Automated reconstruction of whole genome phylogenies from short sequence reads

Frederic Bertels1,2,*, Olin K. Silander1, Mikhail Pachkov1, Paul B. Rainey2,3, and Erik van Nimwegen1

1Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel 4056, Switzerland
2New Zealand Institute for Advanced Study and Allan Wilson Centre for Molecular Ecology and Evolution, Massey University at Albany, Auckland 0745, New Zealand
3Max Planck Institute for Evolutionary Biology, Plön 24306, Germany

* To whom correspondence should be addressed. Tel: +41 61 267 16 21 Email: frederic.bertels@gmail.com

ABSTRACT

Studies of microbial evolutionary dynamics are being transformed by the availability of affordable high-throughput sequencing technologies, which allow whole-genome sequencing of hundreds of related taxa in a single study. Reconstructing a phylogenetic tree of these taxa is generally a crucial step in any evolutionary analysis. Instead of constructing genome assemblies for all taxa, annotating these assemblies, and aligning orthologous genes, many recent studies (1) directly map raw sequencing reads to a single reference sequence, (2) extract single nucleotide polymorphisms (SNPs), and (3) infer the phylogenetic tree using maximum likelihood methods from the aligned SNP positions. However, here we show that, when using such methods to reconstruct phylogenies from sets of simulated sequences, both the exclusion of non-polymorphic positions and the alignment to a single reference genome, introduce systematic biases and errors in phylogeny reconstruction. To address these problems, we developed a new method that combines alignments from mappings to multiple reference sequences, and show that this successfully removes biases from the reconstructed phylogenies. We implemented this method as a webserver named REALPHY (Reference sequence Alignment based Phylogeny builder), which fully automates phylogenetic reconstruction from raw sequencing reads.

INTRODUCTION

One of the unifying goals across fields as diverse as evolutionary biology, epidemiology, and ecology is to understand the evolutionary relationships between different taxa (Preston et al. 1998; Gill et al. 2005; Ishii et al. 2006; Chun et al. 2009; Ogura et al. 2009; Harris et al. 2010), which are typically quantified by constructing phylogenetic trees (Nei and Kumar 2000). Recently, our ability to resolve such trees has greatly improved due to the rate at which sequence data can be generated via high-throughput sequencing methods. However, using high-throughput sequencing data to precisely determine phylogenetic relationships between taxa is not trivial.

Traditionally, phylogenies are reconstructed from whole genome sequence data by (1) assembling sequence reads into contigs; (2) annotating open reading frames; (3) identifying orthologous open reading frames across all genomes; (4) aligning orthologous coding regions; and (5) reconstructing a phylogenetic tree from these multiple alignments, e.g. (Touchon et al. 2009; Luo et al. 2011;
Rodriguez-R et al. 2012). Subsequently, a phylogenetic tree is then reconstructed from the alignments, typically by applying maximum likelihood methods such as RAxML (Stamatakis 2006) or PhyML (Guindon et al. 2010), or Bayesian methods such as PhyloBayes (Lartillot et al. 2009) or MrBayes (Huelsenbeck and Ronquist 2001).

Although it is generally accepted that this method allows accurate reconstruction of phylogenetic trees (Rosenberg and Kumar 2003), the series of steps involved is not only time consuming, but requires a sophisticated combination of bioinformatics methods.

Recently, an alternative method that is simpler and less time consuming has been applied in several large-scale microbial studies (Harris et al. 2010; Cui et al. 2012; Epstein et al. 2012; Harris et al. 2012; McCann et al. 2013). In this method, raw short sequence reads from each taxon are directly mapped to the genome sequence of a single reference. Homologous sites from all taxa (and in some studies only those sites containing single nucleotide polymorphisms (SNPs)) are then concatenated into a multiple sequence alignment from which the phylogenetic tree is reconstructed.

There are reasons to suspect that such reference-mapping based phylogeny reconstruction methods might introduce systematic errors. First, multiple alignments are traditionally constructed progressively, i.e. starting by aligning the most closely related pairs and iteratively aligning these subalignments, e.g. (Notredame et al. 2000; Chenna et al. 2003). Aligning all sequences instead to a single reference is likely to introduce biases. For example, reads with more SNPs are less likely to successfully and unambiguously align to the reference sequence, as is common in alignments of more distantly related taxa. This mapping asymmetry between strains that are closely and strains that are distantly related to the reference sequence may affect the inferred phylogeny, and this has indeed been observed (Spencer et al. 2007). Second, as maximum likelihood methods explicitly estimate branch lengths, including only alignment columns that contain SNPs and excluding (typically many) columns that are non-polymorphic, may also affect the topology of the inferred phylogeny. This effect has been described before for morphological traits (Lewis 2001) and is one reason long-branch attraction can be alleviated with maximum likelihood methods when non-polymorphic sites are included in the alignment (Felsenstein 1981). Furthermore, the more general issue of selectively leaving out data from multiple sequence alignments has been studied recently and found to affect tree topology (Shavit Grievink et al. 2013).

