Reconstructing Histories of Complex Gene Clusters on a Phylogeny

Tomáš Vinař[1], Broňa Brejová[1], Giltae Song[2], and Adam Siepel[3]

[1] Faculty of Mathematics, Physics and Informatics, Comenius University,
Mlynská Dolina, 842 48 Bratislava, Slovakia
[2] Center for Comparative Genomics and Bioinformatics, 506B Wartik Lab,
Penn State University, University Park, PA 16802, USA
[3] Dept. of Biological Statistics and Comp. Biology, Cornell University, Ithaca,
NY 14853, USA

Abstract. Clusters of genes that have evolved by repeated segmental duplication present difficult challenges throughout genomic analysis, from sequence assembly to functional analysis. These clusters are one of the major sources of evolutionary innovation, and they are linked to multiple diseases, including HIV and a variety of cancers. Understanding their evolutionary histories is a key to the application of comparative genomics methods in these regions of the genome. We propose a probabilistic model of gene cluster evolution on a phylogeny, and an MCMC algorithm for reconstruction of duplication histories from genomic sequences in multiple species. Several projects are underway to obtain high quality BAC-based assemblies of duplicated clusters in multiple species, and we anticipate use of our methods in their analysis. Supplementary materials are located at http://compbio.fmph.uniba.sk/suppl/09recombcg/

1 Introduction

Segmental duplications cover about 5% of the human genome (Lander et al., 2001). When multiple segmental duplications occur at a particular genomic locus they give rise to complex gene clusters. Many functionally important families residing in such clusters are linked to various diseases, e.g. UGT1A (colorectal cancer; Tang et al. (2005)), UGT2 (prostate cancer; Hajdinjak and Zagradisnik (2004)), APOBEC3 (HIV; An et al. (2004)), CCL3 (HIV; Degenhardt et al. (2009)), HLA (multiple sclerosis; Bitti et al. (2001)), CST (Alzheimer's disease; Finckh et al. (2000)). Gene duplication is often followed by functional diversification (Ohno, 1970), and, indeed, genes overlapping segmental duplications have been shown to be enriched for positive selection (Gibbs et al., 2007).

In this paper, we describe a probabilistic model of evolution of gene clusters on a phylogeny, and devise an algorithm for inference of highly probable duplication histories and ancestral sequences. We apply our algorithm to simulated sequences on the human-chimp-macaque phylogeny, as well as to real clusters assembled from available BAC sequencing data.

F.D. Ciccarelli and I. Miklós (Eds.): RECOMB-CG 2009, LNBI 5817, pp. 150–163, 2009.

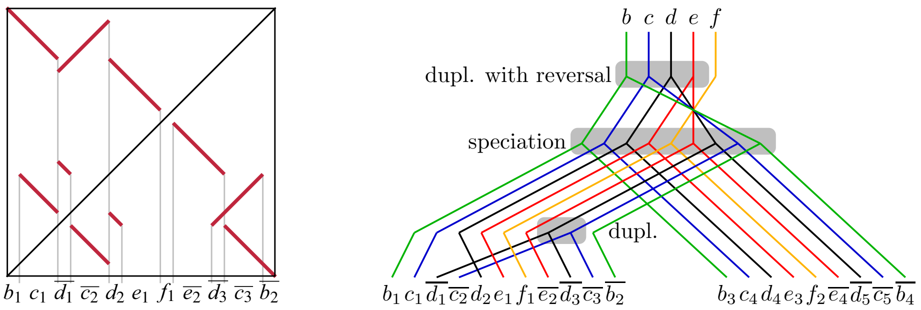

b_1 c_1 $\overline{d_1}$ $\overline{c_2}$ d_2 e_1 f_1 $\overline{e_2}$ $\overline{d_3}$ $\overline{c_3}$ $\overline{b_2}$
b c d e f
dupl. with reversal
speciation
dupl.
b_1 c_1 $\overline{d_1}$ $\overline{c_2}$ d_2 e_1 f_1 $\overline{e_2}$ $\overline{d_3}$ $\overline{c_3}$ $\overline{b_2}$
b_3 c_4 d_4 e_3 f_2 $\overline{e_4}$ $\overline{d_5}$ $\overline{c_5}$ $\overline{b_4}$

Fig. 1. Sequence atomization and segment trees. (left) Self-alignment of a sequence resulting from two duplication events. Lines represent local alignments. There are five types of atomic segments (b, c, d, e, f). For example, type d has three copies: one on the forwards strand (d_2) and two on the reverse strand $(\overline{d_1}, \overline{d_3})$. (right) Segment trees of individual atom types organized in a tube tree (inspired by similar graphics of Brian Raney). Tube trees visualize a duplication history of several atomic segments in the context of the species tree, and their locations in the extant and ancestral sequences. The figure shows a tube tree with two duplications and one speciation.

