Genome Biology

CORRESPONDENCE

Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses

Albert J Vilella, Ewan Birney, Paul Flicek and Javier Herrero*

Comment on Milinkovitch *et al.*: http://genomebiology.com/2010/11/2/R16

A response to 2x genomes - depth does matter by MC Milinkovitch, R Helaers, E Depiereux, AC Tzika and T Gabaldón. *Genome Biol* 2010, **11**:R16.

Low-coverage genomes and phylogeny

The Mammalian Genome project [1] sequenced several placental mammalian species at a low coverage (about 2x). These datasets are characterized by a large number of assembly gaps and a larger fraction of sequencing errors than high-coverage genome sequences. As a result, many gene models will miss entire exons, and several codons will be miscalled. These sequences present new challenges for phylogenetic studies.

In their article, Milinkovitch *et al.* [2] study the effect of low-coverage genomes in standard phylogenetic reconstructions. They show how the addition of these genomes results in extra duplication and gene loss events. They base their conclusion on their earlier study of the human Phylome database [3]. In essence, every human gene is aligned to its homologs in other eukaryotes. They build a phylogenetic tree for each multiple sequence alignment with PhyML [4] using several protein evolutionary models and select the best tree. They use a strict gene versus species tree reconciliation method [5] to call duplication events. Milinkovitch *et al.* [2] also show a figure with a substantial accumulation of duplication calls in the EnsemblCompara GeneTrees database [6] with respect to the Phylome database [3] (Figure 4 in Milinkovitch *et al.* [2]). This figure uses data from an early version (version 41; October 2006) of the EnsemblCompara GeneTrees database [6], namely our initial attempt to include low-coverage genomes in the phylogenetic trees. We realized following this initial attempt that a strict gene versus species tree reconciliation method did not work well with the new genomes and implemented a new method, which has itself been since refined, in Ensembl version 42 (December 2006) [7].

Dubious duplications

We have previously described how relying on a strict reconciliation method after using PhyML leads to an excess of duplication event calls [6]. In that same article, we showed how TreeBeST [8] produces more biologically consistent results. We classify duplication events as either supported or dubious duplications [6]. Dubious duplication events are those without two genes from the same species in two child sub-trees. In other words, a dubious duplication has no duplication in any extant species to support an ancient duplication event. Instead, the gene tree and the species tree disagree. TreeBeST, using the species tree as input, penalizes both duplication and gene loss events when reconstructing the gene tree. As such, TreeBeST reduces drastically the number of dubious duplication events (Figure 1) [6]. Starting from release 42 of the Ensembl, this method has been used for phylogenetic inferences.

Milinkovitch *et al.* [2] find a large number of dubious duplications at the root of the eutherian (placental mammalian) tree, and claim that most of them are due to the addition of low-coverage genomes. The addition of new placental mammal genomes in the gene trees increases the probability of finding a gene misplaced in the tree. Necessarily, the lower sequence coverage and higher fraction of errors in the low-coverage genomes will only increase this probability. We argue that using TreeBeST can overcome this problem to a large extent. Remaining dubious duplication nodes, especially when the number of sampled taxa is large, can be safely considered as speciation nodes. In line with this argument, we use the concept of 'apparent orthologs' to describe homologous genes related by a dubious duplication node in Ensembl [9].

We repeated the analysis described in Milinkovitch *et al.* [2] using successive Ensembl releases and looking at duplication events at the root of the eutherian tree. Releases 39 to 41 of the database were built using PhyML

*Correspondence: jherrero@ebi.ac.uk
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