By simulating sequence evolution across small phylogenies of known topology, we identify parameter regimes where the combination of single-taxon reference mapping and SNP extraction generally leads to severe errors in phylogeny reconstruction. These simulations also show that even when including non-polymorphic sites in an alignment, the effect of mapping to a single reference can lead to systematic errors. In particular, we find that when some taxa are diverged by more than 5-10% from the reference, the distance to the reference is systematically underestimated. This can generate incorrect tree topologies, especially when other branches in the tree are short. Moreover, using data from a set of 21 Escherichia coli genomes, a set of 19 Pseudomonas syringae genomes, and a set of 32 Sinorhizobium meliloti genomes, we show that biases due to mapping to a single reference and exclusion of non-polymorphic sites significantly affect the inferred phylogenetic trees for realistic datasets.
To alleviate these problems, we present a method that combines alignments obtained by mapping reads to not one but to multiple reference sequences. Applying this method to both the simulated and real datasets suggests that by combining sequence mappings to multiple references, mapping biases can be avoided and accurate phylogenies can be reconstructed when each taxon is close (i.e. less than 5% divergence) to at least one of the reference sequences. To make this phylogeny reconstruction procedure available to researchers, including experimental biologists without specific expertise in bioinformatics, we have implemented this method as a webserver called REALPHY (Reference sequence Alignment based Phylogeny builder, available at http://realphy.unibas.ch). REALPHY takes as input raw short sequence read datasets and reconstructs phylogenies by aligning the reads to one or more reference sequences.

RESULTS AND DISCUSSION

The inference of phylogenetic trees from collections of polymorphic sites identified by mapping short sequence reads from multiple genomes to a single reference genome is an increasingly common practice (Harris et al. 2010; Cui et al. 2012; Epstein et al. 2012; Holt et al. 2012; McAdam et al. 2012; Okoro et al. 2012). However, as indicated in the introduction, there are several reasons to suspect that this method may introduce systematic errors.

To test in what situations this method may result in incorrect tree reconstruction, we simulated sequence evolution along known phylogenies, systematically varying both topology (i.e. the placement of the reference genome) and branch lengths. For each dataset we then compared the true tree topology with the tree topologies inferred from (1) the correct and complete alignment of the evolved sequences; (2) the alignment obtained after mapping short reads and retaining only SNP positions; and (3) the alignment after mapping short reads and retaining both SNPs and non-polymorphic sites.

Sequence simulation

Tree shapes and branch lengths

To allow a systematic exploration of parameter space, we restricted our analysis to unrooted four taxon trees, which have only five branches and only three possible topologies: (A,B),(C,D); (A,C),(B,D); and (A,D),(B,C) (Fig. 1A). Throughout, we use taxon A (Fig. 1A) as the reference sequence to which short sequence reads from all other taxa are mapped. To understand the effects of differences in branch length on tree reconstruction, we considered all ways of partitioning the five branches into two subsets (the red and blue branches in Fig. 1C). Because our four taxon tree is asymmetric (the sequence from one taxon is designated as the reference sequence), there are 11 possible groupings of the branches, which we call tree shapes. For each tree shape, we varied the branch lengths in these two groups over a range of values (0.5%, 1%, 2%, 4% and 8% divergence). This gave rise to 25 possible branch length combinations, shown as grid points in Figure 1B. Varying tree shape and branch lengths in this manner gave rise to a total of 275 (25 × 11) different trees.

To refer to any of these trees individually, we specify each of the parameters varied above: first, the tree shape (1-11), followed by the divergence level of the majority set of branches (i.e. the blue
branches in Fig. 1C), and finally the divergence level of the minority set of branches (i.e. the red branches in Fig. 1C). We represent the results for different branch length combinations in a matrix, for example trees with a divergence of 0.5% over blue branches and 8% over red branches correspond to the bottom right corner of Fig. 1B.

**Recombination**
Recombination (gene conversion) occurs frequently in bacterial species (Didelot and Maiden 2010). Thus, in addition to varying tree shape and branch lengths, we investigated the effect of short read mapping on phylogeny reconstruction in the presence of recombination. To simulate this process, 10% of the nucleotides in the reference sequence were replaced with the orthologous nucleotides from the sequence of its cousin taxon (taxon D in Fig. 1; using taxon C would yield identical results). Thus, with the inclusion of recombination, we simulated sequence evolution over a total of 550 trees (2 × 275).

**The impact of SNP extraction and read mapping bias on tree topology**

**Accurate phylogenetic reconstruction when using the true alignment**
We first tested whether the correct tree topology could be reliably recovered from the true alignment, (the evolution of 100,000 nucleotides simulated along a four taxon tree). We found that when there was no recombination, all tree topologies were reconstructed correctly by PhyML (Guindon et al. 2010), a maximum likelihood tree inference program. Not surprisingly, when a sufficient amount of recombination was incorporated, phylogeny reconstruction was no longer error-free (Suppl. Fig. 1).

