Reconstructing the Evolutionary History of Complex Human Gene Clusters

Yu Zhang[1,2], Giltae Song[1], Tomáš Vinař[3], Eric D. Green[4], Adam Siepel[3], and Webb Miller[1]

[1] Center for Comparative Genomics and Bioinformatics, 506B Wartik Lab, Penn State University, University Park, PA 16802, USA
[2] Department of Statistics, Penn State University, University Park, PA 16802, USA
[3] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
[4] Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Abstract. Clusters of genes that evolved from single progenitors via repeated segmental duplications present significant challenges to the generation of a truly complete human genome sequence. Such clusters can confound both accurate sequence assembly and downstream computational analysis, yet they represent a hotbed of functional innovation, making them of extreme interest. We have developed an algorithm for reconstructing the evolutionary history of gene clusters using only human genomic sequence data. This method allows the tempo of large-scale evolutionary events in human gene clusters to be estimated, which in turn will facilitate primate comparative sequencing studies that will aim to reconstruct their evolutionary history more fully.

1 Introduction

Gene clusters in a genome provide substrates for genomic innovation, as gene duplication is often followed by functional diversification [1]. Also, genomic deletions associated with nearby segmental duplications cause several human genetic diseases [2]. One surprising discovery emerging from the sequencing of the human genome was the large extent of recent duplication in the human lineage. Analysis of the human genome sequence revealed that 5% consists of recent duplications [3]; subsequent studies have further found extensive copy-number variation among individuals [4].

Recently duplicated genomic segments are exceedingly difficult to sequence accurately and completely. Even the "finished" human genome sequence [5] contains about 300 gaps, many of which reflect regions harboring nearly identical tandemly duplicated segments. The situation with mammalian genomes sequenced by a whole-genome shotgun sequencing strategy [6] is typically much worse, with recently duplicated segments often grossly misassembled. The development of computational methods for analyzing gene clusters has therefore

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 29–49, 2008.

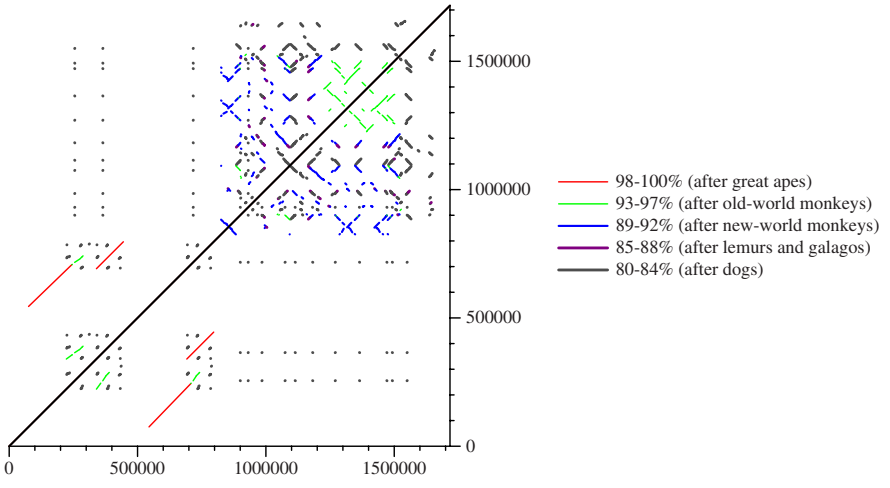

Fig. 1. Dot-plots of self-alignments of the human UGT2 cluster exceeding thresholds of percent identity chosen to roughly correspond to the divergence of the human lineage from great apes (98%), old-world monkeys (93%), new-world monkeys (89%), prosimians (85%) and dogs and other laurasiatherians (80%). We estimate that 2, 27, 51, 59, and 82 duplications respectively are needed to produce the current configuration from a duplication-free sequence (no deletions were predicted), suggesting a sustained growth of the cluster along the human lineage, with a burst of activity around the time that humans and apes diverged from old-world monkeys. The sequence alignments were computed using blastz [11] and post-processed as described in the text.

lagged far behind that for analyzing single-copy regions, due in part to the lack of accurate sequence data. Even the basic problem of formally defining what is meant by a multi-species sequence "alignment" of a region harboring a gene cluster (much less actually generating an accurate alignment of such a region) has only recently been addressed [7,8]. While the recent testing of several alignment methods with comparative sequence data representing 1% of the human genome [9] suggested adequate performance, a closer examination of the resulting alignments for those regions containing tandem gene clusters (e.g., both globin clusters) showed significant imperfections [10].

Here, we describe an algorithm for producing a theoretical ancestral sequence and a parsimonious set of duplication and deletion events explaining the observed state of a gene cluster. We start by setting a lower bound for the percent identity in self-alignments of a gene cluster (e.g., 93%; Fig.1). This defines the set of duplications that have occurred in a given time interval (such as the last 25 million years) and that have not subsequently been deleted. The ancestral configuration of each gene cluster is then deduced at several evolutionary points, and predictions are made about the parsimonious sets of duplications and deletions that converted the ancestral configuration into the extant one.

Similar problems have been studied before. Elemento *et al.* [12] and Lajoie *et al.* [13] developed algorithms for reconstruction of evolutionary histories of gene

families allowing tandem duplications and inversions. Their basic assumption is that a gene is always duplicated as a whole unit and duplicated copies are always immediately adjacent to their sources. These assumptions are routinely violated in the real data, and thus their methods have limited applicability in genome-wide studies. In addition, Elemento *et al.* do not consider inversions, while Lajoie *et al.* only consider single gene duplications. Jiang *et al.* [14] recently used methods developed for repeat identification to infer ancestral "core duplicated elements". Their results provide useful insights about duplication histories, but without detailed reconstructions. In this paper, we aim to provide event-by-event reconstructions of duplication and deletion histories using local sequence alignments, allowing both tandem and interspersed duplications (potentially with inversions).

We have applied our algorithm to 25 human gene clusters, in each case predicting the evolutionary scenarios corresponding to five major divergence points along the lineage leading to human.[1] Our results provide distributions of the predicted sizes of rearranged segments. Also, using percent-identity thresholds associated with large increases in the estimated number of duplications and deletions, we can estimate dates of rapid cluster expansion.

In future work, we plan to use such estimates to examine a large number of human gene clusters in conjunction with experimental data on gene-family size in various primates, as generated by array comparative genome hybridization (aCGH) [15,16]. Our aim is to design a larger primate comparative sequencing project that will more deeply examine the evolutionary history of a set of human gene clusters. In turn, the availability of such comparative sequence data should provide important insights about primate genome evolution and catalyze the development of computational methods for analyzing gene clusters.

2 Problem Statement and Data Preparation

Our goal is to reconstruct the evolutionary history that has generated a gene cluster in the human genome. Given the cluster's DNA sequence in a single species, we first identify all local self-alignments in both forward and reverse-complement orientations using blastz [11]. We can visualize the identified alignments using a dot-plot, and our goal is equivalent to providing a set of instructions for generating the observed dot-plot from a duplication-free sequence using a series of evolutionary events (duplications and deletions).

We preprocess the initial dot-plot to satisfy the *transitive closure property*. That is, if the dot-plot contains local alignments for region A and B, and for region B and C, then the dot-plot must also contain a local alignment for region A and C. We also *maximize each alignment*, i.e., we ensure that the alignments cannot be extended at either end. Finally, a local alignment can be broken into smaller pieces by mutations and interspersed repeats. We have developed an

[1] We have also extended this analysis to 165 biomedically interesting clusters and the results are presented in Appendix C.

accurate algorithm to determine the transitive closure of a dot-plot and to *chain alignments* together if they are broken by these events.