Previously, Elemento et al. (2002) and Lajoie et al. (2007) studied the reconstruction of gene family histories in a simplified model, where gene clusters contain several tandemly repeated copies of a single gene. Elemento et al. (2002) consider tandem duplications only, while Lajoie et al. (2007) also consider inversions of variable lengths. However, more complex models are necessary to address evolution of gene clusters in the human genome. In recent work, genes have been replaced by generic *atomic segments* (Zhang et al., 2008; Ma et al., 2008; Bertrand et al., 2008). Briefly, a self-alignment is constructed by a local alignment program (e.g., blastz (Schwartz et al., 2003)), and only alignments above certain threshold (e.g., 93% for human-macaque split) are kept. Alignment boundaries mark *breakpoints*, and the sequences between neighboring breakpoints are considered atomic segments (Fig.1). Under reasonable evolutionary models, the sequence similarities between atomic segments are *transitive*, and the set of atomic segments can be decomposed into equivalence classes, or *atom types*, such that all segments of the same type have similar sequence. In this way, the nucleotide sequence is transformed into a simpler sequence of atoms.

The task of *duplication history reconstruction* is to find a sequence of events (duplications, deletions, and speciations) that starts with an ancestral sequence of atoms, where each atom occurs once, and ends with atomic sequences of extant species. Such a history directly implies *segment trees* of individual atomic types, implicitly reconciled with the species tree (Fig.1). Each segment tree represents duplication and speciation events concerning one atom type, similarly as gene tree represents history of a single gene. A common way of looking at these histories is from the most recent events back in time. In this context, we can start from the extant sequences, and *unwind* events one-by-one, until the ancestral sequence is reached.

Zhang et al. (2008) sought parsimonious solutions of this problems given the sequence from a single species. In particular, they proved a necessary condition to identify candidates for the latest duplication operation, assuming no reuse of breakpoints. After unwinding the latest duplication, the same step is repeated to identify the second latest duplication, etc. Zhang *et al.* showed that any sequence of candidate duplications leads to a history with the same number of duplication events under no-breakpoint-reuse assumption. As a result, there may be an exponential number of most parsimonious solutions to the problem, and it may be impossible to reconstruct a unique history. Recently, Zhang et al. (2009) extended the same approach to simultaneous inference in multiple species. A similar parsimony problem has also been recently explored by Ma et al. (2008) in the context of much larger sequences (whole genomes) and a broader set of operations (including inversions, translocations, etc.). In their algorithm, Ma *et al.* reconstruct phylogenetic trees for every atomic segment, and reconcile these segment trees with the species tree to infer deletions and rooting. The authors give a polynomial-time algorithm for the history reconstruction, assuming no-breakpoint-reuse and correct atomic segment trees. Both methods make use of extensive heuristics to overcome violations of their assumptions in real data.

The no-breakpoint-reuse assumption is often justified by the argument that in long sequences, it is unlikely that the same breakpoint is used twice (Nadeau and Taylor, 1984). However, there is evidence that breakpoints do not occur uniformly throughout the sequence, and that breakpoint reuse is in fact frequent (Peng et al., 2006; Becker and Lenhard, 2007). Moreover, breakpoints located close to each other lead to short atoms that cannot be reliably identified by sequence similarity algorithms and categorized into atom types. For example, in our simulated data (Section 4), approximately 2% of atoms are shorter than 20bp. These short atoms may appear as additional breakpoint reuses. Thus, no-breakpoint-reuse can be a useful guide, but cannot be relied on in application to real data sets. We have also examined the assumption of correctness of segment trees inferred from sequences of individual segments. For segments shorter than 500bp (39% of all segments in our simulations) 69% of the trees were incorrectly reconstructed, and even for segments 500-1,000bp long, a substantial fraction (46%) is incorrect (Fig.2).

Here, we propose a probabilistic model for sequence evolution by duplication, and we design a sampling algorithm that explicitly accounts for uncertainty in the estimation of segment trees and allows for breakpoint reuse. The results of Zhang et al. (2008) suggest that, in spite of an improved model, there may still be many solutions of similar likelihood. The stochastic sampling approach allows us to examine multiple solutions in the same framework and extract expectations for quantities of interest (e.g., the expected number of events on individual branches of the phylogeny, or local properties of the ancestral sequences).