**Phylogenies reconstructed using only SNP positions are unreliable**
We then tested for parameter regimes that led to incorrect phylogenies when mapping to a single reference, extracting SNP positions only, and reconstructing a phylogeny using maximum likelihood. We identified 131 different parameter settings for which incorrect topologies were inferred for a fraction of the datasets, even in the absence of recombination (Fig. 2). Up to 100% of all inferred tree topologies were incorrect for some parameter sets (e.g., for tree shape 1 at 1% and 4% divergence; Fig. 2). This contrasted strongly with the results using the true alignment, for which no incorrect topologies were inferred for any of the parameter settings.

When recombination was included, the reliability of phylogenetic reconstruction using only SNP positions decreased further. There were 140 parameter sets for which incorrect topologies were inferred and the number of incorrectly inferred trees increased from 6641 to 8871 out of a total of 27,500 datasets.

Importantly, we also found that the choice of the reference taxon affected error rates (Suppl. Fig. 1). For example, although tree shapes three and four have identical branch lengths and differ only in the position of the reference sequence, the accuracy of tree reconstruction differed considerably. When the reference taxon was on a short branch (0.5% divergence), and all other branches were long (8% divergence), no errors were made in inferring the topology. In contrast, when the sister taxon of the reference was on a short branch, and all other branches were long, errors were made in 82% of all cases.
Including non-polymorphic sites improves reliability

The above analyses were performed on alignments containing only SNP positions. When non-polymorphic positions were included in the alignments (i.e. all non-polymorphic positions that were successfully mapped to the reference genome) the accuracy of phylogenetic inference improved. Erroneous topologies were reconstructed for only a single parameter set, in tree shape eight: when the branch of the reference’s sister taxon and the internal branch were short (0.5%), and all other branches were long (8%) the incorrect topology was inferred in 12% of all simulations (Fig. 2). When recombination was included, the accuracy again decreased strongly for five parameter combinations compared to the true alignments (Suppl. Fig. 1).

Thus, when aligning short reads to a single reference genome, there were still some parameter sets for which trees could not be reliably reconstructed. However, for the same parameter sets, no inaccuracies arose when trees were inferred without reference mapping (i.e. using the correct and complete alignment). This demonstrates that the inaccuracy in phylogenetic reconstruction was due to biases that arose in mapping short reads to a single reference sequence.

It is likely that the inaccurately inferred tree topologies are caused primarily by a combination of two factors: (1) Short read aligners such as Bowtie2 can only map sequences closely related to the reference, such that sequences with too many mismatches are discarded; and (2) the relative distance to the reference is important, as regions that are on average more closely related to the reference are less likely to be discarded than regions that are more distant to the reference. Figure 3 qualitatively illustrates how biases are introduced. Assuming that the alignment algorithm only allows a single mismatch between the query and reference sequence within a short region, only a single mutation in the branch leading to the reference would be allowed, and any additional mutations in the region would cause the alignments to be discarded (Fig. 3A). In contrast, three separate mutations would be allowed in the terminal branches that lead to the other leaves in the tree (Fig. 3C). As a consequence, the fraction of columns having identical nucleotides in all taxa would be inflated, whereas the fraction of columns in which all nucleotides would be equal except for the nucleotide in the reference is underestimated (Suppl. Fig. 2). As shown in Supplementary Figure 3, such biases decrease the extent to which the likelihood function supports the correct phylogeny over incorrect alternative topologies, and this is most dramatic for the problematic tree shape 8 (Suppl. Fig. 3B).

Branch lengths are highly inaccurate when using SNP positions only

To analyze how these biases affect branch lengths, we quantified branch lengths from all phylogenetic trees that were correctly reconstructed: (1) from the true alignments; (2) from alignments obtained after short read mapping and SNP extraction (without non-polymorphic sites); and (3) from the full alignments obtained after short read mapping (including all non-polymorphic sites). Since we obviously expected to infer longer overall branch lengths when including SNP positions alone, we assessed the accuracy of the inferred relative branch lengths instead of total branch lengths. We defined the relative length of a branch as its length divided by the sum of all branch lengths within the tree. To quantify the effects of reference mapping and SNP extraction on tree reconstruction, we determined, for each branch in the tree, the ratio of its relative length after mapping and SNP extraction, to the relative length of the branch inferred from the true and complete alignment.
We found again that accuracy was low when using single reference-mapped alignments containing SNP positions alone. The inferred branch lengths in these phylogenies differed considerably from the true branch lengths (Fig. 4A), and we found that even at relatively low levels of divergence (5.9%), in at least 13% of all reconstructions, each of the five branches was estimated to be less than one tenth of their true relative length.

By including both non-polymorphic and SNP positions, the tree reconstruction accuracy increased considerably (Fig. 4B). When total divergence across the true tree was less than 10%, branch length estimation was generally accurate, and branch lengths were only rarely underestimated or overestimated by more than 10%. However, at higher divergence levels accuracy decreased rapidly. This decrease in accuracy also exposed a consistent bias, in which the lengths of the interior branch and the branch leading to the reference were consistently underestimated, while the lengths of the other branches in the tree were consistently overestimated. This confirmed our qualitative considerations above regarding the systematic biases that mapping to a single reference introduces.

