

SIMULTANEOUS HISTORY RECONSTRUCTION FOR COMPLEX GENE CLUSTERS IN MULTIPLE SPECIES*

YU ZHANG

*Department of Statistics,
326 Thomas Building, Penn State University,
University Park, PA 16802, USA
E-mail: yuzhang@stat.psu.edu*

GILTAE SONG, CHIH-HAO HSU, WEBB MILLER

*Center for Comparative Genomics and Bioinformatics,
506B Wartik Lab, Penn State University,
University Park, PA 16802, USA
E-mail: {gsong, chih-hao, webb}@bx.psu.edu*

Genomic intervals that contain a cluster of similar genes are of extreme biological interest, but difficult to sequence and analyze. One goal for interspecies comparisons of such intervals is to reconstruct a parsimonious series of duplications, deletions, and speciation events (a putative evolutionary history) that could have created the contemporary clusters from their last common ancestor. We describe a new method for reconstructing such an evolutionary scenario for a given set of intervals from present-day genomes, based on the statistical technique of Sequential Importance Sampling. An implementation of the method is evaluated using (1) artificial datasets generated by simulating the operations of duplication, deletion, and speciation starting with featureless “ancestral” sequences, and (2) by comparing the inferred evolutionary history of the amino-acid sequences for the CYP2 gene family from human chromosome 19, chimpanzee, orangutan, rhesus macaque, and dog, as computed by a standard phylogenetic-tree reconstruction method.

1. Introduction

Repeated duplications within a cluster of similar genes provide a mechanism for rapid evolution, making those clusters particularly interesting. However, reconstructing evolutionary scenarios that include the operations of duplication and deletion has proven difficult; a number of partial solutions have been proposed^{1,2,3,4}, but none provide an explicit evolutionary

*This work is supported by NHGRI grant HG002238.

reconstruction for multiple species under reasonably general conditions. A recent report⁵ attempts to reconstruct an even more general set of operations that includes fissions/fusions of chromosomes and translocations. Under certain idealized conditions, it guarantees a most-parsimonious reconstruction, and efforts are on-going to build a practical reconstruction pipeline based on that theory.

We recently proposed an algorithm for reconstructing a hypothetical ancestral sequence and a parsimonious set of evolutionary events to explain an observed gene cluster in a given species⁶. Here we generalize our single-species algorithm to simultaneously reconstruct an evolutionary history of orthologous gene clusters in multiple species, i.e., an ordered series of duplication, deletion, and speciation events that reproduces the current orthologous gene-cluster configurations from an ancestral progenitor. We start by setting a lower bound for the percentage identity for both within-species and between-species pairwise alignments. This defines a time bound (such as 25 millions of years ago) after which the recent histories of multi-species gene clusters can be reconstructed. Our belief is that attacking this limited but critical problem of ancestral reconstruction and by employing a fundamentally different approach based on statistical methods, we may be able to compute reconstructions more economically and under more general assumptions than when using purely combinatorial methods⁵.

2. Methods

Our approach takes multi-species DNA sequences of orthologous gene clusters as input. The sequences are first processed by the following pipeline: 1) use blastz⁷ to construct all combinations of self-alignment and pairwise-alignment dot-plots; 2) filter out weak alignments with percentage identity less than a threshold, say 70%, that corresponds to an evolutionary separation greater than that between the most distant species pair under consideration; 3) process the dot-plots such that all local alignments satisfy the “transitive closure property”, i.e., given local alignments between regions (A,B) and (B,C) , there must be a local alignment between (A,C) , where the regions may be located in different species; the transitive closure property ensures the completeness of information; 4) chain together local alignments of similar percentage identity broken by small insertions/deletions or post-duplication insertion of interspersed repeats. For instance, Fig. 1 shows the resulting self-alignment dot-plots for five species in the CYP2 cluster.