Our problem is also closely related to the problem of reconstruction of gene trees and their reconciliation with species trees. Even though the recent algorithms for gene tree reconstruction (e.g., Wapinski et al. (2007)) consider genomic context of individual genes as an additional piece of information, our

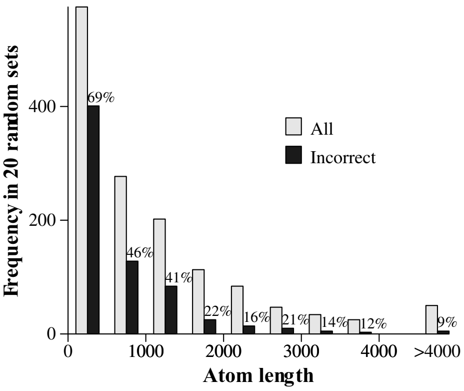


Fig. 2. Distribution of atomic segment lengths and accuracy of segment tree inference in 20 simulated fast-evolving clusters (see Section 4). The gray bars show the numbers of segment types. The black bars show the percentages of segment types for which the highest posterior probability unrooted segment tree inferred by MrBayes (Ronquist and Huelsenbeck, 2003) does not match the correct segment tree.

new methods aim to fully explain genomic context of individual genes through reconstructed duplication histories.

2 Probabilistic Model

In this section, we give a probabilistic model of evolution of gene clusters through segmental duplication on a given species tree T. Later, we use this model for inference of duplication histories and to generate simulated gene clusters.

We start with an ancestral sequence of length N. The sequence evolves by duplications, deletions and substitutions. A *duplication* copies a source region and inserts the new copy at a target position in the sequence, either on the forward strand (with probability $1 - P_i$) or on the reverse strand (with probability P_i). A duplication can be characterized by four coordinates: a *centroid* (the midpoint of the region between the leftmost and rightmost end of the duplication), the *length* of the source region, the *distance* between the source and the target, and a *direction* (from left to right or from right to left). The centroid is chosen uniformly, and the length and distance are chosen from given distributions (see below). Note that some centroid, distance, and length combinations are invalid; those combinations are rejected. Similarly, a *deletion* removes a portion of the sequence, and can be characterized by its *centroid* and *length.* Each event is a deletion with probability P_x, and a duplication with probability $(1 - P_x)$. This process straightforwardly defines the probability $P(E \mid \mathrm{len})$ of any duplication or deletion event E. Here, len is the length of the sequence just before the event E. The number of events on each branch is governed by a Poisson process with rate λ, and thus the probability of observing k events on a branch of length ℓ is $P_n(k, \ell) = (\lambda\ell)^k e^{-\lambda\ell}/k!$.

A duplication history H generated in this way implies a set $\sigma(H)$ of *atomic segments* of several types and a segment tree T_x for each atom type x. The substitutions in the nucleotide sequences of atom x are governed by the HKY substitution model along the segment tree T_x.

We can compute the joint probability $p(H, X)$ of a given set of extant sequences X and a history H (up to a normalization constant) as follows. Let T be a species tree with branches $b_1, b_2, \ldots$ Then:

$$p(H, X) \propto \prod_{b_i \in T} P(H, b_i) \times \prod_{x \in \sigma(H)} P(X_x \,|\, T_x), \tag{1}$$

where $P(H, b_i)$ is the probability of events of history H that occur on branch b_i of the species tree, X_x represents nucleotide sequences of atoms of type x, and $P(X_x \,|\, T_x)$ is the probability of these sequences given tree T_{x_i}. For a sequence of events $E_1, \ldots, E_k$ on branch b_i, the probability $P(H, b_i)$ is simply:

$$P(H, b_i) = P_n(k, \ell) \prod_{j=1}^{k} P(E_j \,|\, \mathrm{len}(j-1)) \tag{2}$$

where $\mathrm{len}(j-1)$ is the length of the sequence before event E_j.

To reduce the number of model parameters, we use geometric distributions to model lengths and distances of duplication events. To estimate these distributions, we have used the lengths and distances estimated by Zhang et al. (2008) from human genome gene clusters (mean length 14,307, mean distance 306,718). The geometric distributions seem to approximate the observed length distributions reasonably well. Similarly, we estimated the probability of duplication with inversion as $P_i = 0.39$ from the same data, we set the probability of deletion as $P_x = 0.05$, and the length distribution of deletions matches the distribution of duplication lengths. In our model, we do not allow inversions that are not accompanied by a duplication.

Note that for our application, the normalization constant for $p(H, X)$ does not need to be computed. We assume a uniform prior on length distribution of ancestral lengths. This has only a small effect for fixed extant sequences, since the ancestral sequence should contain a single occurrence of each segment type, and therefore the ancestral length is determined mostly by the length of individual atomic segment types. Some combinations of centroids, distances, and lengths will be rejected, but we assume that in long enough sequences, the effect of this rejection step will be negligible and we ignore it altogether.