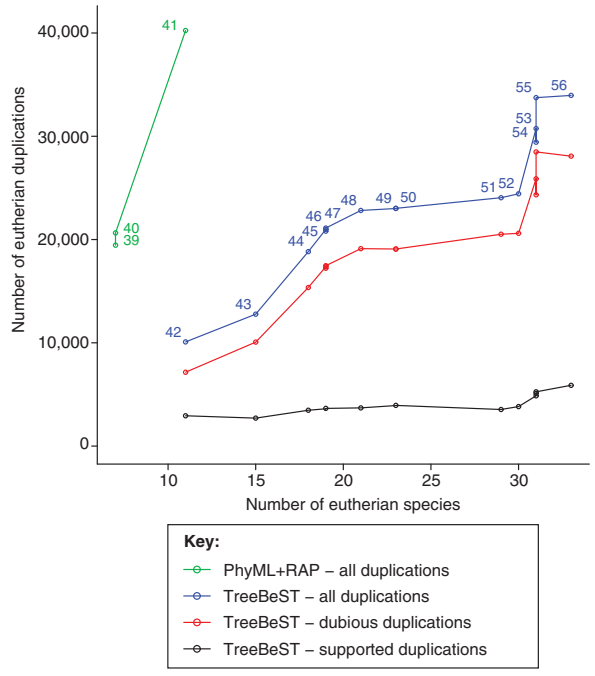

Figure 1. Duplications at the root of the eutherian tree. The number of duplications found at the root of the eutherian tree are shown for different Ensembl releases (colored numbers) with respect to the number of eutherian species included in the analysis. Other duplications found in sub-clades (for instance, primate or rodent duplications) are excluded. In the first releases with gene trees (39-41), PhyML+RAP were used. The first low-coverage genomes were added in release 41. Starting from release 42, TreeBeST was used to infer the gene trees and dubious duplications were called when the duplication nodes were not supported by any extant duplication.

for tree estimation and RAP [10] for gene and species tree reconciliation. In release 41, with the addition of the first four low-coverage genomes, we observe a large increase in the number of duplications (Figure 1, green). Starting from release 42, TreeBeST has been used by Ensembl to infer the gene trees. The remaining dubious duplications that TreeBeST does not resolve are indicated in red in Figure 1. We find that the number of supported duplications stays approximately constant even after tripling the number of species (Figure 1, black). We also observe that despite the increase in the number of dubious duplications, the total number of observed duplications when classifying 33 placental mammals with TreeBeST is still lower than when classifying 11 of them with PhyML.

Low-coverage genomes in EnsemblCompara GeneTrees

Ensembl release 51 saw the addition of six new low-coverage genomes (kangaroo rat, rock hyrax, megabat, bottle-nosed dolphin, tarsier and alpaca) and the upgrade of the guinea pig from low (about 2x) to high coverage (6.7x). Despite this large increase in the number of low-coverage genomes, the total number of duplications at the root of the eutherian mammals increases by only 4.5%. Moreover, the total number of supported duplications is reduced by only 10%. This suggests that the addition of low-coverage genomes in the Ensembl GeneTrees does not necessarily correlate with an increase in the number of duplications. Other important factors, such as taxon sampling and the short length of the internal branches at the root of the eutherian tree, can have a stronger effect in resolving gene tree topologies.

In conclusion, we argue that low-coverage genomes are useful for inferring phylogenetic trees, but only if the phylogenetic methods account for the difficulties of analyzing these data, especially for challenging clades with short speciation branches, such as the mammalian clade. TreeBeST can successfully resolve the majority of the problematic cases and we can detect and handle most of the remaining ones when post-processing the trees.

As mentioned in Milinkovitch *et al.* [2], 2x genomes cause additional problems, such as lack of coverage across an entire gene. For instance, the human and mouse genomes contain 22,379 and 23,117 predicted protein-coding genes, respectively, whereas the tree shrew contains only 15,458. Gaps in the assembly and not evolutionary processes are responsible for most of the missing genes. Wherever possible, methods are provided in Ensembl to maximize the use of information. TreeBeST helps to overcome problems arising from the low quality of the sequence. The methods are documented both in publications and on the website, but users should remain cautious of surprising results, particularly with more challenging datasets such as 2x genomes.

Michel C Milinkovitch, Raphaël Helaers, Eric Depiereux, Athanasia C Tzika and Toni Gabaldon respond:

Low-coverage genomes and phylogeny

Vilella *et al.* correctly indicate that low-coverage mammalian genomes are characterized by a large number of assembly gaps and sequencing errors. In our original study [2], we quantified the impact of low-coverage genomes on inferences pertaining to gene gains and losses when analyzing eukaryote genome evolution through gene duplication. We demonstrated that low-coverage genomes generate a massive number of false gene losses, but also striking artifacts in gene duplication inference at the most recent common ancestor of low-coverage genomes. Vilella *et al.* assert that our conclusions are only based on