Since after preprocessing the alignments are maximized and have the transitive closure property, we can represent the original sequence by a sequence of *atomic segments* that are separated by boundaries of the alignment (*atomic boundaries*). We will denote the atomic segments by letters $a, b, c, \ldots$, and their reverse complements by $\overline{a}, \overline{b}, \overline{c}, \ldots$. The atomic segments that are aligned to each other will have the same letter with different subscripts (e.g., $xa_1yb_1c_1z\overline{c_2}a_2\overline{b_2}w$ has 10 atomic segments, two of which are reverse complements; a_1 and a_2 are aligned, and so are b_1 and b_2, and c_1 and c_2).

We say that the two adjacent atomic segments xy can be *collapsed* into a single atomic segment z, if y is always immediately preceded by x, and x is always immediately followed by y (we also consider $\overline{x}$ and $\overline{y}$ in the reverse orientation). In such case, we can replace all occurrences of xy with z, and all occurrences of $\overline{yx}$ with $\overline{z}$. Since initially all alignments are maximized, our initial representation will have no collapsible atomic segments.

We will be looking at sequences of duplication events in reversed order of time, i.e., starting from the latest duplication. A duplication event copies region P of the sequence (which can consists of several consecutive atomic segments) to another location (possibly with reversal). Thus, we can always identify the latest duplication by a pair of regions (P, D), where D is a region identical to P except for atomic segment subscripts and perhaps orientation (e.g., $(a_1b_1, \overline{b_2}\overline{a_4})$). If correctly identified, we can *unwind* a duplication (P, D) by removing segment D from the sequence, then collapsing all collapsible atomic segments. By unwinding all duplications, we obtain an atomic segment representation of the ancestral sequence. We are now ready to state our problem formally.

Definition 1 (Parsimonious reconstruction of duplication events). *Given a representation of the present-day DNA sequence by atomic segments, find the shortest sequence of duplication events* $(P_1, D_1), (P_2, D_2), \ldots, (P_k, D_k)$ *such that if we unwind these duplications, we obtain a sequence containing only a single atomic segment.*

3 Basic Combinatorial Algorithm

We first present a simple combinatorial algorithm that can correctly reconstruct all the duplication events (except for their order and orientation) under the following assumptions:

(1) A duplication event copies (possibly with reversal) a region of the sequence to any location except inside the originating region.
(2) The sequence evolves only by duplications (including duplications with reversal and tandem duplications). In particular, there are no deletions.
(3) No atomic boundaries are reused as duplication boundaries, except in tandem duplications. Here, boundaries of two aligned atomic segments (e.g. a_1 and a_2) are considered to be the same atomic boundary.

These assumptions are much more permissible than those of Elemento et al. [12], yet they are still often violated in the real data. Therefore, we also offer a more practical solution based on the sequential importance sampling in the next section. Note that assumption (3) is a stronger version of the commonly used no-breakpoint-reuse assumption [17] and can be justified by the usual arguments.

Definition 2 (Candidate alignments). *We call a pair of regions* (P, D) *a* candidate alignment *if* P *and* D *are identical except for subscripts and orientation, and if, after removing* D*, the atomic segment pair flanking* D *and the two pairs flanking each boundary of* P *can be collapsed.*

For example, for $xa_1yb_1c_1z\overline{c_2}a_2\overline{b_2}w$, the alignment (a_1, a_2) is a candidate alignment. This is because after removing a_2, the flanking atomic segment pair, $\overline{c_2}\overline{b_2}$ can be collapsed into a single atomic segment. Additionally, the atomic segment pairs flanking boundaries of a_1 (xa_1 and a_1y) can also be collapsed.

Lemma 1. *In a sequence of atomic segments that arose by the process satisfying the assumptions (1)-(3), the latest duplication is always among the candidate alignments.*

Lemma 1 suggests a simple and efficient *basic algorithm* for reconstructing a sequence of duplications:

1. Find a candidate alignment (P, D).
2. Output (P, D) as the latest duplication and unwind (P, D) by removing D from the sequence and collapsing all collapsible atomic segments.
3. Repeat until there is only a single atomic segment left.

Depending on the choice of candidate alignments in step 1, we can produce several duplication histories that could lead to the present-day sequence as represented by the sequence of atomic segments. Lemma 1 shows that one of those possible solutions is the real sequence of duplications. We can further show that all the other solutions produced by the basic algorithm are equally good solutions of the problem (proof relegated to Appendix A and B):

Theorem 1. *If assumptions (1)-(3) are met then the basic algorithm will always succesfully recover a sequence of duplications that will collapse the whole sequence into a single atomic segment, regardless of the order of choice of candidate alignments in step 1. Moreover, all of these solutions have the same number of events and they represent all parsimonious solutions of the duplication event reconstruction problem.*

For example, to apply the basic algoritm to $xa_1yb_1c_1z\overline{c_2}a_2\overline{b_2}w$, we note that alignment (a_1, a_2) is the only candidate alignment; $(b_1, \overline{b_2})$ and $(c_1, \overline{c_2})$ do not satisfy the definition of candidate alignment at this moment. We remove a_2 to obtain a new sequence $xa_1yb_1c_1z\overline{c_2}\overline{b_2}w$, and we remove the corresponding local alignment (a_1, a_2). We collapse the new sequence into a simpler form $ue_1z\overline{e_2}w$, where $u = xa_1y$,$e_1 = b_1c_1$, $\overline{e_2} = \overline{c_2}\overline{b_2}$. Now only one local alignment remains, which can be resolved by repeating the above procedure. Since both e_1 and $\overline{e_2}$ can be deleted, deleting either of them leads to a duplication-free sequence with different configurations.

4 Sequential Importance Sampling

The assumptions required for the basic algorithm to work are often violated in practice. In particular, large scale deletions in the gene clusters violating assumption (2) are likely to occur, and atomic boundary reuses violating assumption (3) are not uncommon. Once a boundary reuse occurs, regardless of its causes, we can no longer reconstruct the correct evolution history or even predict the true number of events. Even if assumptions (1)-(3) are satisfied, there are always multiple ways of reconstructing the history of a gene cluster. The number of the events will be the same, but the order of the events and the ancestral duplication-free sequence will be different among solutions. To make inference about the evolution history of a gene cluster, we need to summarize the feature of interest from all possible histories. However, enumerating all possible histories would be computationally expensive.

To address the atomic boundary reuse and to model deletions, we propose a stochastic algorithm that first samples many possible histories of a gene cluster from a target distribution, and then makes inference of evolutionary features from the collected samples. We use the target distribution to define the scope of histories and their relative contributions. For example, to make inference exclusively from histories that have no atomic boundary reuse, the target distribution can be uniform on all such histories and 0 otherwise. In practice, we will use more flexible target distributions to accommodate practical complications. To reconstruct a possible history from the target distribution, we use sequential importance sampling (SIS) [18]. SIS sequentially samples one event at a time from a pool of possible events until all local alignments in a dot-plot are resolved. We represent a history of the gene cluster by a series of T events $\mathcal{O}_T = (O_1, \ldots, O_T)$ reconstructed by SIS in reverse order of time. Here, both $\mathcal{O}_T$ and T are unknown. The basic algorithm is a special case in which every reconstructed event O_i corresponds to a *candidate alignment*. By repeating the SIS procedure, we obtain many possible histories. We then summarize the desired features by taking a weighted average, with weights calculated as the difference between the target distribution and the actual sampling distribution.

Given a gene cluster X, we specify the target distribution of histories to be $\pi(\mathcal{O}_T \mid X) \propto e^{aT+br}$, where T is the number of events, r is the number of reused atomic boundaries, and a, b are two penalty parameters. We chose $a = b = -5$; thus histories with fewer evolutionary events and boundary reuses will contribute more to the inference. The penalty (-5) was chosen to allow suboptimal solutions. When the penalty approaches $-\infty$, only the most parsimonious solutions with the least boundary reuse will influence the result. Note that we only need to specify the target distribution up to a normalizing constant.