In the inference algorithm below, we compute likelihood $P(X_x \,|\, T_x)$ and branch lengths for each segment tree separately. This independence assumption simplifies computation and allows variation of rates and branch lengths between atoms. This is desirable, since sequences of different functions may evolve at different substitution rates, and selection pressures may change the proportions of individual branch lengths. Nonetheless, branch lengths tend to be correlated among segment trees when individual atoms are duplicated together, and this information is lost by separating the likelihood computations. We are working on a more systematic solution to the problem of rate and branch length variation.

3 Metropolis–Hastings Sampling

For inference of duplication histories, we use the Metropolis–Hastings Markov chain Monte Carlo algorithm (Hastings, 1970) to sample from the posterior probability distribution $p(H \mid X)$ defined in the previous section, conditional on the extant sequences X and their atomization. The result of the algorithm is a series of samples that can be used to estimate expectations of quantities of interest (e.g., the number of events on individual branches, posteriors of individual segment trees, and particular ancestral sequences), or to examine high likelihood histories.

Briefly, the Metropolis–Hastings algorithm defines a Markov chain whose stationary distribution is the target distribution, but the moves of the Markov chain are defined through a different *proposal distribution* due to the difficulties of sampling from the target distribution directly. We start by initializing sample history H_0. In each iteration, we use a randomized *proposal algorithm* described below to propose a candidate history H'_i according to a distribution conditional on sample H_{i-1}. Sample H'_i is either *accepted* ($H_i := H'_i$) with probability $\alpha(H_{i-1}, H'_i)$, or *rejected* ($H_i := H_{i-1}$) otherwise. The acceptance probability $\alpha(H, H')$ is used to ensure that the stationary distribution of the Markov chain is indeed the target distribution (Hastings, 1970):

$$\alpha(H, H') = \min\left(1, \frac{p(H'|X)q(H \mid H')}{p(H|X)q(H' \mid H)}\right), \tag{3}$$

where $q(H' \mid H)$ is the probability of proposing history H' if the previous history was H.

The proposal algorithm starts by sampling an unrooted *guide tree* T_x for every atom type x. The segment trees implied by the proposed history will be later rooted and refined from these guide trees. Guide tree T_x is sampled from the posterior distribution of the trees conditional on a fixed multiple alignment of all instances of atom type x. Sampling guide trees accounts for uncertainty in the estimation of segment trees. We collapse branches with fewer than 5 expected substitutions over the length of the atom sequence, since such short branches usually cannot be estimated reliably. Thus, the guide trees for shorter atoms, where uncertainty is high, will be close to uninformative star trees, while the trees for longer atoms will remain more resolved.

The proposal algorithm samples a history consistent with the given set of guide trees, starting at the leaves of the trees, and progressively sampling groups of atom pairs to merge, until only a single copy of each atom remains. Merging of two groups of atoms corresponds to unwinding one duplication. To obtain a valid history consistent with the guide trees, each of the two groups has to be a contiguous subsequence in the current atomic sequence. Also, the corresponding atoms of the two groups must be of the same type. Finally, the corresponding atoms must be cherries in their guide trees (see also Fig.3; the leaves x_i and x_j are cherries in T_x if they have the same parent.)

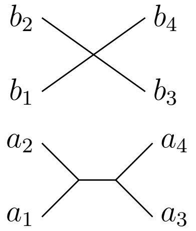


Fig. 3. Consistency of histories with the guide trees. Example of guide trees for an atomic sequence $a_1b_1a_2b_2a_3b_3a_4b_4$. The segment tree T_b was collapsed into an uniformative star tree, perhaps since atom b is too short. Duplication (a_1b_1, a_2b_2) is consistent with the guide trees since (a_1, a_2) and (b_1, b_2) are cherries in the corresponding guide trees, but duplication (a_1b_1, a_3b_3) is inconsistent, as is $(a_1b_1a_2b_2, a_3b_3a_4b_4)$.

For example, if the most recent duplication copied atoms a_1b_1 to atoms a_2b_2 then a_1 and a_2 must be cherries in the tree T_a, and b_1 and b_2 must be cherries in the tree T_b. Unwinding of this duplication will correspond to removal of a_2 and b_2 from the trees T_a and T_b and from the atomic sequence. Now, the same conditions can be applied to the second latest duplication. In this way, a particular set of guide trees can significantly restrict the set of possible histories.

The sampling distribution over candidate groups of atoms is determined by a series of heuristic penalties described in Appendix A, favoring longer duplications, those not inducing breakpoint reuse, and those that were used in the previous sample H_{i-1}. Even though this algorithm employs a number of heuristics to improve the overall performance of the sampler, they only affect the mixing rate and convergence properties of the sampling, not the asymptotic correctness of the MCMC algorithm.