Combining alignments from mappings to multiple reference taxa allows for accurate and unbiased phylogeny reconstruction

Although the inclusion of non-polymorphic sites in the alignment considerably improved the accuracy of tree reconstruction, there were parameter regimes where topologies could still not be reconstructed correctly. Furthermore, relative branch lengths were inferred to be up to two times longer/shorter than they ought to be (Fig. 4B). Above we have argued that this bias is caused by the relative position of the reference to the other sequences in the phylogeny. This suggests that this bias may be overcome by using multiple references that are more evenly distributed across the tree. However, as detailed in Methods, care has to be taken that no other systematic biases are introduced when combining alignments from mappings to multiple references. We therefore developed an iterative procedure for merging alignment columns from mappings to different references into a final non-redundant alignment, ensuring that each genomic position from each reference occurs in at most one column of the final alignment, and that conflicts between the mappings using different references are resolved.

To test whether this strategy allowed us to create more accurate phylogenies, we focused on the parameter setting that caused incorrect topologies even when non-polymorphic sites are included (tree shape 8 with 0.5% and 8% divergence) and built four separate alignments by using each of the taxa as a reference. After merging the alignments, the correct phylogeny was reconstructed in 100% of the cases. Furthermore, whereas relative branch lengths differed by up to two fold from the true alignment when using a single reference, when reconstructing the phylogeny from the merged alignment, relative branch lengths differed by at most 18% (Fig. 5). This demonstrated that mapping sequences to multiple reference taxa allows for much more accurate tree reconstruction, even for substantially divergent sequences.

We have now implemented this new method as a webserver, REALPHY (Reference sequence Alignment based Phylogeny builder), in order to make this resource widely available.

Application to bacterial genome sequences
The simulations presented above show under what conditions mapping to a single reference and inferring phylogenetic trees from SNP positions can lead to errors even for simple four taxon trees. Here we show that these errors do typically occur in realistic datasets and that by merging alignments from multiple references, REALPHY avoids such errors. We analyzed three published datasets with sequences from *Escherichia coli* (Touchon et al. 2009), *Pseudomonas syringae* (Baltrus et al. 2011), and *Sinorhizobium meliloti* (Epstein et al. 2012). The first two datasets demonstrate how biases from mapping to a single reference can affect the inferred phylogeny. In addition, they allow us to compare phylogenies constructed by REALPHY with those using classical alignment methods (Touchon et al. 2009; Baltrus et al. 2011) and demonstrate that, as a consequence of the larger number of sites that are included in REALPHY analyses, we obtain more accurate phylogenies. The *Sinorhizobium meliloti* dataset illustrates how the use of only SNP positions can lead to errors in the reconstructed tree.

**Data from E. coli**

Touchon *et al.* determined the phylogeny of 20 fully sequenced *Escherichia coli* and *Shigella* strains as well as one *Escherichia fergusonii* strain (Touchon et al. 2009) using classical methods. They used whole genome data and reciprocal best BLASTP to identify 1,878 genes that were present in all genomes. These orthologs were aligned (covering ~40% of the shortest genome’s length), and a distance matrix based on this alignment was calculated and used to build a neighbor-joining tree.

We applied REALPHY with default parameters to the same dataset, performing 21 separate runs using each of the 21 taxa as a single reference sequence. The topologies of the inferred trees were almost identical. We identified only one branch point at which not all 21 phylogenies agreed. This branch point concerned the sub-clade containing *E. coli* APEC O1, UTI89 and S88 (Fig. 6). Here, in 9 out of the 21 cases, instead of APEC O1 clustering with S88, APEC O1 clustered with UTI89.

We also merged all reference alignments using our merging procedure (1,896,194 bp total length; 170,886 SNP positions, covering 43% of the shortest genome’s length) and inferred a tree for the combined alignment. The tree inferred for this merged alignment was identical to the consensus tree obtained with 12 of the 21 different references.

We found that REALPHY’s tree differed at two branch points (both in clade B2) from the tree calculated by Touchon *et al.* (Fig. 6A and (Touchon et al. 2009)). The first branch point concerned the aforementioned UTI89, APEC O1 and S88 clade. The fact that this branch point was only supported by 12 of the 21 reference alignments suggests the data is indeed not completely unambiguous, and this is substantiated by a bootstrap experiment with 100 repeats showing a support of 88% for this branch, whereas all other branches have 100% support. The second branch point that differed between ours and the Touchon tree concerned the placement of *E. coli* 536, which was placed at the root of the B2 clade in our reconstruction, but clustered with CFT073 and ED1a in the tree reconstructed by Touchon *et al.* Notably, Touchon *et al.* also presented a second tree based on a MAUVE alignment (Touchon et al. 2009). In this case, 536 was placed as outgroup to the B2 subclade, as our consensus and merged-data tree did. Furthermore Touchon *et al.* showed that the support for this branch is only about 92% compared to 100% for the rest of the tree.