For such preprocessed data, we propose a new algorithm to reconstruct

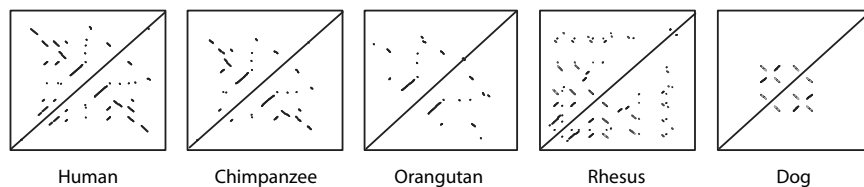


Figure 1. Self-alignment dot-plots for orthologous CYP2 clusters of cytochrome P450 genes in human, chimpanzee, orangutan, rhesus, and dog.

the detailed evolutionary history of orthologous gene clusters in multiple genomes. We consider tandem and non-tandem duplications, duplications with inversions, deletions, and speciation events. We first describe our previously developed algorithms for single-species reconstruction, then generalize the method to multiple species. For multi-species reconstruction, a speciation event corresponds to coalescence between the two sibling species backward in time. We demonstrate how to ensure coalescence in multi-species history reconstruction.

2.1. A Basic Algorithm for Single-Species Reconstruction

Given a preprocessed gene cluster in a single species, there is a simple combinatorial algorithm⁶ that correctly reconstructs all duplication events. The underlying assumption is that the gene clusters have been exclusively generated by segmental duplications, and the alignment boundaries produced by past duplication events are not reused. The no-alignment-boundary-reuse assumption is a stronger version of the commonly used no-breakpoint-reuse assumption in genome rearrangement analysis, which is first established in Nadeau and Taylor's landmark paper⁸. Although this assumption is still in debate⁹, from an inference point of view, it is needed for resolving ambiguities in duplication reconstruction.

The basic algorithm reconstructs duplication events backward in time. In each step, the algorithm finds a local alignment corresponding to a duplication event that satisfies the following criterion: when a duplication is "unwound" by removing one matched DNA segment from the sequence, the breakpoints of the corresponding event do not coincide with remaining alignment boundaries. An example is shown in Fig. 2. If several alignments satisfy the above criterion, we randomly choose one and unwind it. We further merge and extend remaining alignments after each step of reconstruction. The above procedure is repeated until all local alignments in the dot-plot are resolved. The output is thus a reconstructed history

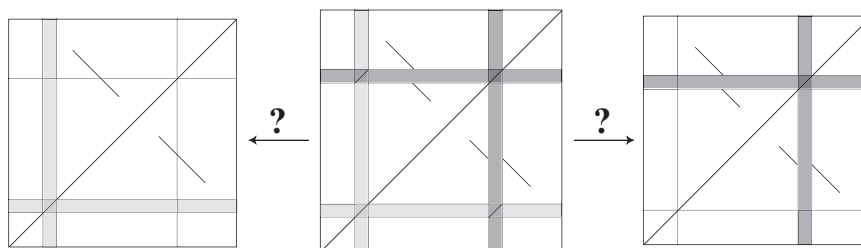


Figure 2. A gene cluster configuration is shown in the middle dot-plot. Lightly and darkly shaded segments are matched, corresponding to a duplication event. To reconstruct the duplication event, we either remove the darkly shaded segment to obtain the left dot-plot, or we remove the lightly shaded segment to obtain the right dot-plot. According to the basic algorithm, the left dot-plot is a correct reconstruction, as breakpoints of the reconstructed event do not coincide with existing alignment boundaries. The right dot-plot is incorrect because there are two breakpoints coincide with existing alignment boundaries.

consisting of a sequence of duplication events, which if unwound reduces the gene cluster sequence to a duplication-free progenitor.

Despite its simplicity, we have proven that the basic algorithm always has solutions, and all solutions are optimal⁶. In particular, there is at least one local alignment at each step of the reconstruction that satisfies the criterion. Depending on the alignments chosen in each step, different histories can be reconstructed that generate the current gene cluster sequence from some duplication-free progenitors. One of the reconstructions will correspond to the real history, while all reconstructions will contain exactly the same number of events, the same duplication sizes, yet different ancestral sequence configurations.

