-Genome Sequences. *Nature*, 475, 493–496.

Li, H., and Homer, N. (2010). A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing. *Briefings in Bioinformatics*, 11, 473–83.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13, 2178–89.

Li, R., Zhu, H., Ruan, J., et al. (2010). De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing. *Genome Research*, 20, 265–72.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009). Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science*, 324, 1561–64.

Comment [INT35]: Unlinked

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Molecular Phylogenetics and Evolution. *Molecular Phylogenetics and Evolution*, 53. Elsevier Inc., 320–28.

Löytynoja, A., and Goldman, N. (2005). An Algorithm for Progressive Multiple Alignment of Sequences with Insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10557–62.

Löytynoja, A., and Milinkovitch, M. C. (2001). SOAP, Cleaning Multiple Alignments From Unstable Blocks. *Bioinformatics*, 17, 573–74.

Maddison, W., and Knowles, L. (2006). Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology*, 55, 21–30.

Mallatt, J. M., Garey, J. R., and Shultz, J. W. (2004). Ecdysozoan Phylogeny and Bayesian Inference: First Use of Nearly Complete 28S and 18S rRNA Gene Sequences to Classify the Arthropods and Their Kin. *Molecular Phylogenetics and Evolution*, 31, 178–191.

Mamanova, L., Coffey, A. J., Scott, C. E., et al. (2010). Target-Enrichment Strategies for Next-Generation Sequencing. *Nature Methods*, 7, 111–18.

Manske, M., Miotto, O., Campino, S., et al. (2012). Analysis of Plasmodium Falciparum Diversity in Natural Infections by Deep Sequencing. *Nature*, 487, 375–79.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. (2013). Molecular Phylogenetics and Evolution. *Molecular Phylogenetics and Evolution* 66. Elsevier Inc., 526–38.

McVean, G. A. T., and Cardin, N. J. (2005). Approximating the Coalescent with Recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 1387–93.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational Methods for Discovering Structural Variation with Next-Generation Sequencing. *Nature Methods*, 6, S13–S20.

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, 95, 315–27.

Morrison, D. A., and Ellis, J. T. (1997). Effects of Nucleotide Sequence Alignment on Phylogeny Estimation: a Case Study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution*, 14, 428–41.

Mullikin, J. C., and Ning, Z. (2003). The Phusion Assembler. *Genome Research*, 13, 81–90.

Nguyen-Dumont, T., Pope, B. J., Hammet, F., Southey, M. C., and Park, D. J. (2013). A High-Plex PCR Approach for Massively Parallel Sequencing. *BioTechniques*, 55, 69–74.

Nichols, R. (2001). Gene Trees and Species Trees Are Not the Same. *Trends in Ecology and Evolution*, 16, 358–64.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007). Recent and Ongoing Selection in the Human Genome. *Nature Reviews Genetics*, 8, 857–68.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP Calling From Next-Generation Sequencing Data. *Nature Reviews Genetics*, 12, 443–51.

Nosenko, T., Schreiber, F., Adamska, M., et al. (2013). Deep Metazoan Phylogeny: When Different Genes Tell Different Stories. *Molecular Phylogenetics and Evolution*, 67, 223–233.

Nylander, J. A. A., RONQUIST, F., Huelsenbeck, J. P., and Nieves-Aldrey, J.-L. (2004). Bayesian Phylogenetic Analysis of Combined Data. *Systematic Biology*, 53, 47–67.

Ogden, T. H., and Rosenberg, M. S. (2006). Multiple Sequence Alignment Accuracy and Phylogenetic Inference. *Systematic Biology*, 55, 314–28.

Page, R. D., and Charleston, M. A. (1997). From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. *Molecular Phylogenetics and Evolution*, 7, 231–40.

Pagel, M., and Meade, A. (2004). A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology*, 53, 571–81.

Parkhill, J. (2002). The Importance of Complete Genome Sequences. *Trends in Microbiology* 10, 219–20–authorreply220.

Penny, D., McComish, B. J., Charleston, M. A., and Hendy, M. D. (2014). Mathematical Elegance with Biochemical Realism: the Covarion Model of Molecular Evolution. *Journal of Molecular Evolution* 53, 711–23.