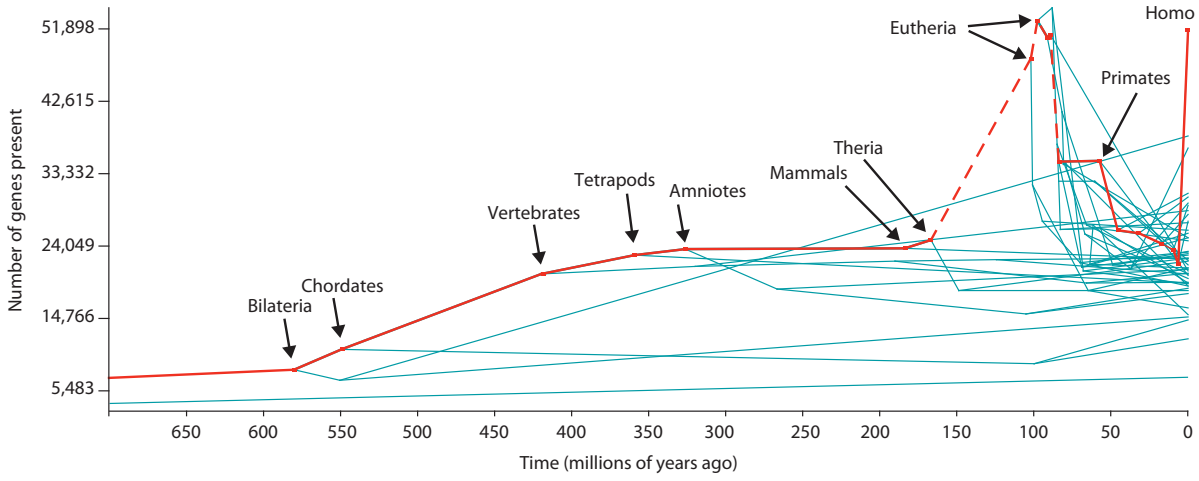

Figure 2. Increase in gene number through evolutionary time for all lineages (blue lines) in the latest Ensembl version (version 59). The red line indicates the lineage leading to the human genome. The dashed part of the line indicates the artifactual peak of duplications at the eutherian nodes.

the 'Phylome' database (PhylomeDB) and on an early version (version 41) of their EnsemblCompara GeneTrees database. We disagree with these assertions and argue that all our conclusions remain valid even when considering the later versions of Ensembl to which Vilella *et al.* contributed. For example, Figure 2 of this Correspondence indicates that the striking artifactual peak of false duplications at the eutherian nodes is still present when plotting the increase in gene number through evolutionary time using a recent version of Ensembl (version 59), exactly as shown in Figure 2 of our study [2] (which used Ensembl version 49). This shows that the use of TreeBeST in the most recent versions of Ensembl cannot resolve the problem generated by incorrect gene topologies. In our study [2], PhylomeDB was used as a reference because it does not contain any low-coverage genomes; the differences in gene tree inference methods between the PhylomeDB and Ensembl pipelines are irrelevant here.

Dubious duplications

We illustrated in our study (Figure 7a in Milinkovitch *et al.* [2]) how incorrect gene trees can generate false duplication events and false losses. One of us (TG) has previously suggested [3] that the species-overlap score (the fraction of shared species over the total of species in post-duplication nodes) provides a means for assessing the validity of inferred duplication nodes. Vilella *et al.* ([6] and here) have used basically the same approach (which they call the 'duplication consistency score') in EnsemblCompara. We indicated (Figure 7b in Milinkovitch *et al.* [2]) that the eutherian node shows one of the three worst duplication confidence values of all nodes in the species phylogeny. Figure 1 of this Correspondence in fact indicates that TreeBeST confirms our results: a majority of eutherian duplications are dubious and therefore a majority of the gene trees are wrong. Moreover, even though duplications with high species-overlap scores are more likely to be correct than those with low scores, it would be inappropriate to set an arbitrary threshold for this score below which all duplications are assumed to be dubious. This is because any threshold will inevitably lead to some true duplications being given low scores because of real differential losses after duplication (these will thus be false negatives). This will be especially true for the case of phylogenies including low-coverage genomes because missed genes and wrong topologies can produce wrong (low) duplication scores. For all these reasons, the use of low-coverage genomes results in an increased difficulty in reconstructing ancestral events, regardless of the use of any duplication-calling method.

Vilella *et al.* also claim that the 'addition of new placental mammal genomes in the gene trees increases the probability of finding a gene misplaced in the tree.' We had anticipated that possibility in our study (page 6, second paragraph in Milinkovitch *et al.* [2]), and we therefore performed large-scale simulations demonstrating that low coverage *per se* of genome sequences suffices to generate artifactual gains at a predictable node in the species phylogeny. Starting with exclusively high-coverage genomes (whose analysis does not generate a peak of false duplications at the eutherian node), we randomly introduced protein sequence ambiguities in three eutherian species according to a distribution approximating that observed in real low-coverage genomes. Re-analysis of this perturbed dataset generates a massive number of

artifactual duplications, with the most affected node being the basal eutherian lineage, that is, the common ancestor of the three perturbed species (Figure 8 in Milinkovitch *et al*. [2]). No species has been added in these simulations; thus, and contrary to Vilella *et al.*'s claim, the artifactual gains uncovered with our simulations cannot have been caused by increased taxon sampling. Incidentally, it is known that increased taxon sampling tends to improve rather than decrease the accuracy of phylogenies [11-13]. We therefore maintain that a significant proportion of artifactual gains at the eutherian node (and elsewhere in the species tree) might be due to genome sequence low coverage *per se*.