The multiple alignment for each segment type is created by MUSCLE (Edgar, 2004). Even though it is possible to sample multiple alignments to prevent potential alignment errors from propagating throughout the whole analysis (Holmes and Bruno, 2001), such sampling is by itself computationally intensive. Given that in this paper we consider sequences of greater than 90% similarity, we do not expect multiple alignments to be a major source of error in our reconstructions. Trees, branch lengths, and HKY nucleotide substitution model parameters are sampled by MrBayes (Ronquist and Huelsenbeck, 2003) with uniform prior over tree topologies, and default priors for the other parameters. For each segment type x, all the tree samples are precomputed in a run of 10,000 iterations with a burn-in of 2,500 samples, keeping every 10th sampled tree. In every iteration of the history proposal algorithm, we keep the previous guide tree with 95% probability, otherwise we choose a new tree randomly from the pre-computed samples.

Deletion operations cannot be easily dated, and some of them cannot be even observed in the extant sequence. To address this problem, we attach each deletion to the most recent overlapping duplication or speciation and in the proposal algorithm, described in Appendix A, we always propose duplications and corresponding deletions together. To keep the running time feasible, we assume that there is at most one deletion following each duplication and that it does not extend beyond the boundaries of the corresponding duplication segment.

4 Experiments

We have implemented the MCMC sampler described above and verified its functionality on simulated data. For the simulations, we have estimated branch lengths and HKY model parameters (equilibrium frequencies and transition/transversion ratio) from the UCSC syntenic alignments (Karolchik et al., 2008) of human, chimp, and macaque on human chromosome 22.

We created 20 simulated gene clusters in each of the following two categories: slow evolving and fast evolving (duplication rate at 200 and 300 times substitutions per site, respectively). We have applied our algorithm to atomic segments derived from the simulation, with short ($<$ 500bp) atomic segments removed to emulate the increase in breakpoint reuse due to imperfect identification of alignment boundaries in real data sets (see Tab.1 for the data set overview). For each cluster, we ran two chains of the sampler from random starting points for up to 10,000 iterations each, discarding the first 2,500 samples as burn-in. The sampler seems to converge reasonably quickly (supplementary Fig.B1).

In the majority of cases, we predict the correct number of events (Tab.2; 14 out of 20 for slowly evolving, 15 out of 20 for fast evolving clusters). Note that in some cases the predicted number of events is lower than the actual number of events: this is likely due to events that become invisible in the extant sequences because of subsequent deletions. We have compared our results on human lineage to Zhang et al. (2008), and on the whole tree to Zhang et al. (2009). The performance has improved, especially in the case of fast evolving clusters. We also examined the correctness of distribution of events along the phylogeny (supplementary Tab.B1). Finally, we compared predicted and actual ancestral atomic sequences. To quantify the differences between the sequences, we have counted the number of breakpoints required to transform the predicted ancestral sequence to the actual ancestor (Fig.4). In the majority of cases (31 out of 40), the expected number of breakpoints is smaller than 0.5.

Beyond the simulated data, we have applied our algorithm to the following gene cluster sequences: PRAME (human-macaque phylogeny), AMY (human-macaque

Table 1. Overview of simulated and real data sets

	slow rate			**fast rate**			PRAME	AMY	UGT1A
	min	max	mean	min	max	mean			
Seq. len (kb)	91	295	172	120	387	219	1000, 200	221, 170	210, 210, 250
No. atom types	15	53	36	39	57	48	39	44	55
No. duplications	5	24	15	18	29	23	34.9 ± 0.8	23.4 ± 0.8	22.9 ± 0.8
No. deletions	0	3	0.8	0	3	1.1	9.4 ± 1.9	15.2 ± 1.9	20.2 ± 1.3
Species	H,C,R			H,C,R			H,R	H,R	H,C,O

Table 2. Performance evaluation. The table shows the histogram of differences between the real number of events and the predicted number of events along the human lineage and on the whole tree on the 40 simulated data sets (20 with slow duplication rate 200, 20 with fast duplication rate 300). MCMC: rounded expected number of events from all samples. ML: highest likelihood sample. We compare to results of Zhang et al. (2008) on single species (Z2008) and Zhang et al. (2009) on the whole tree (Z2009). Note that the results of the two programs are not directly comparable, since our program was run on correct atomization with short atoms filtered out (giving Z2008 and Z2009 advantage of smaller amount of breakpoint reuse in the data), while Z2008 and Z2009 used their own built-in atomization method (giving advantage to our program, since the results of their atomization may be potentially incorrect).