Perkel, J. (2008). SNP Genotyping: Six Technologies That Keyed a Revolution. *Nature Methods*, 5, 447–453.

Philip, G. K., Creevey, C. J., and McInerney, J. O. (2005). The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal Than for Animal and Fungi and Stronger Support for the Coelomata Than Ecdysozoa. *Molecular Biology and Evolution*, 22, 1175–84.

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

Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution*, 22, 1246–53.

Phillips, M. J. (2004). Genome-Scale Phylogeny and the Detection of Systematic Biases. *Molecular Biology and Evolution*, 21, 1455–58.

Pisani, D. (2004). Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: an Example From the Arthropoda. *Systematic Biology*, 53, 978–89.

Pisani, D., Cotton, J. A., and McInerney, J. O. (2007). Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes. *Molecular Biology and Evolution*, 24, 1752–60.

Pons, J., Barraclough, T., Gómez-Zurita, J., et al. (2006). Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*, 55, 595–609.

Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population Genetic Inference From Genomic Sequence Variation. *Genome Research*, 20, 291–300.

Posada, D., and Buckley, T. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology*, 53, 793–808.

Qiu, Y.-L., Li, L., Wang, B., et al. (2006). The Deepest Divergences in Land Plants Inferred From Phylogenomic Evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 15511–16.

Rannala, B., and Yang, Z. (2003). Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics*, 164, 1645–56.

Rannala, B., and Yang, Z. (2008). Phylogenetic Inference Using Whole Genomes. *Annual Review of Genomics and Human Genetics*, 9, 217–31.

Rhaesa, A. S., Bartolomaeus, T., Lemburg, C., Ehlers, U., and Garey, J. R. (1998). The Position of the Arthropoda in the Phylogenetic System. *Journal of Morphology*, 238, 263–285.

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., et al. (2007). Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. *Systematic Biology*, 56, 389–99.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies. *Nature*, 425, 798–804.

Rosenberg, M. S., ed. (2011). *Sequence Alignment: Methods, Models, Concepts, and Strategies*. Oakland, CA: University of California Press.

Rosenberg, N. A., and Nordborg, M. (2002). Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. *Nature Reviews Genetics*, 3, 380–90.

Roth, A. C., Gonnet, G. H., and Dessimoz, C. (2009). Algorithm of OMA for Large-Scale Orthology Inference. *BMC Bioinformatics*, 10, 220.

Roure, B., Baurain, D., and Philippe, H. (2012). Impact of Missing Data on Phylogenies Inferred From Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 30, 197–214.

Salichos, L., and Rokas, A. (2014). Inferring Ancient Divergences Requires Genes with Strong Phylogenetic Signals. *Nature*, 497, 327–31.

Sankoff, D., Morel, C., and Cedergren, R. J. (1973). Evolution of 5S RNA and the Non-Randomness of Base Replacement. *Nature New Biology*, 245, 232–34.

Scheinfeldt, L. B., and Tishkoff, S. A. (2013). Recent Human Adaptation: Genomic Approaches, Interpretation and Insights. *Nature Reviews Genetics*, 14, 692–702.

Schiffels, S., and Durbin, R. (2014). Inferring Human Population Size and Separation History From Multiple Genome Sequences. *Nature Genetics*, 46, 919–925.

Schneeberger, K., Ossowski, S., Ott, F., et al. (2011). Reference-Guided Assembly of Four Diverse Arabidopsis Thaliana Genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 10249–54.

Scholtz, G. (2002). The Articulata Hypothesis–or What Is a Segment? *Organisms Diversity and Evolution*, 2, 197–215.

Shapiro, B. (2005). Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Molecular Biology and Evolution*, 23, 7–9.

Simpson, J. T., and Durbin, R. (2010). Efficient Construction of an Assembly String Graph Using the FM-Index. *Bioinformatics*, 26, i367–73.

Simpson, J. T., and Durbin, R. (2012). Efficient De Novo Assembly of Large Genomes Using Compressed Data Structures. *Genome Research*, 22, 549–556.

Simpson, J. T., Wong, K., Jackman, S. D., et al. (2009). ABySS: a Parallel Assembler for Short Read Sequence Data. *Genome Research*, 19, 1117–23.