Directly sampling histories from the target distribution is often intractable, and thus SIS is used. Suppose we already reconstructed t most recent events, we sample the next event O_{t+1} from a trial distribution $g_t(O_{t+1} \mid \mathcal{O}_t)$. Our goal in choosing the trial distribution is to allow easy sampling while resembling the target distribution as closely as possible. By sampling events until all alignments are resolved, we obtain a possible history $\mathcal{O}_T$, and by repeating this procedure we

collect many possible histories. However, the collected histories will not follow the target distribution $\pi(\mathcal{O}_T \mid X)$, but instead $\prod_{t=0}^{T-1} g_t(O_{t+1} \mid \mathcal{O}_t)$. To correct this bias, we calculate weight $w = \pi(\mathcal{O}_T \mid X) / \prod_{t=0}^{T-1} g_t(O_{t+1} \mid \mathcal{O}_t)$ determining how much reliance we shall put on each reconstructed history. Finally, given m histories $\mathcal{O}_{T_1}^{(1)}, \mathcal{O}_{T_2}^{(2)}, \ldots, \mathcal{O}_{T_m}^{(m)}$ and their weights $w_1, \ldots, w_m$, we make a statistical inference about evolutionary features by approximating the expectation of any function $u(\mathcal{O}_T)$ of histories as $E[u(\mathcal{O}_T)] = \left(\sum_{i=1}^{m} w_i u(\mathcal{O}_{T_i}^{(i)})\right) / (\sum_{i=1}^{m} w_i)$. For example, $u(\mathcal{O}_T) = T$ gives the number of events.

The choice of the trial distribution directly determines the efficiency of history reconstruction. For example, if assumptions (1)-(3) are met, we can let $g_t(O_{t+1} \mid \mathcal{O}_t)$ be uniform on all events O_{t+1} that involve a candidate alignment and 0 on all other events. As a result, the SIS algorithm will efficiently and precisely produce the same number of events as the basic algorithm.

We used simulations to choose a set of good trial distributions. In particular, we used $g_t(O_{t+1} \mid \mathcal{O}t) = (L-\ell)^{-k-2} f(s,\delta)/Z$ for duplication, and $g_t(O_{t+1} \mid \mathcal{O}t) = (L+\ell)^{-1} e^{-\ell/\lambda} f(s,\delta)/Z$ for deletion. For duplication $O_{t+1} = (P, D)$, $k \in \{0, 1, 2, 3\}$ denotes the number of reused atomic boundaries, i.e. the number of non-collapsible atomic segment pairs that flank D and the boundaries of P after removing D. Furthermore, L and ℓ denote the current sequence length and the duplication size, respectively. For deletion, ℓ and L denote the actual and the expected deletion size, respectively. We only consider deletions without atomic boundary reuse, and $\lambda = 10000$. Intuitively, we prefer to sample longer duplications and shorter deletions in each SIS step. We also prefer alignments with higher percent identity and those that resolve more local alignments, which is represented by function $f(s,\delta) = e^{(\delta-(100-s))/5}$ of the alignment percentage identity $s \in [0, 100]$ and the number δ of alignments resolved by O_{t+1}.

We only consider a deletion event if the atomic segment pair flanking a deletion site appears elsewhere in the sequence. Otherwise, no deletion information is available. For example, suppose $a_1 b_1$ flanks a deletion site, and we observe a_2 and b_2 elsewhere, then the region between a_2 and b_2 can be inserted in between $a_1 b_1$ to unwind a deletion. The relative orientation between a_1 and b_1 must match that between a_2 and b_2, and $a_1 b_1$ must not be located between a_2 and b_2. If all conditions are met, we calculate the percentage identity s from the flanking alignments (a_1, a_2) and (b_1, b_2), and the deletion event can be reconstructed. Finally, Z denotes the normalizing constant for the trial distribution. Compared with the normalizing constant for the target distribution, Z is much easier to calculate, because we can easily enumerate all possible events given $\mathcal{O}_t$.

5 Application to Human Gene Clusters

We have identified 457 duplicated regions in the human genome assembly hg18, based on alignments from UCSC browser self-chains [19] of length at least 500 bp, with at least 70% identity, and with both segments located within 500 Kbp of each other. The regions were defined by clustering overlapping duplications; only regions of substantial size (at least 50 Kbp) and non-trivial complexity (at least

Table 1. Estimated numbers of duplications and deletions in 25 human gene clusters following divergence from great apes (GA), old world monkeys (OWM), new world monkeys (NWM), prosimians (LG), and dogs and other laurasiatherians (DOG)

Name (possible disease association)	Location	GA	OWM	NWM	LG	DOG	gaps
PRAMEF	chr1p36.21	7	23	32	48	63	3
HIST2H (asthma; atrial fibrillation)	chr1q21.1-2	21	41	68	101	107	6
FCGR (systemic lupus erythematosus)	chr1q23.3	3	3	5	6	6	0
CFH (macular degeneration)	chr1q31.1	4	6	18	22	25	0
CCDC;CFC1 (left-right laterality defects)	chr2q21.1	3	5	12	12	15	0
UGT1A (neonatal hyperbilirubinemia)	chr2q37.1	0	2	13	17	23	0
UGT2 (prostate cancer)	chr4q13.2-3	2	27	51	59	82	1
SMA;SMN (motor neuron disease)	chr5q13.2	23	25	25	25	25	0
HIST1H;BTN (coronary heart disease)	chr6p22.2-1	0	1	9	19	35	0
HLA;TRIM (multiple sclerosis)	chr6p22.1-21.33	0	2	29	45	58	0
HLA;BAT (type 1 diabetes)	chr6p21.33	0	4	12	17	28	0
HLA-D (rheumatoid arthritis)	chr6p21.32	0	1	14	21	26	0
HLA-D;COL11A (acute lymphoblastic leukemia)	chr6p21.32	0	0	0	7	14	0
CCL;CTF2;PMS2 (rheumatoid arthritis)	chr7q11.23	21	31	38	40	45	1
IFN (cervical cancer)	chr9p21.3	0	11	15	20	41	0
SFTPA (tuberculosis)	chr10q22.3	6	7	8	10	12	1
OR5;HB;TRIM (thalassemia; sickle cell anemia)	chr11p15.4	4	6	10	10	27	0
KLR (immunological diseases)	chr12p13.2	0	1	1	2	3	0
CHRNA;KIAA (schizophrenia)	chr15q13.3-1	15	38	47	56	58	2
CYP1;DKFZ (lung cancer; macular degeneration)	chr15q24.1-3	2	14	23	26	28	0
LOC (rheumatoid arthritis)	chr16p11.2	3	6	6	6	8	0
NF1;EVI2 (intestinal neuronal dysplasia; autism)	chr17q11.2	3	9	10	10	10	0
CYP2 (lung cancer; esophageal cancer)	chr19q13.2	0	5	14	17	19	0
KIR;LILR (hepatitis C; liver cancer)	chr19q13.42	0	16	30	43	65	0
WFDC	chr20q13.12	0	0	0	1	2	0

two duplications) were retained. These regions cover ∼215 Mbp (7%) of the human genome. We targeted 165 biomedically interesting clusters (∼111 Mbp) that either overlap genes associated with a human disease (genetic association database [20]), or contain groups of similarly named genes [21].

Clusters were processed through a pipeline that included: (1) self-alignment by blastz; (2) production of subsets of the alignments roughly corresponding to duplications in the human lineage after divergence from great apes ($\geq$ 98% identity), old-world monkeys (93%), new-world monkeys (89%), lemurs (85%), and dogs (80%); (3) adjusting alignment endpoints to avoid predicting spurious tiny duplications; (4) chaining (i.e., local alignments of similar percent identity broken by small insertions/deletions or post-duplication insertion of interspersed repeats. For each of the resulting 825 combinations of gene cluster and divergence threshold, we estimated the number of duplications or deletions in the human lineage subsequent to the divergence. Selection of the results is shown in Table 1.

Table 1 reveals large differences in the evolutionary tempo among the gene clusters. For instance, the cluster of SMN genes appears to have been quiescent through almost all of primate evolution, then experienced an explosion of duplications in the last six million years. On the other hand, the cluster containing HLA-D appears to have changed little for 50 million years, while that containing UGT2 may have accumulated duplications fairly consistently throughout primate evolution, but with a surge of activity about 10-40 MYA.