2.2. An Alternative Probabilistic Solution

The exclusive duplication assumption and the no-alignment-boundary reuse assumption are often violated in real analysis. Genomic deletions are likely to occur, and alignment boundaries may be reused due to either misalignments or rearrangement hotspots. When an alignment boundary has to be reused during reconstruction, the algorithm can no longer guarantee a correct history reconstruction or even estimate the true number of events. In addition, there are often many equivalent history reconstructions for a gene cluster, which differ by the event orders and thus produce very different progenitor sequences. To accommodate deletion events and data complications, and to infer gene cluster evolution from a large pool of plausible

histories, we developed a more practical probabilistic algorithm⁶.

The approach is to treat evolutionary events as random, where alignment boundaries can be reused with small probabilities. We first specify a target distribution for gene cluster histories, which defines the scope of histories that we think plausible. For example, to make inference exclusively from histories without alignment boundary reuse, the target distribution should be uniform on all such histories and 0 otherwise. In practice, we define the target distribution as $\pi(\vec{O}_T | X) \propto e^{-5(T+r)}$, where $\vec{O}_T = (O_1, O_2, \dots, O_T)$ denotes a series of evolutionary events ordered backward in time, with O_i describing the type, the location, and the size of each evolutionary event, T denotes the total number of events occurred, and X denotes the gene cluster data. The exponential formula is arbitrarily chosen such that the number of events T and the number of alignment boundary reuses r are linearly penalized in the log-likelihood scale. Parameter “-5” is chosen to allow suboptimal solutions. This target distribution favors histories with fewer events and fewer alignment boundary reuses.

Directly sampling histories from the target distribution is computationally intractable. We therefore turn to sequential importance sampling (SIS)¹⁰. Suppose that t most recent events have been reconstructed, the SIS algorithm samples the next event O_{t+1} backward in time from a trial distribution $g_t(O_{t+1} | \vec{O}_t, X)$, where \vec{O}_t denotes the t events. The SIS algorithm sequentially samples events from a series of trial distributions until all alignments in the dot-plot are resolved, which then produces a history reconstruction \vec{O}_T . We repeat the SIS algorithm to obtain many history reconstructions, and we calculate a weight for each reconstruction as $w = \pi(\vec{O}_T | X) / \prod_{t=0}^{T-1} g_t(O_{t+1} | \vec{O}_t, X)$. The weight is used to adjust for the sampling bias. Given m reconstructions $\vec{O}_{T_1}^{(1)}, \vec{O}_{T_2}^{(2)}, \dots, \vec{O}_{T_m}^{(m)}$ and their weights w_1, \dots, w_m , we infer evolutionary parameters using a weighted average of function $u(\vec{O}_T)$ as $E[u(\vec{O}_T)] \cong \left(\sum_{i=1}^m w_i u(\vec{O}_{T_i}^{(i)}) \right) / \left(\sum_{i=1}^m w_i \right)$. Letting $u(\vec{O}_T) = T$, for example, allows us to estimate the number of events that occurred in a gene cluster evolution. The basic algorithm in the previous section is a special case of the SIS approach, with $g_t(O_{t+1} | \vec{O}_t, X)$ uniform on all events O_{t+1} satisfying the the basic algorithm's criterion and 0 otherwise. If the no-alignment-boundary-reuse assumption hold true, the SIS algorithm will efficiently and precisely produce the same duplication reconstructions as given by the basic algorithm. Additional details about the SIS algorithm can be found in our previous paper⁶.