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., and Brumfield, R. T. (2013). Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales. *Systematic Biology*, 63, 83–95.

Sousa, V., and Hey, J. (2013). Understanding the Origin Ofspecies with Genome-Scale Data:Modelling Gene Flow. *Nature Reviews Genetics*, 14, 404–14.

Spang, A., Saw, J. H., Jørgensen, S. L., et al. (2015). Complex Archaea That Bridge the Gap Between Prokaryotes and Eukaryotes. *Nature*, 521, 173–79.

Stamatakis, A. (2014). RAxML Version 8: a Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 30, 1312–13.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology* 57, 758–71.

Steel, M. (2005). Should Phylogenetic Models Be Trying to "Fit an Elephant"? *Trends in Genetics*, 21, 307–9.

Struck, T. H., Paul, C., Hill, N., et al. (2011). Phylogenomic Analyses Unravel Annelid Evolution. *Nature*, 471, 95–98.

Suchard, M. A., and Rambaut, A. (2009). Many-Core Algorithms for Statistical Phylogenetics. *Bioinformatics*, 25, 1370–76.

Swain, M. T., Tsai, I. J., Assefa, S. A., et al. (2012). A Post-Assembly Genome-Improvement Toolkit (PAGIT) to Obtain Annotated Genomes From Contigs. *Nature Protocols*, 7, 1260–84.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic Inference. In *Molecular Systematics*, ed. D M Hillis, Craig Moritz, and Barbara K Mable. Sunderland, MA: Sinauer Associates.

Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The Inference of Gene Trees with Species Trees. *Systematic Biology*, 64, e42–e62.

Taylor, D. J. (2004). An Assessment of Accuracy, Error, and Conflict with Support Values From Genome-Scale Phylogenetic Data. *Molecular Biology and Evolution*, 21, 1534–37.

Telford, M. J., Bourlat, S. J., Economou, A., Papillon, D., and Rota-Stabelli, O. (2008). The Evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 363, 1529–37.

Tewhey, R., Warner, J. B., Nakano, M., et al. (2009). Microdroplet-Based PCR Enrichment for Large-Scale Targeted Sequencing. *Nature Biotechnology*, 27, 1025–31.

The 1000 Genomes Project Consortium (2013). An Integrated Map of Genetic Variation From 1,092 Human Genomes. *Nature*, 490, 56–65.

Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. Edited by Jonathan Badger. *PLoS ONE*, 6, e18093.

Thompson, J. F., and Milos, P. M. (2011). The Properties and Applications of Single-Molecule DNA Sequencing. *Genome Biology*, 12, 217.

Timme, R. E., Bachvaroff, T. R., and Delwiche, C. F. (2012). Broad Phylogenomic Sampling and the Sister Lineage of Land Plants. *PLoS ONE*, 7, e29696.

Treangen, T. J., and Salzberg, S. L. (2011). Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions. *Nature Reviews Genetics*, 13, 36–46.

Trivedi, U. H. (2014). Quality Control of Next-Generation Sequencing Data Without a Reference. *Frontiers in Genetics*, 5, 111.

Turner, E. H., Ng, S. B., Nickerson, D. A., and Shendure, J. (2009). Methods for Genomic Partitioning. *Annual Review of Genomics and Human Genetics*, 10, 263–84.

Vilella, A. J., Severin, J., Ureta-Vidal, A., et al. (2008). EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates. *Genome Research*, 19, 327–35.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47, 97–120.

Watson, M. (2014). Quality Assessment and Control of High-Throughput sequencing Data, *Frontiers in Genetics*, 5, 235.

Westesson, O., Barquist, L., and Holmes, I. (2012). HandAlign: Bayesian Multiple Sequence Alignment, Phylogeny and Ancestral Reconstruction. *Bioinformatics*, 28, 1170–71.

Wheeler, W. C., and Gladstein, D. S. (1994). MALIGN: a Multiple Sequence Alignment Program. *Journal of Heredity*, 85, 417–418.

Whelan, N. V., Kocot, K. M., Moroz, L. L., and Halanych, K. M. (2015). Error, Signal, and the Placement of Ctenophora Sister to All Other Animals. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 5773–78.