We also estimated the size, spacing, and orientation of duplication events. Fig.2 shows estimated distributions of the size of the duplicated region and the

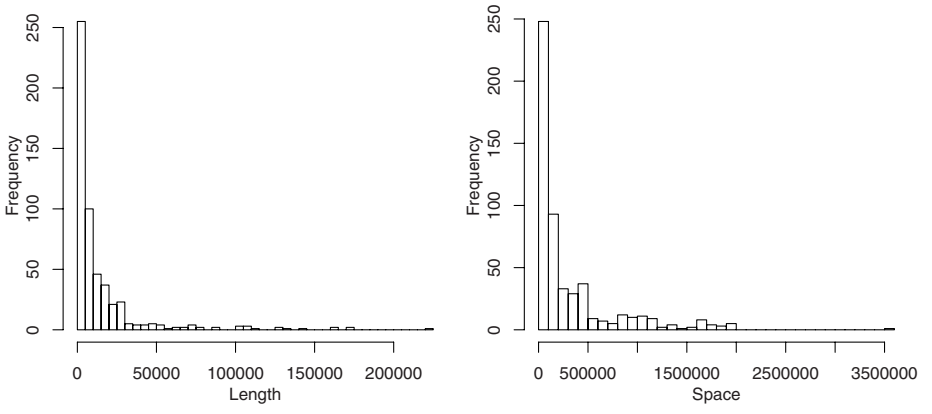

Fig. 2. Distribution of duplication lengths (left) and distances between the original and duplicate segments (right) for duplications with at least 93% sequence identity

spacing between the original and duplicated segments for duplications with at least 93% identity. For those duplication events, the copy was in the reverse orientation relative to the original segment in 39% of the cases.

We used these observed distributions and inversion rates to simulate the evolution of gene clusters, providing data to evaluate our pipeline. Starting from a 500 Kbp sequence, we simulated the formation of gene clusters via 10-100 duplications. For each event, we chose a random left end and length from the observed distribution. The procedure then chose an insertion point at a distance selected from the observed spacing distribution, and a copy of the "source" interval (or its reverse complement at a frequency of 0.39) was inserted. We also simulated deletions with frequency equal to 2% of the duplication rate (the observed frequency), using random left ends and length drawn from the empirical distribution. By simulating $N = 10, 20, 30, \ldots, 100$ events, we created 10 gene clusters for each N. The results of our pipeline were compared to the actual number of simulated events. Fig.3 shows that our algorithm accurately predicted the true number of events for the simulated gene clusters. The predicted numbers of events were slightly larger (4% on average) than the true number of events.

6 Discussion

We have designed and implemented a method to predict the duplication history of a gene cluster using sequence data from only one species. Our goal was to measure the tempo of cluster expansions throughout primate evolution for every human gene cluster, so as to help prioritize the selection of notably interesting gene clusters for more detailed comparative genomics studies. Our future plans include performing comparative sequence analysis of a series of human gene clusters, which will involve isolating and accurately sequencing the orthologous genomic regions in multiple primates.

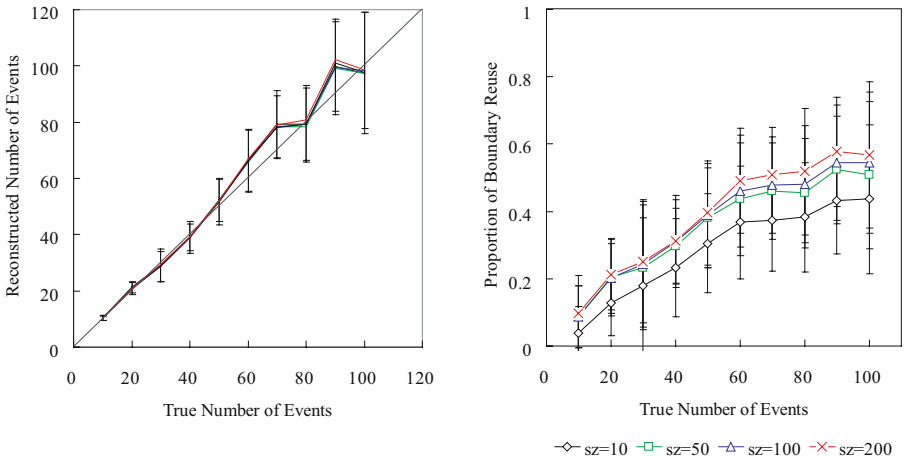

Fig. 3. Left: Actual versus reconstructed number of events with standard errors. Right: Proportion of breakpoint reuses within the reconstructed histories. For each simulated gene cluster, we used four minimum alignment length thresholds (sz): 10 bp, 50 bp, 100 bp, and 200 bp, as indicated (shorter alignments were omitted).

It will be fascinating to compare cluster dynamics in certain lineages to observed phenotypic differences among primates. For instance, Hurle et al. [22] look for correlations between differences in the WFDC cluster and several phenotypes, including female promiscuity. Note that Table 1 indicates a lack of recent WFDC expansions in the human lineage. Another potential use is illustrated by the PRAME cluster, where three gaps remain in the human assembly (Table 1). The rhesus cluster was straightforward to assemble because it lacks recent duplications [23], paving the way for evolutionary studies to help understand the cluster's function.

In addition, such sequence data should reveal differences among primate species of possible relevance for selecting species for further biomedical studies. Sequence data has already been gathered from primate orthologs of the HLA cluster, showing a large expansion in the macaque lineage [24,23], and effects of differences among the rhesus, cynomolgus, and pigtail macaque MHC clusters may be relevant for clinical studies of AIDS progression [25,26]. Similarly, the KLR cluster has been sequenced in marmoset by Averdam et al. [27] to help determine the value of that species as a primate model for immunological research. Our planned systematic project will provide a deeper understanding of primate genome evolution than would piecemeal studies of this sort.

The data should also fuel the development of computational methods for handling the complexities associated with comparative sequence data that include closely related duplicated segments. The approach described here is just one way of approaching this fascinating class of problems.

Acknowledgements. This project has been funded in part by NHGRI grant HG002238 to WM, and in part from support provided to EDG from the NHGRI

Intramural Program. Jian Ma produced a set of 141 clusters used at the start of this study, and Richard Burhans performed further analysis on those clusters.

References

1. Ohno, S.: Evolution by Gene Dupplication. Springer, Berlin (1970)
2. Lupski, J.R.: Genomic rearrangements and sporadic disease. Nat. Genet. 39(7 Suppl), 43–47 (2007)
3. Lander, E.S., et al.: Initial sequencing and analysis of the human genome. Nature 409(6822), 860–921 (2001)
4. Wong, K.K., de Leeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E., Lam, W.L.: A comprehensive analysis of common copy-number variations in the human genome. Am. J Hum. Genet. 80(1), 91–104 (2007)
5. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature 431(7011), 931–935 (2004)
6. Green, E.D.: Strategies for the systematic sequencing of complex genomes. Nat. Rev. Genet. 2(8), 573–573 (2001)
7. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., Miller, W.: Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14(4), 708–715 (2004)
8. Raphael, B., Zhi, D., Tang, H., Pevzner, P.: A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res. 14(11), 2336–2336 (2004)
9. Margulies, E.H., et al.: Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 17(6), 760–764 (2007)
10. Hou, M.: (unpublished data, 2007)
11. 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)
12. Elemento, O., Gascuel, O., Lefranc, M.P.: Reconstructing the duplication history of tandemly repeated genes. Mol. Biol. Evol. 19(3), 278–278 (2002)
13. 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)
14. Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., Eichler, E.E.: Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat. Genet. 39(11), 1361–1368 (2007)
15. Wilson, G.M., Flibotte, S., Missirlis, P.I., Marra, M.A., Jones, S., Thornton, K., Clark, A.G., Holt, R.A.: Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome Res. 16(2), 173–181 (2006)
16. 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)
17. 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)

18. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, New York (2001)
19. Kuhn, R.M., et al.: The UCSC genome browser database: update 2007. Nucleic Acids Res 35(Database issue), D668–D673 (2007)
20. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nat. Genet. 36(5), 431–432 (2004)
21. Ma, J.: personal communication (2007)
22. Hurle, B., Swanson, W., Green, E.D.: Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res. 17(3), 276–276 (2007)
23. The Rhesus Macaque Genome Sequencing and Analysis Consortium: Evolutionary and biomedical insights from the rhesus macaque genome. Science 316(5822), 222–224 (2007)
24. Daza-Vamenta, R., Glusman, G., Rowen, L., Guthrie, B., Geraghty, D.E.: Genetic divergence of the rhesus macaque major histocompatibility complex. Genome Res. 14(8), 1501–1505 (2004)
25. Krebs, K.C., Jin, Z., Rudersdorf, R., Hughes, A.L., O'Connor, D.H.: Unusually high frequency MHC class I alleles in Mauritian origin cynomolgus macaques. J Immunol. 175(8), 5230–5239 (2005)
26. Smith, M.Z., Fernandez, C.S., Chung, A., Dale, C.J., De Rose, R., Lin, J., Brooks, A.G., Krebs, K.C., Watkins, D.I., O'Connor, D.H., Davenport, M.P., Kent, S.J.: The pigtail macaque MHC class I allele Mane-A*10 presents an immundominant SIV Gag epitope: identification, tetramer development and implications of immune escape and reversion. J Med. Primatol 34(5-6), 282–283 (2005)
27. Averdam, A., Kuhl, H., Sontag, M., Becker, T., Hughes, A.L., Reinhardt, R., Walter, L.: Genomics and diversity of the common marmoset monkey NK complex. J Immunol. 178(11), 7151–7151 (2007)

A Proof of the Basic Algorithm

Proof of Theorem 1: Denote the present day sequence of atomic segments S and the series of k duplications that created this sequence $O_1, O_2, \ldots, O_k$. To prove the claim, we will first show that for any candidate alignment (P, D), sequence S can also be created by a sequence of duplications $O'_1, O'_2, \ldots, O'_k$ of the same length (also satisfying assumptions (1)-(3)), where the last duplication O'_k is (P, D). All claims of the theorem are a direct consequence of this claim, proven simply by induction on the number of duplication events.

Now consider a candidate alignment (P, D) in sequence S. If we look at the duplication history in reverse, we can show that D will be always a D-segment of some candidate alignment until one of the following happens (see Lemma 2): (A) either D is deleted by unwinding a duplication (P', D), or (B) all the P-segments matching D are unwound, and the role D-segment is in fact gained by a duplication (D, P').

In case (A), we can find a segment P'' matching D such that there exist a sequence of k duplications that will create sequence S, where (P'', D) is the latest duplication (Lemma 3). Since both (P'', D) and (P, D) are candidate alignments in S, we can replace (P'', D) with (P, D) in the last duplication and still obtain the same sequence S with k duplications.

In case (B), the role of the D-segment has been gained by a duplication $O_i = (D, P')$ at time i. Immediately after this event, (D, P') must be a candidate alignment (Lemma 1). Since (P', D) is also a candidate alignment, we can replace $O_1, \ldots, O_i$ with some sequence of duplications $O'_1, \ldots, O'_i$ such that we obtain the same intermediate atomic segment sequence at time i, where $O'_i = (P', D)$ (Lemma 4). Using the sequence of duplications $O'_1, \ldots, O'_i, O_{i+1}, \ldots, O_k$, we reduce case (B) to case (A), for which we have already proven the claim. □

We present the proofs of the following supporting lemmas in Appendix B.

Lemma 2. *If we consider duplication operations in reverse order, the D-segment of a candidate alignment will remain a D-segment of some (not necessarily the same) candidate alignment until either this D segment is removed from the sequence by unwinding a duplication (P, D), or all segments matching D are deleted, in which case the segment gains the role of D-segment by duplication (D, P).*

Lemma 3. *Let S be a sequence of atomic segments created by k duplications $O_1, \ldots, O_k$, and let $O_i = (P, D)$ for some i. If D is a D-segment of a candidate alignment in all intermediate sequences after duplication O_i, as well as in S (possibly with different P-segments, say P'), we can always find a sequence of duplications $O'_1, \ldots, O'_k$ leading to S such that $O'_k = (P', D)$.*

Lemma 4. *Let S be a sequence of atomic segments created by k duplications $O_1, \ldots, O_k$, where the last duplication is $O_k = (D, P)$. If (P, D) is also a candidate alignment, there exists a sequence of k duplications $O'_1, \ldots, O'_k$ such that the last operation is $O'_k = (P, D)$, and it creates the same sequence of atomic segments S.*

B Proofs of Supporting Lemmas

Lemma 5. *For a candidate alignment (P, D), with $D = u|a_1 \cdots b_1|v$ and $P = x|a_2 \cdots b_2|y$, the D segment will not overlap with any other alignments unless (P, D) is a forward tandem duplication.*

Proof. Without loss of generality, we assume there is a copy of $u|a_1$ in the sequence, say $u_3|a_3$. If $u_3|a_3$ lies within or outside either $|a_1 \cdots b_1|$ or $|a_2 \cdots b_2|$, it will remain in the sequence after removing D. Since $x|a_2$ is collapsible after removing D, $u_3|a_3$ must equal to $x|a_2$, which means $u = u3 = x$, but this contradicts with the maximum alignment assumption.

Alternatively, either $u_3|a_3$ or $x|a_2$ is deleted when removing D. If $u_3|a_3$ is deleted by D, it must lie on the boundary $b_1|v$ of D, i.e., either $b_1|v \equiv u_3|a_3$ or $b_1|v \equiv \overline{a_3}|\overline{u_3}$; either way we will have the atomic pair flanking D non-collapsible after removing D. On the other hand, if $x|a_2$ is deleted by D, we must have either a forward tandem duplication $u|a_1 \cdots b_1|a_2 \cdots b_2|y$ or a backward tandem duplication $\overline{v}|\overline{b_1} \cdots \overline{a_1}|a_2 \cdots b_2|y$. The latter leads to a contradiction because $u = \overline{a_2}$ means $u_3|a_3 = \overline{a_2}|a$, and hence $\overline{v}|a_2$ is not collapsible after removing D. □

Lemma 6. *D_1 of a candidate alignment (P_1, D_1) cannot lie within either P_2 or D_2 of another candidate alignment (P_2, D_2), but they can represent the same region, i.e., $D_1 \equiv D_2$.*

Proof. By Lemma 5, the statement is true if (P_1, D_1) is not a forward tandem duplication. When (P_1, D_1) is a forward tandem duplication, without loss of generality, assume (P_1, D_1) has the form $D_1|P_1 = u|a_1 \cdots b_1|a_2 \cdots b_2|y$. Suppose there is another candidate alignment (P_2, D_2), in which either P_2 or D_2 covers D_1. If D_1 completely lies within either P_2 or D_2 and shares no boundaries with them, then there is a second copy of $b_1|a_2$, say $b_3|a_3$ in the sequence. After removing D_1, we should have $u|a_2$ collapsible, which is impossible due to $b_3|a_3$. On the other hand, suppose D_1 lies within either P_2 or D_2 and they share the boundary $u|a_1$; then the same arguments apply. Instead, if D_1 shares the boundary $b_1|a_2$ with either P_2 or D_2, there are two situations:

Situation 1: P_2 covers D_1. In this case, after removing D_2, we should have $b_1|a_2$ collapsible, which is impossible due to $b_2|y$ in P_1.