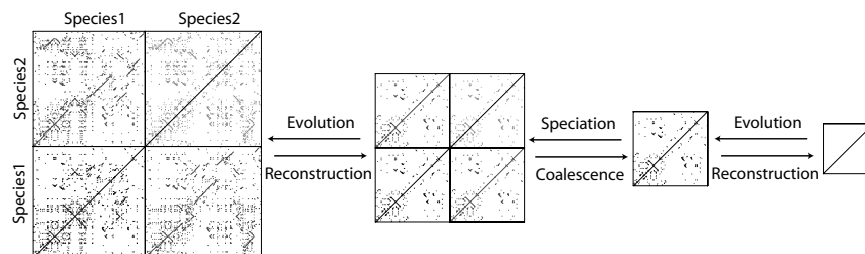


Figure 3. Two-species history reconstruction for gene clusters evolved after 150 duplications and 1 speciation. From left to right: 1) joint dot-plot of self-alignments (lower-left and upper-right quadrants) and inter-species pairwise alignments (upper-left and lower-right); 2) joint dot-plot immediately after speciation; 3) joint dot-plot immediately before speciation; 4) original duplication-free ancestor.

2.3. Extension to Multi-Species Reconstruction

Under certain conditions, it is easily shown that the basic algorithm for single-species history reconstruction can be directly applied to solve the multi-species history reconstruction problem. The conditions are similar to those for a single species, except that we now assume gene clusters evolve exclusively through segmental duplications in all species, and alignment boundaries are neither reused within species nor across species. The key idea is to treat speciation events as duplications that double the ancestral genomes. We first concatenate the orthologous gene clusters of all species together, then treat the joint sequence as a new gene cluster sequence for an unknown species. As a result, the single-species basic algorithm can be used to reconstruct the duplication history of the joint gene cluster. It further follows that any reconstructed duplication histories for the joint gene cluster is optimal. In practice, the SIS algorithm is used instead to infer deletion events and handle data complications. Since we know exactly how the multi-species sequences are concatenated, it is straightforward to convert the solution back to a history reconstruction for multi-species gene clusters. Our algorithm allows identification of speciation events as the entire sequence of one species being removed in reconstruction. The total number of events in a multi-species history reconstruction equals the number of duplications and deletions plus the number of speciation events. An example of two-species history reconstruction is shown in Fig. 3.

2.4. *Coalescence between Species*

If the no-alignment-boundary-reuse assumption is violated, the above multi-species reconstruction algorithm can no longer guarantee correct history reconstructions. A more challenging yet critical issue is the coalescence between species. A coalescence event is the inverse of a speciation event, where two sibling species coalesce into their common ancestor. By concatenating gene clusters sequences together, we may reconstruct some histories that are biologically impossible. For example, an event may be reconstructed as duplicating DNA segments across species, or a speciation event is reconstructed as multiple partial duplications. Those are just algorithm artifacts and must be corrected in reconstruction. Intuitively, we should impose a constraint to reconstruction stating that no alignments in pairwise (non-self) alignment dot-plots should be resolved unless it corresponds to a double-genome speciation event.

We use ortholog information to resolve the coalescence problem. Here, orthologs refer to DNA sequences between a pair of species that evolved from a common ancestor by speciation. Using orthologs, we can learn the occurrence time of an event relative to speciation. By the definition of orthologs between two species, it is obvious that no orthologous sequences should be resolved before reconstructing the speciation event, but all non-orthologous sequences should be resolved. If we restrict the algorithm to only reconstruct events within non-orthologous regions between pairs of species, eventually two alternative situations will occur: either all non-orthologous sequences between two species are resolved and thus their sequences become identical, or no more duplication and deletion events can be reconstructed to resolve the remaining non-orthologous sequences. When either case happens, the two species should be coalesced.

Identifying orthologs in two species is a hard problem in the presence of tandem gene clusters. In each step of reconstruction, we impute orthologous sequences between pairs of species using the corresponding pairwise alignments. In particular, given a pairwise alignment dot-plot of two species, we treat all forward pairwise alignments as matches and all reverse pairwise alignments and unmatched regions as gaps and mismatches. We then calculate a rough global pairwise alignment between the two species using this information, and we treat the DNA segments matched along the best global alignment path as orthologs between the two species. We currently do not consider segmental inversions, and thus reverse pairwise alignments cannot form orthologs. Since gaps are mostly created by duplications and

deletions, we only penalize mismatches but not gaps. The recursive scoring function of dynamic programming is given by

$$S[x_i^e, y_i^e] = \max_j \{S[x_j^e, y_j^e] + u(x_i^e - x_i^s) - \gamma \max(c(x_j^e, x_i^s), c(y_j^e, y_i^s))\}$$