Whelan, S. (2008). Spatial and Temporal Heterogeneity in Nucleotide Sequence Evolution. *Molecular Biology and Evolution*, 25, 1683–94.

Whitelaw, C. A., Barbazuk, W. B., Pertea, G., et al. (2003). Enrichment of Gene-Coding Sequences in Maize by Genome Filtration. *Science*, 302, 2118–20.

Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., et al. (2011). Episodic Radiations in the Fly Tree of Life. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 5690–95.

Wilkinson, M. (2006). Identifying Stable Reference Taxa for Phylogenetic Nomenclature. *Zoologica Scripta*, 35, 109–12.

Williams, T. A., Foster, P. G., Cox, C. J., and Embley, T. M. (2014). An Archaeal Origin of Eukaryotes Supportsonly Two Primary Domains of Life. *Nature*, 504, 231–36.

Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J., and Embley, T. M. (2012). A Congruent Phylogenomic Signal Places Eukaryotes Within the Archaea. *Proceedings of the Royal Society B: Biological Sciences*, 279, 4870–79.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment Uncertainty and Genomic Analysis. *Science*, 319, 473–76.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biology*, 15, R46.

Wu, M., and Eisen, J. A. (2008). A Simple, Fast, and Accurate Method of Phylogenomic Inference. *Genome Biology*, 9, R151.

Wu, M., Chatterji, S., and Eisen, J. A. (2012). Accounting for Alignment Uncertainty in Phylogenomics. *PLoS ONE*, 7, e30288.

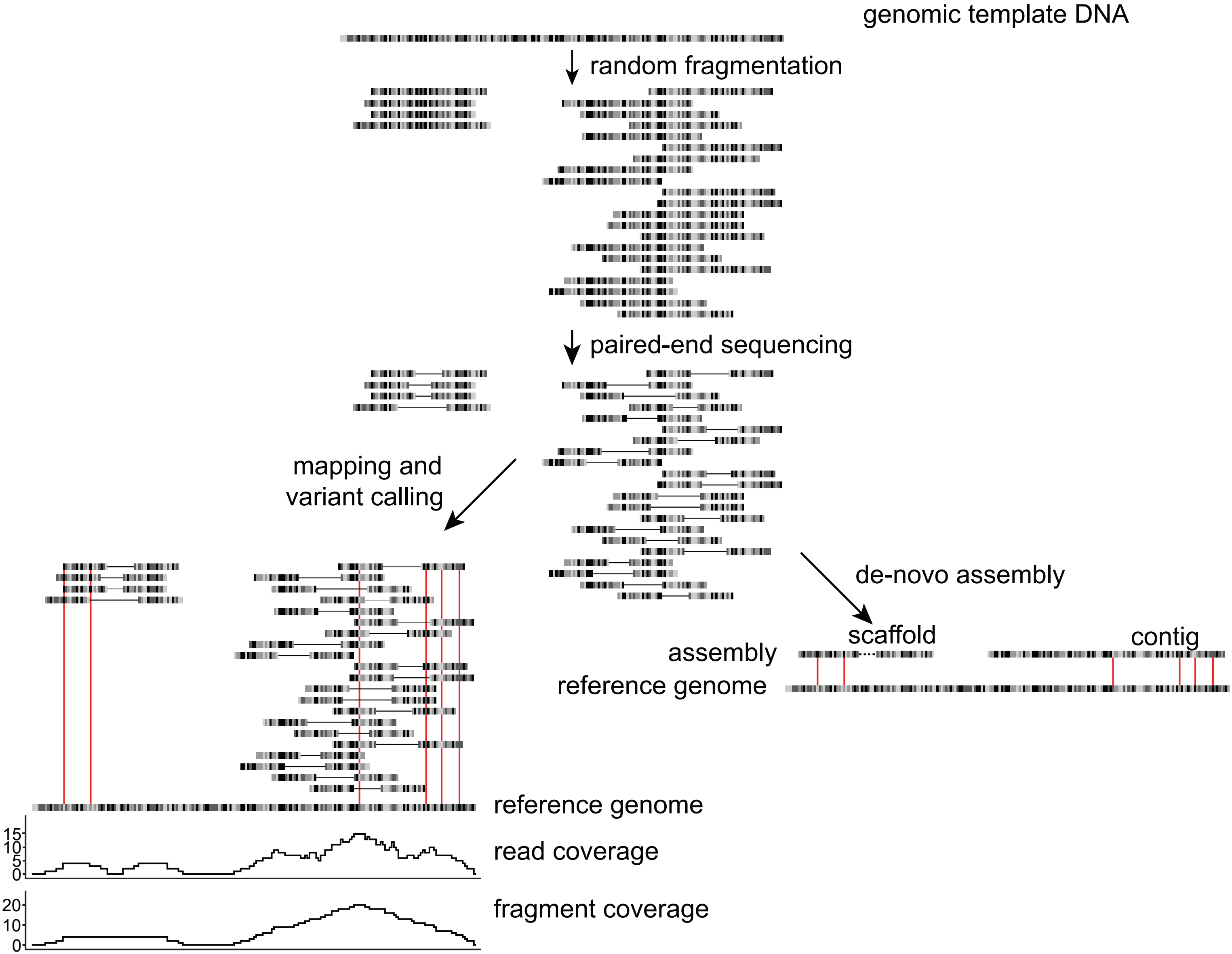
Yalcin, B., Adams, D. J., Flint, J., and Keane, T. M. (2012). Next-Generation Sequencing of Experimental Mouse Strains. *Mammalian Genome*, 23, 490–98.