	human lineage										**whole tree**									
	slow rate					**fast rate**					**slow rate**					**fast rate**				
Method	< 0	0	1	2	> 2	< 0	0	1	2	> 2	< 0	0	1	2	> 2	< 0	0	1	2	> 2
MCMC	1	15	1	3	0	1	16	1	2	0	3	14	2	0	1	2	15	2	1	0
ML	1	15	1	2	1	1	16	0	3	0	3	14	1	1	1	2	13	2	3	0
Z2008	3	14	2	1	0	1	6	4	3	6										
Z2009											0	8	5	2	5	0	0	3	2	15

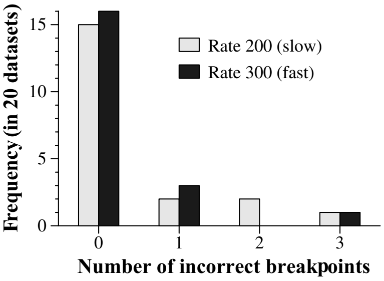


Fig. 4. Histogram of expected number of incorrect breakpoints on the 40 simulated data sets. The number of breakpoints required to transform predicted sequences to actual sequences is computed over all MCMC samples and the average is rounded to the closest integer.

phylogeny), and UGT1A (human-chimp-orang phylogeny). PRAME cluster (preferentially expressed antigen in melanoma) is one of the most active gene clusters in the human genome, and shows strong evidence of positive selection (Birtle et al., 2005; Gibbs et al., 2007). AMY cluster contains five amylase genes that are responsible for digestion of starch. It appears to have expanded much faster in humans than in other primates, according to aCGH experiments (Dumas et al., 2007). The UGT1A cluster consists of multiple isoforms of a single gene, instrumental in transforming small molecules into water-soluble and excretable metabolites. This gene has at least thirteen unique alternate first exons resulting from duplications at various stages of mammalian evolution. UGT1A provides an unusual opportunity for studying promoter evolution.

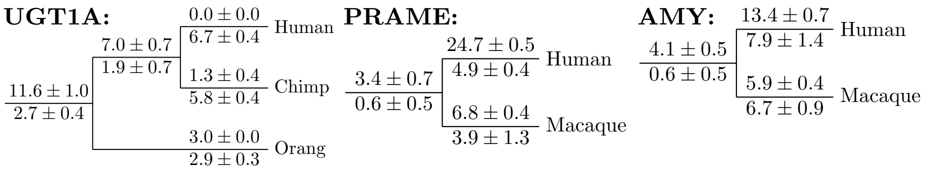


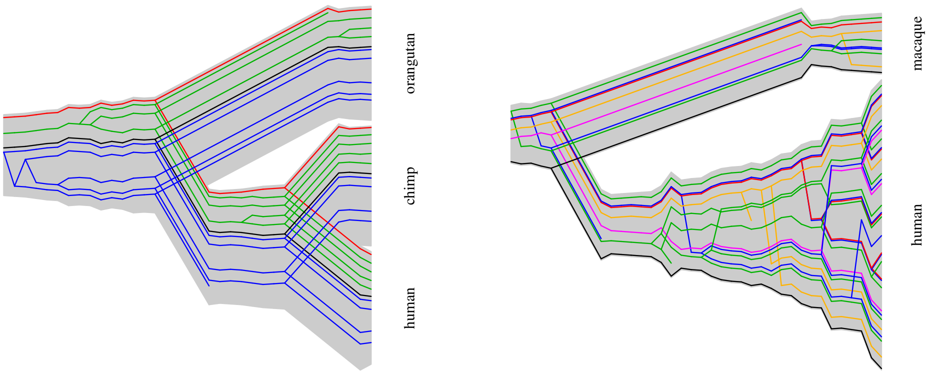
Fig. 5. Estimated numbers of events. For each cluster, we show the posterior mean and standard deviation of the number of duplications (above the branch) and deletions (below the branch) as assessed by MCMC sampling. The root branch shows events up until 90% sequence similarity cutoff.

Recently duplicated clusters are grossly missassembled in shotgun based genomes (Green, 2001; Zhang et al., 2008). To prepare our data sets, we have first screened sequenced BACs from chimp, orangutan, and macaque for similarity with the corresponding human sequences, assembled BACs into longer contigs, and selected subregions whose ends showed clear homology with upstream and downstream sequences of the human cluster. To identify atomic segments, we divide the sequences into equally sized 500bp windows, and for each window we find approximate copies in all available sequences at 90% identity cutoff. The atoms are assigned in a greedy way (starting from the windows with the largest number of copies), and windows overlapping already assigned atoms are discarded. Finally, atoms that always occur in pairs are merged into longer atoms. Table 1 shows an overview of the resulting sequences and atoms.