Here, (x_i^s, y_i^s) and (x_i^e, y_i^e) denote the starting and ending positions of the i th forward alignment, respectively, while $c(a, b)$ measures the size of unmatched segments within region $[a, b]$, which are regarded as mismatches. Also, u and γ are match and mismatch weights. The maximization is taken over all alignments ending before the starting position of the i th alignment, i.e., $\forall \{j : x_j^e \leq x_i^s, y_j^e \leq y_i^s\}$. We recursively calculate the scores of all forward pairwise alignments between two species, and we do this for all pairs of species.

Having identified orthologous sequences between all pairwise gene clusters, our algorithm will reconstruct duplication and deletion events exclusively from non-orthologous regions in each species. For reconstructing duplications, only self-alignments are used, to prevent reconstructing across-species events. If a self-alignment involves a non-trivial part of non-orthologous regions (e.g. >200bp), we treat the alignment as non-orthologous. This is because different species may have evolved at the same functional regions independently, which will make ortholog identification ambiguous. By iteratively imputing orthologs and reconstruct evolutionary events, pairs of species will eventually have no more reconstructible duplication and deletion events from their non-orthologous regions. We then assume the two species to be identical and coalesce them. To coalesce two species, we first calculate a consensus sequence for the two species using their pairwise alignments with other species in the dot-plot. We then remove one species from the dot-plot and replace the other species with the consensus sequence. Note that as evolutionary events are gradually resolved using orthologs, identifying orthologs will in turn become easier.

In theory, a phylogenetic tree of multiple species is not required by our algorithm, and an unrooted tree can be estimated *a posteriori* from the data. In practice, if different species suggest contradicting events caused by incorrect ortholog identification, genome mis-assembly, alignment errors, or unknown evolutionary events, we should take advantage of known phylogenetic relationships to resolve such problems. A straightforward approach is to iteratively reconstruct the history of two closest species in the tree in a bottom-up approach.

3. Results

We evaluated our algorithm using simulated gene-cluster data for $n = 2, 3, 4$ species. The assumed trees resemble that for human, chimpanzee, orangutan, and rhesus, i.e., each additional species is an outgroup species of the existing ones. Starting from a single ancestral sequence of 500 kbp, we randomly simulated k duplication and deletion events, with frequency 98% and 2%, respectively, until the first speciation. We then doubled the current sequence in correspondence to a speciation event. Between every two adjacent speciation events, we continued to simulate k events in each lineage, and repeated the process until all n species appeared in the tree. Finally, we simulated k additional events in each species. The duplication and deletion events were generated according to distributions estimated from analysis of 165 biomedically interesting human gene clusters⁶, totalling around 111 million bases. Between every adjacent pair of events in each lineage, we further introduced 0.2% mutations to diversify the sequences. When simulating $n = 4$ species with $k = 20$, for example, the sequence percentage identities between species are reduced by between 8% and 24%.

We used our algorithm, including data preprocessing, to estimate the number of duplication, deletion, and speciation events for each dataset. Given $n = 2, 3, 4$ species and $k = 5, 10, 15, 20$ events between adjacent speciations in each lineage, the true number of events for a gene cluster dataset is $kn(n+1)/2 + n - 1$. The results from 20 datasets under each setting are shown in Fig. 4. Our algorithm achieved very high accuracy in estimating the total number of events, with about 1% over-estimation. We further checked the number of events that occurred in downstream species after each speciation. Given m downstream species, the true number of downstream events is $k(m(m+1)/2 - 1) + m - 2$, and our estimations are shown along the lines in Fig. 4. Again, our results indicate accurate reconstruction of both coalescence times and the order of coalescence among multiple species. In terms of computational efficiency, the current implementation of the method can reconstruct a history of 100 events within 30 seconds on a regular personal computer. Since we concatenate the genomes of multiple species into a common dot-plot, the computation time for multiple species is mainly determined by the size of the joint dot-plot.