Situation 2: D_2 covers D_1. In this case, we must have $D_2 = p|c1 \cdots ua_1 \cdots b_1|a_2$, in which the segment $a_1 \cdots b_1$ is D_1, and $P_2 = w|c_2 \cdots u_4 a_4 \cdots b_4|z$. After removing D_2, we have $p|a_2$ collapsible, which means $p = u$. After removing D_1, we should have $u|a_2$ collapsible, which means $(p|c_1) = (u|c_1) = (u|a_2)$, and thus $c_1 = a_2$. However, this means $w|c_2 = w|a_2$ in P_2 must also equal to $u|a_2$, and thus $w = u = p$, which contradicts with the maximum alignment assumption. □

Definition 3 (Coupling). *Two candidate alignments (P_1, D_1) and (P_2, D_2) are coupled if $P_1 \equiv D_2$ and $P_2 \equiv D_1$.*

Lemma 7. *D_1 in a candidate alignment $A = (P_1, D_1)$ cannot share boundaries with P_2 in another candidate alignment $B = (P_2, D_2)$, unless either $D_1 \equiv D_2$ or A is coupled with B.*

Proof. Let $D_1 \equiv u|a_1 \cdots b_1|v$, $P_1 \equiv x|a_2 \cdots b_2|y$, and $D_2 \equiv p|c_1 \cdots d_1|q$, $P_2 \equiv w|c_2 \cdots d_2|z$. Without loss of generality, we assume that D_1 shares boundaries with P_2. There are two situations:

Situation 1: D_1 is adjacent to P_2, in which case we have $(w|c_2) \equiv (b_1|v)$. Since $w|c_2$ is collapsible after removing D_2, we should have $b_2|y$ in P_1 equal to $b_1|v$, and thus $y = v$. However, this contradicts with the maximum alignment assumption. The exception is that either $b_1|c_2$ or $b_2|y$ is deleted when removing D_2. The former indicates $D_1 \equiv D_2$ by Lemma 6. For the latter, if $b_2|y$ is completely removed by D_2, there is another copy of $b_2|y$ in P_2, which still indicates $y = v$ and leads to a contradiction. If D_2 only removes b_2 in $b_2|y$, then D_2 covers P_1 by Lemma 6. In this case, we have either of the following:

1. D_2 and P_1 are in the same orientation:
   In this case, $d_1 = b_2$ and $q = y$. Since $b_2|y$ is collapsible after removing D_1, and $b_2|y = d_1|q$, we must have $d_2|z$ in P_2 equal to $d_1|q$, which contradicts with the maximum alignment assumption. The only exception is that $b_2|y$ is deleted when removing D_1. In this case, (P_1, D_1) is either coupled with (P_2, D_2), or is a forward tandem repeat in the form $P_1|D_1$. The latter is impossible, otherwise after removing D_1, we should have $b_2|c_2$ collapsible, so $b_2|c_2 = p|c_1$, which contradicts with the maximum alignment assumption.
2. D_2 and P_1 are in different orientations:
   In this case, $p \equiv \overline{y}$ and $\overline{b_2} \equiv c_1 = c_2$. However, it indicates that $b_1|c_2 = b_2|\overline{b_2}$ at the boundary of $D_1|P_2$ is not collapsible after removing D_2, and thus (P_2, D_2) is not a candidate alignment. The only exception is when b_1 of $b_1|c_2$ at the boundary of $D_1|P_2$ is deleted when removing D_2, which is impossible due to Lemma 6.

Situation 2: D_1 covers P_2. After removing D_2, $(d_2|z) \equiv (b_1|v)$ in P_2 is collapsible. However, this contradicts with $v \neq y$, unless either $b_1|v$ in P_2 or $b_2|y$ in P_1 is deleted when removing D_2.

1. if $b_1|v$ in P_2 is deleted, then we either have a forward tandem repeat $P_2|D_2$, or a reverse tandem repeat $P_2|\overline{D_2}$. For the former, we must have $u = w$ and $a = c$ following similar arguments as in Lemma 6. As a result, when removing D_2, $w|c_2$ is collapsible and thus $x = w = u$, which contradicts with the maximum alignment assumption. The only exception is when (P_1, D_1) and (P_2, D_2) are coupled. For the latter, we have a reverse tandem repeat $P_2|\overline{D_2}$. Similarly, we can show that $y = \overline{p} = \overline{u}$ and $d = \overline{c}$. Therefore, $w|c$ in P_2 equals to $w|\overline{d}$, and will remain intact after removing D_2. However, after removing D_2, we should have $d|\overline{p}$ collapsible, and thus $w = p$, which contradicts with the maximum alignment assumption unless (P_1, D_1) and (P_2, D_2) are coupled.
2. if $b_2|y$ in P_1 is deleted, then first, $b_2|y$ cannot be completely deleted by D_2, otherwise there is another copy of $b_2|y$ remaining in P_2, and the same arguments that $v \neq y$ can be applied to show a contradiction; second, the y of $b_2|y$ cannot be deleted by D_2 as proved in Situation 1; third, if the b_2 of $b_2|y$ in P_1 is removed by D_2, we have $D_2 \supset P_1$, which leads to coupling because $D_1 \supset P_2$. □

Lemma 8. *Given two candidate alignments (P_1, D_1) and (P_2, D_2), if at least one of them is not a forward tandem repeat, then D_1 will neither overlap with nor be adjacent to D_2. D_1 and D_2 can be coupled (i.e., $D_1 \equiv P_2$ and $D_2 \equiv P_1$), separated or representing the same region.*

Proof. Let $D_1 \equiv u|a_1 \cdots b_1|v$, $P_1 \equiv x|a_2 \cdots b_2|y$, and $D_2 \equiv p|c_1 \cdots d_1|q$, $P_2 \equiv w|c_2 \cdots d_2|z$. By Lemma 5 and Lemma 6, D_1 cannot overlap with, cover, or lie within D_2, unless both alignments are forward tandem repeats or if $D_1 \equiv D_2$. As a result, we only need to show that D_1 and D_2 are not adjacent to each other unless they are coupled. Without loss of generality, assume D_1 and D_2 are adjacent in the form $D_1|D_2 = u|a_1 \cdots b_1|c_1 \cdots d_1|q$.

Situation 1: $w|c_2$ in P_2 remains intact after removing D_1. After removing D_1, $u|v = u|c_1$ should be collapsible, and thus $u = w$. On the other hand, $w|c_2$ in P_2 is collapsible after removing D_2 and $u|a_1$ will remain intact, so we have $(u|a_1) = (w|a_1) = (w|c_2)$, which contradicts with Lemma 5. The only exception is that $w|c_2$ in P_2 is deleted when removing D_2, which indicates either (P_2, D_2) is coupled with (P_1, D_1), or (P_2, D_2) is a forward tandem repeat in the form $D_2|P_2$. The latter is impossible, because $q = c_1$, and after removing D_1, we have $u|c_1$ collapsible (because D_1 is adjacent to D_2), which means $u = d$ and thus $z = c_1 = q$, in which case (P_2, D_2) is not maximized.

Situation 2: $w|c_2$ in P_2 is completely deleted when removing D_1. In this case, we must have a copy of $w|c_2$ in P_1, and thus the same arguments for Situation 1 apply.