Low-coverage genomes in EnsemblCompara

Vilella *et al.* indicate that the incorporation (in Ensembl version 51) of six additional low-coverage eutherian genomes and the promotion of a single low-coverage eutherian genome does not increase dramatically the number of dubious duplication nodes. That is correct and logical. Indeed, our simulations [2] indicated that transformation of only three high-coverage eutherian species to low coverage suffices to generate a striking artifactual peak at the eutherian node. Thus, as we indicated in our study [2], 'major artifacts in gene gains and losses [...] will remain until all low-coverage genomes are promoted to high coverage.' Upgrading only one eutherian genome sequence cannot solve the problem.

Conclusions

The aim of our study [2] was certainly not (contrary to what is implied by Vilella *et al.*) to suggest that the Ensembl project performs poor analyses. Most of the analyses presented were performed with MANTiS [14], an application system that implements a dynamical programming approach for the mapping of gene gains, duplications and losses on the metazoan phylogenetic tree. MANTiS [14] builds a relational database integrating, in an explicit phylogenetic framework, all Ensembl genes, corresponding molecular functions and biological processes, and expression data. As such, MANTiS makes extensive use of the Ensembl database in general, and Vilella *et al.*'s EnsemblCompara database [6] in particular.

We feel that the Ensembl project incorporates among the best analytical tools for phylogeny inference and identification of valid versus ambiguous duplication nodes in gene trees. However, our analyses show that low coverage genome sequence *per se* is likely to generate a large number of artifactual duplications. It is not clear to us why Vilella *et al.* attempt to minimize the importance of good raw data for valid gene tree reconstruction and proper inferences of gene gains, duplications and losses (which all depend on these gene trees). Figure 1 actually indicates that the majority of inferred duplication nodes are wrong. This necessarily means that the majority of the gene trees are wrong. We tend to think that most end-users in the genomic community wish to see the correct gene trees and corresponding valid gene duplication and gene loss events. Our study [2] can also be viewed as a plea for better homogeneity in both taxon sampling and high-coverage sequencing (we urged the promotion of low-coverage genomes to high coverage and not their removal), which will be made easier by the development of next generation sequencing. It is unclear why Vilella *et al.* have interpreted our work [2] as a criticism of genome sequencing and genome database projects. On the contrary, we view it [2] as a strong support for additional sequencing programs as well as for extensive analytical efforts, such as those implemented in the remarkable Ensembl project [6,7,15].

Published: 21 February 2011

References

1. Mammalian Genome Project [http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/mammalian-genome-project]
2. Milinkovitch M, Helaers R, Depiereux E, Tzika A, Gabaldon T: **2X genomes - depth does matter.** *Genome Biol* 2010, **11:**R16.
3. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T: **The human phylome.** *Genome Biol* 2007, **8:**R109.
4. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
5. Zmasek CM, Eddy SR: **A simple algorithm to infer gene duplication and speciation events on a gene tree.** *Bioinformatics* 2001, **17:**821-828.
6. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara Gene-Trees: complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19:**327-335.
7. Ensembl Archives [http://www.ensembl.org/info/website/archives/index.html]
8. TreeSoft: TreeBeST [http://treesoft.sourceforge.net/treebest.shtml]
9. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36:**D707-D714.
10. Dufayard J, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G: **Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases.** *Bioinformatics* 2005, **21:**2596-2603.
11. Wheeler WC: **Extinction, sampling, and molecular phylogenetics.** In *Extinction and Phylogeny.* Edited by Novacek MD, Wheeler QD. New York: Columbia University Press; 1992:205-215.
12. Gatesy J, Milinkovitch M, Waddell V, Stanhope M: **Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa.** *Syst Biol* 1999, **48:**6-20.
13. Zwickl DJ, Hillis DM: **Increased taxon sampling greatly reduces phylogenetic error.** *Syst Biol* 2002, **51:**588-598.
14. Tzika A, Helaers R, Van de Peer Y, Milinkovitch MC: **MANTiS: a phylogenetic framework for multi-species genome comparisons.** *Bioinformatics* 2008, **24:**151-157.
15. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, *et al.*: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39:**D800-D806.

doi:10.1186/gb-2011-12-2-401
Cite this article as: Vilella AJ, *et al.*: Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biology* 2011, **12:**401.