We also evaluated our method using actual sequence data from the CYP2 gene cluster, which is known to have undergone extensive lineage-specific duplications¹¹. The CYP2 genes in humans, chimpanzee, orangutan, rhesus and dog were identified by using data downloaded from

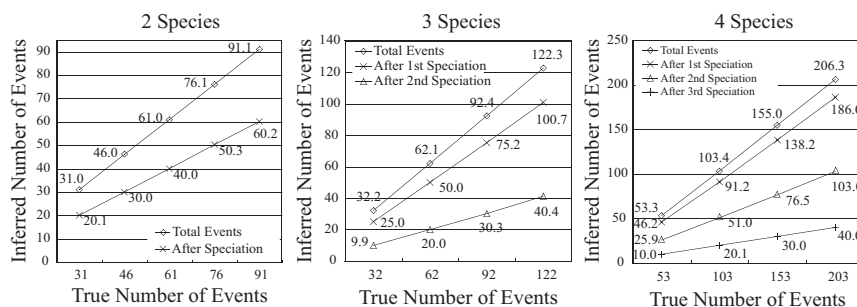


Figure 4. Reconstructed versus true number of events for simulated gene clusters. Given $n = 2, 3, 4$ species and $k = 5, 10, 15, 20$ events between adjacent speciations, the total number of events is $kn(n + 1)/2 + n - 1$, as shown on x-axis. The number of events among m downstream species after each speciation is $k(m(m + 1)/2 - 1) + m - 2$. All estimated numbers are shown along the lines.

the UCSC Genome Browser (<http://genome.ucsc.edu>) and by applying the GeneWise¹⁵ program. The human cluster contains 8 genes including 2 pseudo-genes. We constructed a gene tree using an amino-acid-based maximum-likelihood method, as shown in Fig. 5.

We built a second gene tree based on the evolutionary history reconstructed for the *CYP2* cluster by the method described in this paper. The method inferred 85 events; 6 of the duplication events involved genes. The two trees agreed perfectly. For instance, the maximum-likelihood tree (Fig. 5) and our reconstruction agree that *CYP2A6* and *CYP2A7* split after the human-chimp divergence, though our method gives the additional information that *CYP2A7* was copied and reinserted in the genome to create *CYP2A6*. The two approaches also agree that the common ancestor of *CYP2A13*, *c6*, and *o5* split from the common ancestor of *CYP2A7*, *c1*, and *o1* after the human-rhesus split and before the human-orangutan split. This example indicates how traditional phylogenetic reconstruction methods can be used to validate our results, though our analysis is more informative since it also treats non-coding DNA and infers the source and target of a duplication event.

4. Discussion

Gene duplications are a primary mechanism of evolution¹². Indeed, biologists have sought explanations for why so many duplicated genes are retained for long periods of time¹³. Studies of recent gene duplications have been greatly impeded by that lack of accurate sequence for gene clus-

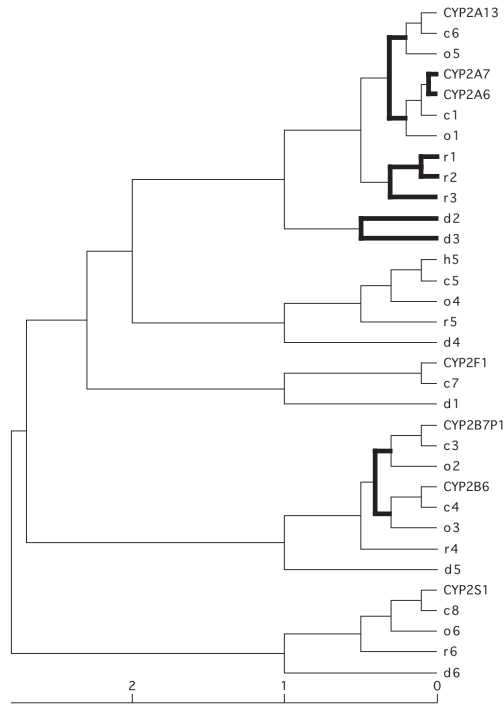


Figure 5. A maximum-likelihood Gene tree of the CYP2 gene cluster for 8 human genes, 8 chimpanzee genes (c1-c8), 6 orangutan genes (o1-o6), 6 rhesus genes (r1-r6) and 6 dog genes (d1-d6). Bold lines indicate duplication events after speciation of dog.