Situation 3: $w|c_2$ in P_2 is partially deleted when removing D_1, i.e., either w or c_2 is removed. In this case, P_2 must share boundaries with D_1, which is impossible due to Lemma 7, except for the coupling relationship or when $D_1 \equiv D_2$. □

Lemma 9. *A candidate alignment (P_1, D_1) cannot be partially deleted or extended when removing another candidate alignment (P_2, D_2). Instead, either P_1 or D_1 can be completely deleted by D_2. If P_1 is deleted by D_2, then there is a third candidate alignment (P_3, D_1). If D_1 is deleted by D_2, then $D_1 \equiv D_2$.*

Proof. Let $A \equiv (P_1, D_1)$ and $B \equiv (P_2, D_2)$ denote the two candidate alignments. By Lemma 8, D_1 and D_2 may be identical, coupled, or separated. The exception is when both A and B are forward tandem repeats, in which case the statement holds true. If $D_1 \equiv D_2$, removing D_2 will completely delete D_1. If D_1 and D_2 are coupled, removing D_2 will completely delete P_1. If D_1 and D_2 are separated, deleting D_2 will only affect (P_1, D_1) if D_2 strictly covers P_1. This is because neither D_2 overlaps with P_1 nor D_2 lies within but share boundaries with P_1, according to Lemma 6, and by Lemma 7, D_2 cannot be adjacent to P_1. Assume D_1 and D_2 are separated, and let $D_1 \equiv u|a_1 \cdots b_1|v$, $P_1 \equiv x|a_2 \cdots b_2|y$, and $D_2 \equiv p|c_1 \cdots d_1|q$, $P_2 \equiv w|c_2 \cdots d_2|z$. Since P_1 is strictly within D_2, we must have a copy of P_1, denoted by $P_3 \equiv x_3|a_3 \cdots b_3|y_3$ in P_2, which will remain intact after deleting D_2. As a result, the third alignment $C = (P_3, D_1)$ must be a candidate alignment. □

Using Lemma 5-9, we are now ready to prove Lemma 2-4 in Appendix A.

Proof of Lemma 2: By Lemma 9, a candidate alignment (P_1, D_1) cannot be partially removed or extended when removing other candidate alignments. We thus only need to show that, when reconstructing duplication in the reverse order, D_1 will continue to be the D segment of some candidate alignments until either D_1 is deleted or all segments matching with D_1 are deleted.

Let $D_1 \equiv u|a_1 \cdots b_1|v$ and $P_1 \equiv x|a_2 \cdots b_2|y$. Assume D_1 becomes an invalid D segment after removing a candidate alignment (P_2, D_2). If removing D_2 deletes P_1, then there is a third candidate alignment (P_3, D_1). If both P_1 and D_1 remain intact after removing D_2, then by Lemma 7 and Lemma 8, the

flanking segments of P_1 and D_1 will remain intact as well. Let $D_2 \equiv p|c_1 \cdots d_1|q$ and $P_2 \equiv w|c_2 \cdots d_2|z$, removing D_2 will produce a new atomic pair $p|q$. To invalidate the D-segment role of D_1, at least one of $x|a_2$, $b_2|y$, $u|v$ pairs must become non-collapsible due to $p|q$. If $u|v$ is affected, without loss of generality, we assume $p = u$. Since $u|v$ is collapsible after removing D_1, $p|c_1$ in D_2 must equal to $u|v$ and thus $c_1 = v$. As a result, $w|c_2 = w|v$ in P_2 must equal to $u|v$, indicating $p = w = u$. This contradicts with the maximum alignment assumption. The only exception is when P_2 and D_2 are adjacent in the form $\overline{P_2}|D_2 \equiv \overline{z}|\overline{d_2} \cdots \overline{c_2}|c_1 \cdots d_1|q$, and thus $p = u = \overline{c_2}$. However, since and $v = c_1$, we have $u|v = \overline{c_2}|c_1$ non-collapsible. Similar arguments can be applied to show contradictions when either $x|a$ or $b|y$ becomes non-collapsible due to $p|q$. In conclusion, D_1 will always be the D segment of some candidate alignment until either D_1 is deleted or all segments matching with D_1 are deleted. □

Proof of Lemma 3: We will prove this lemma by induction on the number of duplication events. First, the lemma holds trivially for the sequences with a single duplication (which must be (P, D)). Now, let us assume that the lemma holds for all duplication sequence of length less than k. We want to prove that it also holds for a sequence of duplication $O_1, \ldots, O_k$ of length k.

If $O_k = (P, D)$, then lemma holds trivially. Therefore, assume that $O_k \neq (P, D)$, and thus (P, D) is among one of $O_1, \ldots, O_{k-1}$. Let S_{k-1} be the atomic segment sequence created by $O_1, \ldots, O_{k-1}$, then according to the induction hypothesis, there exists a segment P' and a sequence of duplication $O'_1, \ldots, O'_{k-1} = (P', D)$ that also creates S_{k-1}.

Let S be the sequence created by the sequence of duplication $O'_1, \ldots, O'_{k-1}, O_k$, i.e., converted from S_{k-1} via one additional duplication O_k. Suppose that $O_k = (P_1, D_1)$, then $D_1 \neq D$ and $P_1 \neq P'$ under the no atomic boundary reuse assumption. Since D is a D-segment in S under the Lemma assumption, we can always find two alternative events $O''_{k-1} = (P'_1, D_1)$ and $O''_k = (P'', D)$ to replace $O'_{k-1} = (P', D)$ and $O_k = (P_1, D_1)$ (i.e., to switch orders of deleting D and D_1), such that S can also be created by the sequence of duplication $O'_1, \ldots, O''_{k-1}, O''_k$. This is a direct result of Lemma 9 and the fact that $D_1 \neq D$. Therefore, S can be created by k duplications with the last operation being (P'', D), even if D is generated by duplication $i(< k)$ in the real history. □

Proof of Lemma 4: Let $P \equiv x|a \ldots b|y$ and $D \equiv p|a \ldots b|q$. If both (P, D) and (D, P) are candidate alignments in S, then by Lemma 5, no other alignments will cover either P or D unless (P, D) is a forward tandem repeat. If (P, D) is not a forward tandem repeat, $(x|a)$, $(b|y)$, $(p|a)$, $(b|q)$ must all be unique pairs in the atomic segment sequence S. In addition, we should have $x|a$ collapsible after removing D, and thus x must be unique in S. Similar arguments can show that y, p, and q are also unique in S. As a result, the two segments P and D are bounded within unique atomic segments and thus forms "two islands". So any previous duplication related with P or D segments must be completely inside of either P or D, and they do not share boundaries with P or D. The same conclusion

applies even if P and D are adjacent to each other. Therefore, to change the latest duplication from $O_k = (D, P)$ to $O'_k = (P, D)$, we simply "redirect" all the duplications that are inside of D to be inside of P, and keep the rest the same. This will create a new sequence of duplication $O'_1, \ldots, O'_{k-1}, O'_k = (P, D)$ that creates S. □

C Duplication Complexity of Selected Gene Clusters

Name	Location	GA	OWM	NWM	LG	DOG	gaps
PRAMEF	chr1:12750851-13626366	7	23	32	48	63	3
PADI	chr1:17423413-17600526	0	0	0	0	0	0
	chr1:22775285-23112635	0	0	0	0	0	0
	chr1:25443774-25537798	0	1	1	1	1	0
CYP4	chr1:47048227-47411959	1	5	5	6	11	0
	chr1:86662627-86892926	0	0	1	1	1	0
GBP	chr1:89244904-89692274	0	5	7	9	22	0
AMY	chr1:103898363-104119006	4	10	14	14	14	0
	chr1:110861483-111018698	0	0	0	0	0	0
	chr1:119739258-119963386	0	0	3	19	20	0
HIST2H	chr1:144651745-148125604	21	41	68	101	107	0
	chr1:150451947-150599304	0	1	1	1	1	0
LCE	chr1:150776235-151067237	0	0	6	7	11	0
SPRR	chr1:151220060-151272246	0	0	0	0	1	0
SPRR	chr1:151278447-151390171	0	0	1	7	8	0
	chr1:153784948-154023311	0	5	14	24	28	0
FCRL	chr1:155406878-156042315	0	2	11	30	40	0
CD1	chr1:156417524-156593228	0	0	0	0	1	0
OR	chr1:156634961-157053841	0	0	0	0	1	0
	chr1:157512882-157835664	0	0	0	0	1	0
FC	chr1:159742726-159915333	3	3	5	6	6	0
	chr1:167848867-167968738	0	0	0	0	0	0
CFH	chr1:194914679-195244603	4	6	18	22	25	0
	chr1:205701588-205958677	1	7	12	12	13	0
ZNF	chr1:245215980-245486993	2	2	2	2	5	0
OR	chr1:245680906-246912147	1	6	23	48	55	0
	chr2:79106193-79240545	0	0	0	1	1	0
CCDC; CFC1	chr2:130461934-131153411	3	5	12	12	15	0
	chr2:166554904-167039157	0	0	0	1	3	0
	chr2:208680310-208736768	0	1	2	2	2	0
	chr2:232893923-233063157	0	3	13	21	24	0
UGT1A	chr2:234140385-234334547	0	2	13	17	23	0
	chr3:38566866-38926662	0	0	0	0	1	0
ZNF	chr3:44463068-44751808	0	1	1	2	2	0
CCR	chr3:45917359-46425558	0	0	0	0	1	0
	chr3:48977485-49396481	0	0	1	1	1	0
OR5	chr3:99254906-99898694	0	1	10	14	27	0
	chr3:134863859-134969704	0	0	0	1	1	0