For each cluster, we ran five chains from different starting points for 5,000–10,000 samples, discarding the first 2,500 samples. We have estimated the number of duplications and deletions overall (Table 1), and on individual branches of the phylogeny (Fig.5). The estimated numbers of duplications for PRAME and AMY are comparable to those of Zhang et al. (2008). With UGT1A, we obtain higher estimates possibly due to differences in our atomization procedure or effects of the additional species in the analysis.

Figure 6a shows the highest likelihood reconstruction of the history of the UGT1A cluster. The cluster contains several isoforms of the same five-exon gene. Exons 2-5 are shared among all the isoforms, while exon 1 is alternatively spliced. The reconstruction shows division of the first exons into three distinct groups, and their ortholog/paralog relationships in human, chimp, and orangutan.

While the duplication history of the UGT1A cluster consists of mostly ancient events, the PRAME cluster (Fig.6b) shows recent large-scale duplications, especially in the human lineage. Figure 6 shows such events by several co-linear bifurcations at the same level of the tube tree. The reconstruction of the history by traditional methods (gene tree/species tree reconciliation) is complicated by the presence of recent duplications (99% similarity), and chimeric genes (Gibbs et al., 2007). We address these issues by considering multiple guide trees for each atom as well as spatial configuration of atoms in multiple species. However, the predicted history is by no means perfect. Rhesus sequence exhibits large regions that apparently arose by a single event, yet we split this event due to mistakes



(a) UGT1A cluster consists of several five exon isoforms of the same gene. Exons 2-5 (red) are shared among all the isoforms, the first exons (blue, black, green) are alternatively spliced.

(b) PRAME cluster consists of multiple copies of the PRAME gene. A typical copy has three coding exons, the highlighted atoms overlap exon 2. Some of the genes were pseudogenized.

Fig. 6. Highest likelihood reconstruction of duplication histories. The branch lengths in the figures do not correspond to the actual branch lengths. The atoms are ordered in their order along the genomic sequences (extant and ancestral).

in atomic representation. We expect that improved procedure for segmenting sequence into atoms will address this problem.

5 Discussion

In this paper, we have introduced a new model of evolution of gene clusters and designed an algorithm to reconstruct high probability evolutionary histories of these clusters. We have tested our method on both simulated and real data. Comparative genomics methods traditionally concentrate on sequences where 1:1 orthology can be established. In case of gene clusters, this is rarely the case. Our efforts in reconstruction of gene cluster histories will support further development of comparative genomic tools to analyze these complex regions.

Gene clusters should not be seen only as a confounding factor. The number of orthologous sequences, their divergence, and phylogenetic relationships greatly impact the accuracy of comparative genomic studies. For example, Kosiol et al. (2008) has shown that the sensitivity of positive selection scans is improved by considering sequences from a complex phylogeny. Studies based on orthologous regions between species can at present use a phylogeny of up to 10 orthologous copies of a particular mammalian gene from genomes sequenced at reasonable quality. On the other hand, some clusters contain many more copies with significantly more complex phylogeny even within a single species (for example, the PRAME cluster contains more than 30 copies in the human genome alone).

Thus, the gene clusters provide an opportunity for refined look at evolution of genes and genomes. Multiple sources of evidence suggest that many interesting developments in genomes happen within the boundaries of gene clusters, which further increases interest in their study. Multiple efforts are currently under way to BAC sequence selected gene clusters in multiple species and in multiple populations (Zhang et al., 2008; Zody et al., 2008). Accurate methods and models for reconstruction of duplication histories of these clusters are essential in understanding the evolution, function, and biomedical implications of these regions.

The general framework of our method allows future developments. One limitation of our sampler is its low sample acceptance ratio, indicating low level of mixing in the Markov chain. We plan to devise a systematic way for tuning the parameters in the proposal distribution towards better acceptance ratios. We also plan to improve the underlying probabilistic model. Currently the branch lengths in segment trees are chosen independently of the duplication history. Instead, we plan to consistently date duplication events on each branch, and use a scaling parameter for each atom type so that we can accurately model correlation between branch lengths of individual atom types and at the same type allow rate variation in different parts of the sequence. We use a simple HKY substitution model with variance in rates allowed between individual atomic segments. In future work, it will be possible to employ more complex models of sequence evolution, such as variable rate site models and models of codon evolution, within the same framework. Such extensions will allow us to identify sites and branches under selection in gene clusters in a principled way, and contribute towards better functional characterization of these important genomic regions. An interesting alternative approach might be to use combinatorial optimization instead of sampling to find maximum likelihood history in the above model.