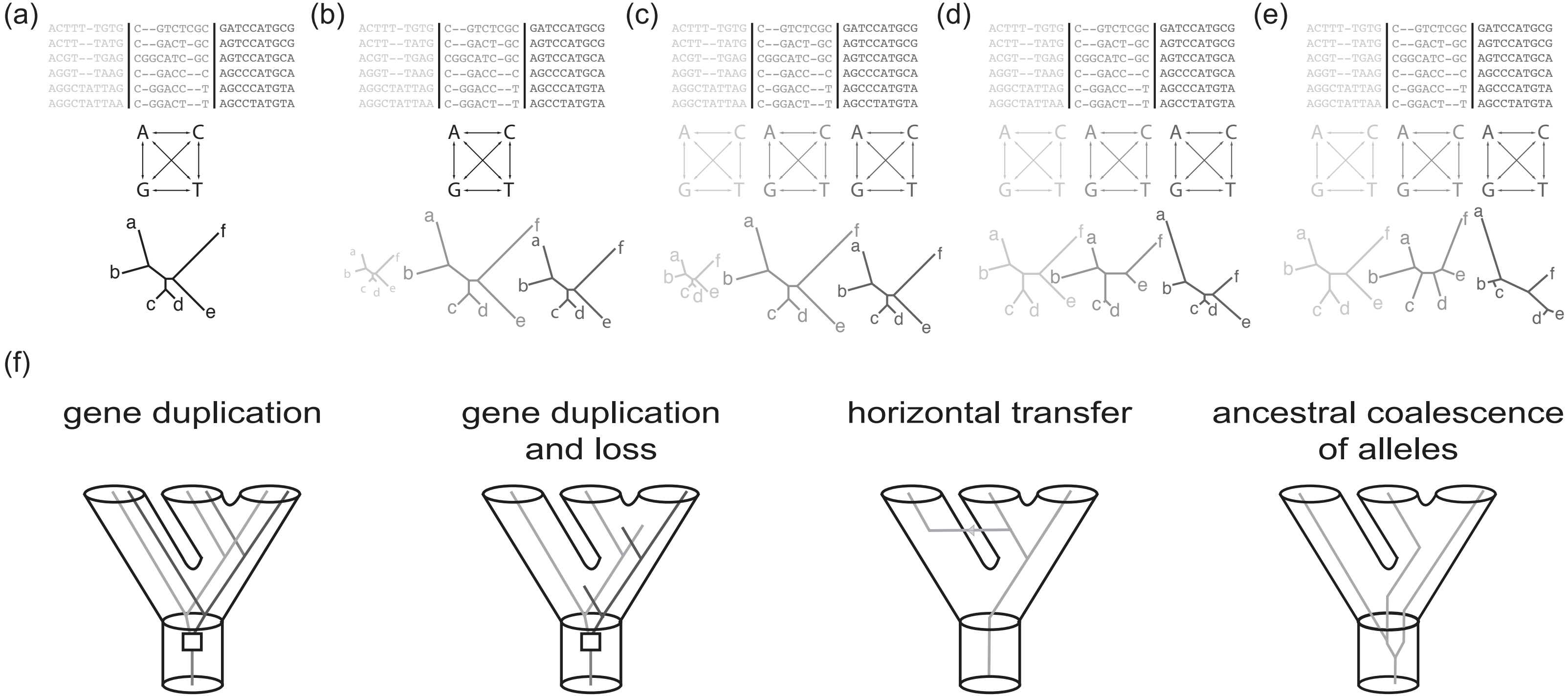
Yang, Z. (1996a). Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *Journal of Molecular Evolution*, 42, 587–96.