The facts that REALPHY uses a larger number of sites (43% of the smallest *E. coli* genome compared to 40% for the Touchon *et al.* data) and that REALPHY’s tree matches the consensus tree
of all reference alignments, suggest that REALPHY's tree may be more accurate. To investigate this further we selected only the seven strains from the B2 clade and reran REALPHY on this dataset. Because of the much higher similarity of this subset of sequences, the reference alignments included a much larger number of sites (covering about 76% of the shortest B2 genome). We found that the tree inferred by REALPHY for the merged alignment was identical to the trees inferred for all seven reference alignments (Suppl. Fig. 4). Moreover, this tree supported all REALPHY's branches from the tree of all 21 taxa, strongly supporting that REALPHY's tree was more accurate than the tree of Touchon et al. Interestingly, the tree built from the B2 clade differs from both REALPHY's and the Touchon et al. tree in the placement of the CFT073 strain, demonstrating that phylogenetic trees can often be further refined by analyzing sequences from subclades separately.

In summary, the analysis of the *E. coli* data showed that the resulting tree can be biased by the reference strain, and that usage of merged alignments from multiple references avoids this bias in this case. It also indicated that REALPHY performs at least as well as classical methods that are more complex and time consuming, and can even outperform these methods when it is using a larger number of sites.

*Data from P. syringae*

In our second analysis we studied a published *P. syringae* dataset (Baltrus et al. 2011), consisting of three fully sequenced genomes and 16 draft genomes in FASTA format. This sequence set was considerably more divergent than the above *E. coli* dataset (~9% compared to ~14%, respectively). As discussed above, we expect the effects of reference mapping bias to increase as sequence divergence increases. Indeed this bias becomes apparent when comparing the reference alignment lengths from *P. syringae* to those of *E. coli*. Whereas for the *E. coli* data approximately 43% of the genomes was covered by the REALPHY alignments, for the *P. syringae* genomes this coverage ranges from 17.6% to 18.9%. As may be expected, this alignment bias is significant enough to affect the inferred *P. syringae* topology. When we used *P. syringae* B728a as the reference sequence, it was placed as most basal taxon in the group II clade instead of Cit7; and when we used *P. phaseolicola* 1448a as reference, *P. phaseolicola* 1448a was placed as most basal taxon in group III instead of *Pae* (Fig. 6B). As both differences concern the clade in which the reference strain is present, it is probable that these differences are the result of a mapping bias to the reference sequence. This bias was removed when we constructed a merged alignment obtained from all three reference genomes. This alignment contained a total of 1,403,205 bp (236,228 SNPs, covering 23% of the smallest reference) and the inferred tree agreed completely with the consensus tree of the three individual reference phylogenies.

Notably, there were some disagreements between the topology of our tree and the multilocus sequence typing tree inferred by Baltrus et al. As the MLST tree is inferred from only a small number of sites, Baltrus et al. inferred another tree based on a concatenated alignment of 324 proteins (corresponding to roughly 6% of the shortest *P. syringae* genome's length) and this phylogeny is more similar to the one inferred by REALPHY. In this case, Pma and Por_1_6 clustered together, as well as Pto_DC3000 and Pla106, agreeing with our inferred topology. As our phylogeny is based on an alignment that contains far more sites than both the protein alignment and the MLST alignment, this
suggests that our phylogeny is likely to model the evolutionary relationships between *P. syringae* strains more accurately than the phylogenies presented by Baltrus et al. This example further confirms that usage of a single reference can significantly bias the resulting topology and that REALPHY's inferred phylogenies are often more accurate than phylogenies constructed from a smaller number of selected sites.

*Data from Sinorhizobium meliloti*

In the previous examples the phylogenies were constructed from all alignment sites, i.e. both SNP sites and non-polymorphic sites. To illustrate reconstruction errors that result from using only SNP sites, we applied REALPHY to a set of *S. meliloti* strains (Epstein et al. 2012). Since this dataset consists of very closely related strains that differ only by a maximum of about 1%, we do not expect to observe significant reference alignment bias (trees inferred from the two references Rm41 and 1021 were identical). However, the usage of SNP sites only may affect the inferred phylogeny. To test this we inferred a phylogeny using PhyML from a complete alignment and an alignment containing only SNP sites (Fig. 7). We found that there is one significant difference between the resulting tree topologies, affecting the placement of T094. In addition, the relative branch lengths of the tree inferred from SNP sites only changed significantly (Suppl. Fig. 5). Interestingly, for the *S. meliloti* dataset, relative branch lengths changed more severely than for the *E. coli* and *P. syringae* datasets, despite the fact that both the *E. coli* and the *P. syringae* dataset are more diverged than the *S. meliloti* dataset. These results further highlight the importance of including non-polymorphic sites in alignments from which phylogenies are inferred using maximum likelihood methods.

**Conclusion**

In recent years, numerous studies (e.g., (Harris et al. 2010; Croucher et al. 2011; Mutreja et al. 2011; Holt et al. 2012; McAdam et al. 2012)) have reconstructed phylogenetic trees for large numbers of closely related bacterial strains by mapping short sequence reads to a reference genome sequence. Here we have analysed the performance of such methods on simulated and real sequence data and have shown that there are two primary pitfalls to this approach. The most readily apparent is that when SNP alignments are used to construct trees with maximum likelihood methods, it can lead to incorrect tree topologies and inaccuracy in the inferred branch lengths. Furthermore, when query sequences are sufficiently divergent from the reference sequence, the most divergent regions of the genome may fail to map, and this mapping bias may lead to incorrect branch lengths and incorrect topologies. Notably, the simulations that we have presented here did not include any variation in mutation rates across the genome; biases in transitions or transversions; or clustering of mutations due to selection; each of these could serve to exacerbate the problem of biased sequence mapping.