Acknowledgements. We would like to thank Devin Locke and LaDeana Hillier at Washington University, St. Louis for providing us with the BAC sequences from chimp, orangutan, and macaque. We would also like to thank Webb Miller and Yu Zhang for helpful discussions on this problem. Research of TV and BB is funded by European Community FP7 grants IRG-224885 and IRG-231025.

References

An, P., et al.: APOBEC3G genetic variants and their influence on the progression to AIDS. J. Virol. 78(20), 11070–11076 (2004)

Becker, T.S., Lenhard, B.: The random versus fragile breakage models of chromosome evolution: a matter of resolution. Mol. Genet. Genomics 278(5), 487–491 (2007)

Bertrand, D., Lajoie, M., El-Mabrouk, N.: Inferring ancestral gene orders for a family of tandemly arrayed genes. J. Comput. Biol. 15(8), 1063–1067 (2008)

Birtle, Z., Goodstadt, L., Ponting, C.: Duplication and positive selection among hominin-specific PRAME genes. BMC Genomics 6, 120 (2005)

Bitti, P.P., et al.: Association between the ancestral haplotype HLA A30B18DR3 and multiple sclerosis in central Sardinia. Genet. Epidemiol. 20(2), 271–273 (2001)

Degenhardt, J.D., et al.: Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in Rhesus Macaques (Macaca mulatta). PLoS Genet. 5(1), e1000346 (2009)

Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R., Sikela, J.M.: Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res. 17(9), 1266–1267 (2007)

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

Elemento, O., Gascuel, O., Lefranc, M.P.: Reconstructing the duplication history of tandemly repeated genes. Mol. Biol. Evol. 19(3), 278 (2002)

Finckh, U., et al.: Genetic association of a cystatin C gene polymorphism with late-onset Alzheimer disease. Arch. Neurol. 57(11), 1579–1583 (2000)

Gibbs, R., et al.: Evolutionary and biomedical insights from the rhesus macaque genome. Science 316(5822), 222–224 (2007)

Green, E.D.: Strategies for the systematic sequencing of complex genomes. Nat. Rev. Genet. 2(8), 573 (2001)

Hajdinjak, T., Zagradisnik, B.: Prostate cancer and polymorphism D85Y in gene for dihydrotestosterone degrading enzyme UGT2B15: Frequency of DD homozygotes increases with Gleason Score. Prostate 59(4), 436–439 (2004)

Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109 (1970)

Holmes, I., Bruno, W.J.: Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics 17(9), 803–810 (2001)

Karolchik, D., et al.: The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res. 36(Database issue), D773–D779 (2008)

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., Siepel, A.: Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4(8), e1000144 (2008)

Lajoie, M., Bertrand, D., El-Mabrouk, N., Gascuel, O.: Duplication and inversion history of a tandemly repeated genes family. J. Comput. Biol. 14(4), 462–468 (2007)

Lander, E.S., et al.: Initial sequencing and analysis of the human genome. Nature 409(6822), 860–921 (2001)

Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., Haussler, D.: The infinite sites model of genome evolution. Proc. Natl. Acad. Sci. USA 105(38), 14254–14261 (2008)

Nadeau, J.H., Taylor, B.A.: Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA 81(3), 814–818 (1984)

Ohno, S.: Evolution by Gene Dupplication. Springer, Berlin (1970)

Peng, Q., Pevzner, P.A., Tesler, G.: The fragile breakage versus random breakage models of chromosome evolution. PLoS Comput. Biol. 2(2), e14 (2006)

Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12), 1572–1574 (2003)

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., Miller, W.: Human-mouse alignments with BLASTZ. Genome Res. 13(1), 103–107 (2003)

Tang, K.S., Chiu, H.F., Chen, H.H., Eng, H.L., Tsai, C.J., Teng, H.C., Huang, C.S.: Link between colorectal cancer and polymorphisms in the uridine-diphosphoglucuronosyltransferase 1A7 and 1A1 genes. World J. Gastroenterol 11(21), 3250–3254 (2005)

Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23(13), i549–i558 (2007)

Zhang, Y., Song, G., Hsu, C.-H., Miller, W.: Simultaneous History Reconstruction for Complex Gene Clusters in Multiple Species. In: Pacific Symposium on Biocomputing (PSB), vol. 14, pp. 162–173 (2009)

Zhang, Y., Song, G., Vinar, T., Green, E.D., Siepel, A., Miller, W.: Reconstructing the Evolutionary History of Complex Human Gene Clusters. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS (LNBI), vol. 4955, pp. 29–49. Springer, Heidelberg (2008)

Zody, M.C., et al.: Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat. Genet. 40, 1076–1083 (2008)