ters; most genomes are being sequenced by the so-called “whole-genome shotgun” approach, which has severe difficulties discriminating gene copies that exceed, say, 95% identity¹⁴, corresponding roughly to duplication events in the last 15 million years. Fortunately, gene clusters are beginning to be sequenced in multiple primates by techniques that resolve recent duplications^{16,17}, and there will soon be ample data to evaluate methods for reconstructing the evolutionary history of tandem gene clusters.

This current paper is motivated by the belief that the most serious deficiency of current whole-genome alignments is inadequate handling of tandem gene clusters. Assuming that a satisfactory approach can be developed for aligning gene clusters, the solution can be “spliced into” whole-genome

alignments computed by other means. However, work remains before the method explored here can be considered totally reliable. For instance, inversions not associated with a duplication event currently cause problems for our approach to identifying orthologs. Another hurdle is presented by gene-conversion events¹⁸, which can present the appearance of an alignment with zones of differing percent identity. We anticipate that promising approaches will be developed by several groups, and that before long biologist will have free access to automatically generated alignments in gene clusters of higher quality than what is currently available.

References

1. O. Elemento, O. Gascuel and M. P. Lefranc, *Mol. Biol. Evol.* **19**, 278–288 (2002).
2. M. Lajoie, D. Bertrand, N. El-Mabrouk and O. Gascuel, *J. Comput. Biol.* **14**, 462–468 (2007).
3. Z. Jiang, H. Tang, M. Ventura, M. F. Cardone, T. Marques-Bonet, X. She, P. A. Pevzner, E. E. Eichler, *Nat. Genet.* **39**, 1361–1368 (2007).
4. J. Ma, A. Ratan, B. J. Raney, B. B. Suh, L. X. Zhang, W. Miller and D. Haussler. *J. Comput. Biol.* To appear (2008).
5. J. Ma, A. Ratan, B. J. Raney, B. B. Suh, W. Miller and D. Haussler. *Proc. Natl. Acad. Sci. USA*. To appear (2008).
6. Y. Zhang, G. Song, T. Vinar, E. D. Green, A. Siepel and W. Miller, *Proceedings of RECOMB* (2008).
7. S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller, *Genome Res.* **13**, 103–107 (2003).
8. J. H. Nadeau and B. A. Taylor, *Proc. Natl. Acad. Sci. USA* **81**, 814–818 (1984).
9. P. Pevzner and G. Tesler, *Proc. Natl. Acad. Sci. USA* **100**, 7672–7677 (2003).
10. J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York (2001).
11. H. Wang, K. M. Donley, D. S. Keeney and S. M. G. Hoffman, *Environmental Health Perspectives* **111**, 1835–1842 (2002).
12. S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin (1970).
13. A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. Yan and J. Postlethwait. *Genetics* **151**, 1531–1545 (1999).
14. E. D. Green, *Nat. Rev. Genet.* **2**, 573–573 (2001).
15. E. Birney, M. Clamp and R. Durbin. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
16. Rhesus Macaque Genome Sequencing and Analysis Consortium. *Science* **316**, 222–234 (2007).
17. B. Hurler, W. Swanson, NISC Comparative Sequencing Program and E. D. Green. *Genome Res.* **17**, 276–286 (2007).
18. J.-M. Chen D. N. Cooper, N. Chuzhanova, C. Ferec and G. P. Patrinos. *Nature Reviews Genetics* **8**, 762–775 (2007).