To address these pitfalls we have presented a new method, REALPHY, which can successfully avoid biases from mapping to a single reference by implementing a procedure for merging alignments obtained by mapping to multiple reference genomes into a single non-redundant alignment.

REALPHY was mainly designed to reconstruct phylogenies for microbial genomes, i.e. bacterial genomes and single-celled eukaryotes such as fungi, but it can in principle be equally applied to data from higher eukaryotic organisms. However, such applications have not been tested yet and, as
described in Materials and Methods, the computational resources that are required increase with the size of the input genomic data, and may become prohibitive for large eukaryotic genomes that contain many repetitive sequences.

To make this method available to a large community of researchers, including pure biologists without bioinformatics expertise, we provide REALPHY through a webserver, allowing the fast and automated generation of multiple sequence alignments from a variety of genome sequence data formats (e.g., Illumina sequence reads, contigs, draft genomes, fully sequenced genomes), and the automatic reconstruction of phylogenies from these alignments.

MATERIAL AND METHODS

REALPHY implementation

A flowchart of the REALPHY implementation is presented in Figure 8.

Pipeline requirements

The REALPHY pipeline requires the user to provide a set of DNA sequences for each taxon to be included in the phylogenetic tree. This set will typically consist of short sequence reads, but may also include larger sequences, such as fully or partially assembled genomes. In addition, REALPHY requires one or more reference sequence sets to which all sequences will be aligned. Each reference sequence set should consist of a whole genome sequence, a set of chromosome sequences, or a set of contigs.

Alignment

Sequence reads from each query genome provided as FASTQ formatted files are directly mapped to each of the reference sequences using Bowtie2. Assembled genomes provided in FASTA or Genbank format are divided into all possible subsequences of 50 bp (default) to be able to efficiently map these sequences to a reference genome with Bowtie2. REALPHY calls Bowtie2 with the default k-mer length of 22, allowing one mismatch within the k-mers to maximize sensitivity. For each short sequence, only the best mappings are retained, i.e. when there are $n > 1$ ‘best’ mappings; each of the mappings is assigned a weight $1/n$. For each reference sequence, the short read mappings for all query genomes are combined into a multiple alignment containing all orthologous positions that can be reliably identified across the reference and query genomes.

It is possible that paralogous fragments from a query genome may map to the same position as an ortholog in a reference sequence. If these paralogous fragments have diverged, reads from the same query genome may report differing nucleotides aligned to the same position of the reference genome. To avoid such inconsistent mappings, only unambiguous positions are included in the final alignment. Unambiguous position assignment results if the weighted sum of mappings from the query genome is $\geq 10$, and $\geq 95\%$ of the mappings show the same nucleotide. This percentage was chosen to make it unlikely that paralogous mappings would pass the cut-off, but would reduce false negatives due to sequencing errors, which are relatively common in high-throughput sequencing data (Nakamura et al.)
2011). By default, only those alignment columns are retained in which a nucleotide from each of the taxa is present.

In some cases, a small number of genomes may be highly diverged from all reference sequences in some genomic regions, resulting in no successful alignments. In other cases, some genomic regions may be missing entirely. This may be due to their absence in the sequencing data set due to uneven sequence coverage or due to gene deletions. Even if only a few strains are affected by these problems, these regions will be missing from the final alignment, as by default REALPHY only includes regions of the genome for which all strains are present in the alignment. Although such situations are by definition problematic and may lead to inaccurate phylogenies, the user can choose to override the default parameters and include columns in the alignment in which either all or a specified proportion of genomes can have ambiguous or missing nucleotides. These missing nucleotides will be represented by gaps. Importantly, it has been shown that under certain circumstances, phylogenetic trees reconstructed from such alignments can be more reliable than trees reconstructed from alignments in which gapped positions are omitted (Shavit Grievink et al. 2013).

Combining alignments from mapping to different reference sequences

The results of the short read mappings consist of a collection of alignment columns where mappings for all taxa exist. The easiest procedure for combining alignment columns that result from mapping to different references would be to collect the union of all alignment columns and apply a phylogenetic reconstruction method to this dataset. However, such a dataset would be highly redundant, with a given position from a given reference occurring multiple times, i.e. once for each reference to which it was mapped. More importantly, certain positions may be represented more frequently than others in a full collection of alignment columns, which is likely to introduce biases in phylogeny reconstruction. For example, it is well known that substitution rates vary over several orders of magnitude for different genes within a genome (as reviewed in (Rocha 2006)). As a consequence, positions from slowly evolving genes may be reliably mapped to distal reference genomes, whereas positions from fast evolving genes can only be mapped to the closest reference genomes. Consequently, positions from slowly evolving genes are likely to be over-represented in the full collection of alignment columns.