Yang, Z. (1996b). Among-Site Rate Variation and Its Impact on Phylogenetic Analyses. *Trends in Ecology and Evolution*, 11, 367–72.

Zerbino, D. R., and Birney, E. (2008). Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. *Genome Research*, 18, 821–29.

Zhou, X., and Rokas, A. (2014). Prevention, Diagnosis and Treatment of High-Throughput Sequencing Data Pathologies. *Molecular Ecology*, 23, 1679–1700.


genomic template DNA
random fragmentation
paired-end sequencing
mapping and
variant calling
de-novo assembly
scaffold
contig
assembly
reference genome
reference genome
15
10
5
0
read coverage
20
10
0
fragment coverage



(a)
ACTTT-TGTG C--GTCTCGC GATCCATGCG
ACTT--TATG C--GACT-GC AGTCCATGCG
ACGT--TGAG CGGCATC-GC AGTCCATGCA
AGGT--TAAG C--GACC--C AGCCCATGCA
AGGCTATTAG C-GGACC--T AGCCCATGTA
AGGCTATTAA C-GGACT--T AGCCTATGTA
A C G T
a b c d e f
(b)
ACTTT-TGTG C--GTCTCGC GATCCATGCG
ACTT--TATG C--GACT-GC AGTCCATGCG
ACGT--TGAG CGGCATC-GC AGTCCATGCA
AGGT--TAAG C--GACC--C AGCCCATGCA
AGGCTATTAG C-GGACC--T AGCCCATGTA
AGGCTATTAA C-GGACT--T AGCCTATGTA
A C G T
a b c d e f
a b c d e f
a b c d e f
(c)
ACTTT-TGTG C--GTCTCGC GATCCATGCG
ACTT--TATG C--GACT-GC AGTCCATGCG
ACGT--TGAG CGGCATC-GC AGTCCATGCA
AGGT--TAAG C--GACC--C AGCCCATGCA
AGGCTATTAG C-GGACC--T AGCCCATGTA
AGGCTATTAA C-GGACT--T AGCCTATGTA
A C G T
A C G T
A C G T
a b c d e f
a b c d e f
a b c d e f
(d)
ACTTT-TGTG C--GTCTCGC GATCCATGCG
ACTT--TATG C--GACT-GC AGTCCATGCG
ACGT--TGAG CGGCATC-GC AGTCCATGCA
AGGT--TAAG C--GACC--C AGCCCATGCA
AGGCTATTAG C-GGACC--T AGCCCATGTA
AGGCTATTAA C-GGACT--T AGCCTATGTA
A C G T
A C G T
A C G T
a b c d e f
a b c d e f
a b c d e f
(e)
ACTTT-TGTG C--GTCTCGC GATCCATGCG
ACTT--TATG C--GACT-GC AGTCCATGCG
ACGT--TGAG CGGCATC-GC AGTCCATGCA
AGGT--TAAG C--GACC--C AGCCCATGCA
AGGCTATTAG C-GGACC--T AGCCCATGTA
AGGCTATTAA C-GGACT--T AGCCTATGTA
A C G T
A C G T
A C G T
a b c d e f
a b c d e f
a b c d e f
(f)
gene duplication
gene duplication and loss
horizontal transfer
ancestral coalescence of alleles