To avoid such biases, REALPHY combines alignment columns from different references into a final set of alignment columns using the following procedure (Suppl. Fig. 6). Alignment columns from all alignments are pooled and then iteratively processed as follows: (1) Randomly select an alignment (column C) from the pool. This column will contain both nucleotides for aligned non-reference genomes (e.g., short sequence read data) as well as nucleotides derived from positions $x_r$ in each of the other reference genomes $r$. (2) For each of these positions $x_r$ occurring in column C, we also select the alignment column $C_r$ of nucleotides mapped to position $x_r$ in the reference $r$ (if this column $C_r$ is present in the pool). (3) All selected columns, i.e. C and the $C_r$ for all other references, are then removed from the pool, and a consensus column is calculated by applying a simple majority rule. (4) This consensus column is then added to the collection of final alignment columns. We continue to select random columns from the pool until there are no columns left. This ensures that each reference genome position occurs in only one of the final alignment columns and that possible disagreements
about which nucleotide from a given taxon should be aligned to a given reference position, are resolved through a simple majority rule.

**Tree building**

Based on the final set of DNA sequence alignment columns, the pipeline determines a phylogenetic tree by applying PhyML ((Guindon et al. 2010), default parameters) or Dnapars (a maximum parsimony method; (Felsenstein 2009)). We chose PhyML as it is optimized for speed in terms of handling large numbers of taxa as well as long sequence alignments. The maximum likelihood method PhyML is run with the general time reversible (GTR) model of nucleotide evolution and gamma distributed rate variation by default. Dnapars from the Phylip program suite is run with its default settings.

**Output**

For each reference genome the output consists of a FASTA and a PHYLIP formatted file that contain an aligned set of orthologous sites (SNPs as well as non-polymorphic sites), a tree file in Newick format, and multiple tab-delimited files (one for each query genome) containing the positions on the reference genome to which the identified SNPs were aligned.

**Computational resources**

The resources REALPHY requires depend mainly on the genome length, the number of genomes, and the number of references. The disk space required (~60MB per Mbase × #genomes × #references) and the computing time (~2 min per Mbase × #genomes × #references) are linearly dependent on these three factors. Furthermore, the amount of RAM required depends primarily on the sequence length and the number of genomes (~250MB per Mbase × #genomes). The computing time required for mapping (which is performed by Bowtie2) will be affected by the repetitiveness of the genomes. As we have not yet tested REALPHY on data from large eukaryotic genomes with many repetitive regions, we currently cannot meaningfully estimate how computational times will scale for such large genomes.

**Implementation**

The pipeline has been fully automated and is provided as a web-server at [http://realphy.unibas.ch](http://realphy.unibas.ch). In addition, a stand-alone implementation in Java can be downloaded from the same website.

**Sequence simulation**

We simulated sequence evolution in a custom made Java program along four taxon trees in which branch lengths were systematically varied between 0.5% and 8% divergence (**Fig. 1**). These sequences were 100,000 bp long, with a GC content of 50%. Evolution occurred with identical transition and transversion rates, i.e. using the elementary Jukes-Cantor model (Jukes and Cantor 1969). For each parameter combination (i.e. the combination of branch lengths in the tree) we repeated the simulation 100 times.

**Phylogenetic analyses**
Multiple sequence alignments were built as described for the REALPHY algorithm. From these alignments as well as the true alignments, phylogenies were reconstructed using the maximum likelihood method PhyML with a general time reversible (GTR) substitution matrix and a gamma-distributed rate heterogeneity model (Guindon et al. 2010).

ACKNOWLEDGEMENT

We thank Chris Field for helpful discussions.

FUNDING

This work was supported by the Royal Society of New Zealand with a James Cook Research Fellowship to PBR; and the Swiss National Science Foundation with an Ambizione Fellowship (grant number PZ00P3-144605) to OKS, and a Swiss National Science Foundation Grant (grant number 31003A_135397) to EvN. Funding for open access charge: Swiss National Science Foundation Grant (grant number 31003A_135397).

REFERENCES


Felsenstein J. 2009. PHYLIP (Phylogeny Inference Package) ver. 3.69.


Figure 1. Tree shapes and branch lengths used to simulate sequence evolution. (A) The three possible topologies in a four taxon tree. (B) The sample space of tree topologies. Each axis indicates the divergence along one set of branches: the divergence of the red branches is indicated along the x-axis and the divergence of the blue branches is indicated along the y-axis. We sampled at five points along each axis, i.e. at 0.5, 1, 2, 4, and 8% divergence, for a total of 25 different combinations of branch lengths. (C) All possible tree shapes considered in the analyses. There are 11 total tree shapes in a four taxon tree that divide the branches into two types (shown here as red and blue). In all our analyses, the reference node is the lower left node of the tree.
Figure 2. Parameter combinations for which incorrect topologies were inferred from mapped alignments excluding (left) and including (right) non-polymorphic sites (without recombination). Left: Fraction of incorrectly reconstructed trees from a total of 100 replicates for all parameter combinations, inferred only on extracted SNPs. Each panel shows data for a different tree shape. Tree shapes are indicated on the left of each panel. Each heatmap shows the divergence (in percent) of red branches across the x-axis and divergences across blue branches on the y-axis. Right: Tree shape 8 with a divergence of 0.5% along the short branches and 8% along the long branches is shown. The percentage above the tree indicates the proportion of trees (out of 100) for which the incorrect topology was inferred.

Figure 3. Mapping to a single reference introduces alignment biases. Assuming, for illustrative purposes, that the alignment algorithm allows only one mismatch between query and reference within a 21 bp region, each panel shows the maximal number of mutations allowed in order for successful mapping of all orthologous fragments to occur, as a function of the positions in the tree where mutations occur. (A) If a single mutation occurs on the reference branch, then the distance from the reference to all other sequences reaches one immediately and no further mutations are allowed. (B) One mutation on the internal branch as well as one mutation on the sister branch are allowed before all three query sequences reach a distance of one to the reference. (C) Three independent mutations...
on each of the external branches are allowed before all query sequences reach a distance of one to the reference.

**Figure 4.** Deviation of relative branch lengths, as inferred from mapped sequence alignments, from the true relative branch lengths for (A) phylogenies inferred using SNP positions only and (B) phylogenies inferred using all positions. For each branch in our simulated four taxon trees, the figure shows the proportion of trees in which the estimated relative branch length deviated from the true relative branch length to a certain degree (colour). The trees were subdivided into six equally sized bins based on the overall divergence level (proportion of columns within the original multiple sequence alignment that contain SNPs) and the branch length ratios were calculated for each divergence class (position on the x-axis). The proportion of trees inferred from mapped sequence alignments that contain relative branch lengths that are more than 10 times greater than those from the true tree are shown in dark blue. Relative branch lengths that are more than 10 times shorter are shown in dark red. Relative branch lengths that are within 10% of the true branch length are shown in white (see legend). The figure shows one plot for each of the five branches within the tree (this branch is indicated in green in the four taxon trees between A and B). The reference sequence is always the taxon on the bottom left of the tree. Trees were only included in the statistics if the mapped tree topology matched the true (known) tree topology.

**Figure 5.** Accuracy of estimated relative branch lengths when inferring a phylogeny from a single reference alignment (grey bars) and from a merged alignment of all four references (white bars). The relative branch length (BL) of a particular branch is defined as the length of the branch divided by the
sum of all branch lengths in the tree. The BL ratio is the ratio of the estimated BL and the BL of the true tree. The bars show the BL ratios for each of the five branches (indicated at the bottom) of the trees inferred in 88 independent trials (all correctly reconstructed topologies) of alignments from tree shape eight with divergences of 0.5% and 8%. Note that the closer the bars are to one, the more similar the estimated tree is to the true tree.

Figure 6. Comparison of REALPHY phylogenies to phylogenies inferred in previous publications. Both REALPHY trees (green) were built using PhyML, with the general time reversible model with gamma distributed rates as substitution model. The annotation on the branch points in black denotes the bootstrap support for the branch points from a total of 100 bootstrap experiments (only shown if <100) for REALPHY trees, Bayesian probabilities for the Baltrus tree (shown if <0.95) and bootstrap values out of 1000 for the Touchon tree (shown if <1000). Annotations in grey show the number of REALPHY single-reference trees that support the particular branch points (only shown if <21 for *E. coli* and <3 for *P. syringae*). Boxed parts of the trees contain differences to the previously published corresponding tree. (A) *E. coli* phylogeny reconstructed by Touchon et al. (left) compared to a phylogeny reconstructed from all 21 merged reference alignments produced by REALPHY. The differences between the two trees are the placements of *E. coli* 536 and S88. (B) *P. syringae* phylogeny reconstructed by Baltrus et al. (left) compared to a phylogeny based on mappings to the three fully sequenced *P. syringae* strains: *P. syringae* B728a, *P. syringae pv. phaseolicola* 1448a and *P. syringae pv. tomato* DC3000. Right: The root of the tree was arbitrarily selected to facilitate comparison between the two topologies. When inferring trees from single reference genome alignments, two branch points are not supported by all three trees (annotated on the corresponding branches). These branch points concern the placement of Cit7 (*P. syringae* B728a as reference) and Pae (*P. syringae pv. phaseolicola* 1448a as reference).
Figure 7. Comparison between two phylogenies inferred from a REALPHY alignment of Sinorhizobium meliloti strains (Epstein et al. 2012) including (left) and excluding (right) non-polymorphic alignment sites. The alignments were created by merging the reference alignments from S. meliloti Rm41 and 1021. The red box highlights differing branch points. Bootstrap support is indicated if below 100%, except for the blue clade where the support is low.

Figure 8. Illustration of the individual steps in the REALPHY pipeline (running from top to bottom). All fully sequenced or assembled genomes (FASTA and Genbank files) are divided into all overlapping 50 bp subsequences. Short sequences are aligned to individual reference sequences with Bowtie2. Alignment columns are created from all pairwise mappings to the references. Individual reference
alignments are merged into a single multiple sequence alignment. A phylogeny is reconstructed from merged and individual reference alignments via PhyML.