Name	Location	GA	OWM	NWM	LG	DOG	gaps
	chr3:152413859-152539276	0	0	0	0	0	0
	chr3:196822567-196963470	1	1	1	1	1	1
	chr4:38451248-38507567	0	0	0	0	0	0
UGT2	chr4:68830737-70547917	2	27	51	59	82	1
CXCL	chr4:74781081-75209572	0	0	0	3	26	0
ADH	chr4:100215375-100612366	0	0	3	8	10	0
SMN	chr5:68787010-70696078	23	25	25	25	25	0
PCDH	chr5:140145736-140851366	0	0	0	1	37	0
	chr6:10322043-10743230	0	1	1	1	1	0
HIST1H; BTN	chr6:25833812-26617296	0	1	9	19	35	0
HIST1H	chr6:27561049-27970197	1	1	3	4	11	0
ZNF; OR	chr6:28161149-29664934	0	0	10	19	33	0
TRIM	chr6:29786467-30568761	0	2	29	44	58	0
BAT	chr6:31267292-31607879	0	4	12	17	28	0
HLA-D	chr6:32514542-32891079	0	1	14	21	26	0
HLA-D	chr6:33082752-33265289	0	0	0	7	14	0
GSTA	chr6:52711832-52960243	0	7	13	27	33	0
TAAR	chr6:132951558-133008844	0	0	0	0	0	0
	chr6:160794897-161275095	0	0	9	17	18	0
	chr6:169347092-169825478	0	0	0	0	0	0
LOC	chr7:71966977-72466918	1	5	8	8	8	0
CCL; CTF2; PMS2	chr7:73565093-76526339	21	31	38	40	45	1
	chr7:86869277-87034269	0	0	0	1	1	0
	chr7:98915207-99500181	0	0	10	20	26	0
	chr7:142134143-142186482	0	1	4	4	4	0
	chr7:142469761-142919050	0	0	0	0	0	0
OR	chr7:143005241-143760083	7	9	9	11	17	0
ZNF	chr7:148389924-149094267	0	4	9	15	17	0
GIMAP	chr7:149794678-150079280	0	4	4	5	5	0
DEF	chr8:6769157-6902786	1	1	1	1	13	0
DEFB10; DEFB	chr8:7069563-7953918	5	8	8	10	14	1
	chr8:22933046-23139154	0	0	6	21	30	0
	chr8:82518183-82604430	0	0	0	0	0	0
ZNF; ZNF	chr8:145901725-146244938	0	0	0	0	2	0
IFN	chr9:21048760-21471698	0	11	15	20	41	0
OR13	chr9:106305453-106535416	0	0	2	2	3	0
OR	chr9:124279100-124603579	0	0	1	1	2	0
	chr9:134962296-135122729	0	0	0	0	0	0
AKR1C	chr10:4907977-5322660	0	5	7	13	32	0
	chr10:26458036-27007198	0	4	7	17	19	0
	chr10:53701853-54315804	0	0	0	1	1	0
SFTPA	chr10:80936018-81672884	6	7	8	10	12	1
	chr10:88319645-89246594	2	2	3	3	3	0
IFIT	chr10:91051661-91168336	0	0	0	0	1	0
	chr10:96426730-96897127	1	2	18	18	20	0
	chr10:118205218-118387999	0	0	1	3	7	0
	chr10:135086124-135244057	2	2	2	2	2	0
	chr11:1065614-1239359	0	0	0	0	0	1
OR5; HB; TRIM	chr11:4124149-6177952	4	6	10	10	27	0

Name	Location	GA	OWM	NWM	LG	DOG	gaps
OR	chr11:6745853-6899767	0	0	1	1	2	0
	chr11:24900251-25670383	0	0	0	0	1	0
OR4	chr11:48193633-48622537	0	0	7	17	20	0
	chr11:48865105-49870196	1	12	15	15	18	0
OR	chr11:54833085-56562513	0	1	14	46	61	0
OR	chr11:57390332-58032285	0	1	1	1	2	0
OR	chr11:58833693-59274730	0	0	2	2	6	0
	chr11:66900400-67551984	0	2	4	4	4	0
MMP	chr11:102067847-102343167	0	0	0	0	0	0
OR	chr11:123129479-123988274	0	3	5	7	15	0
	chr12:9099391-9319709	0	0	0	0	0	0
KLR	chr12:10446112-10497748	0	1	1	2	3	0
TAS2R	chr12:10845284-11475585	0	6	26	36	64	0
	chr12:20846959-21313050	0	0	0	11	24	0
KRT	chr12:50852169-51586146	0	2	4	8	15	0
OR	chr12:53795147-54317866	0	0	1	2	2	0
	chr12:55040623-55490902	0	0	0	0	2	0
	chr12:111828405-111931464	0	0	0	0	0	0
ZNF; ZNF	chr12:132011584-132289534	0	0	0	0	0	0
	chr13:19614743-19695656	0	0	0	0	0	0
	chr13:51634776-51849914	0	1	1	2	2	0
OR	chr14:19250951-19781765	0	0	0	1	3	0
RNASE	chr14:20319257-20525050	0	3	6	8	8	0
	chr14:20692977-21208956	1	1	1	2	3	0
C14orf	chr14:23177922-23591420	1	5	8	9	11	0
	chr14:24044573-24173288	0	0	0	0	0	0
C14orf	chr14:73073807-73175062	0	1	1	3	3	0
SERPINA	chr14:93850088-94034351	0	0	1	1	1	0
SERPINA	chr14:94099676-94182828	0	0	0	0	0	0
	chr14:105101878-105397048	2	17	20	20	21	0
CHRNA; KIAA	chr15:26168691-30570226	15	38	47	56	58	2
CYP1; DKFZ	chr15:71687352-74071019	2	14	23	26	28	0
	chr16:1211147-1279180	0	2	2	2	2	0
ZNF	chr16:3105811-3428601	0	0	0	0	4	0
	chr16:20234773-20711192	2	6	6	6	7	0
LOC	chr16:28560127-29404514	3	6	6	6	8	0
MT	chr16:55181257-55275655	0	0	0	4	18	0
	chr16:85101437-85170740	0	0	0	0	0	0
	chr16:88526416-88690103	0	0	0	0	1	0
OR	chr17:2912380-3289105	1	3	4	5	10	0
	chr17:6501152-6854467	0	1	1	1	1	0
MYH	chr17:10145620-10499991	1	2	7	11	25	0
	chr17:22979762-23370074	0	2	4	4	5	0
NF1; EVI2	chr17:25940349-27337990	3	9	10	10	10	0
CCL	chr17:29605831-29711075	0	0	0	0	0	0
CCL	chr17:31334805-31886998	4	7	7	8	9	1
KRT	chr17:36069761-37038364	0	9	13	20	30	0
	chr17:59292402-59355509	0	4	5	5	5	0
ABCA	chr17:64375713-64805977	0	1	1	1	3	0

Name	Location	GA	OWM	NWM	LG	DOG	gaps
CD300	chr17:70033428-70220651	0	0	0	0	2	0
DS	chr18:26828138-26991601	0	0	0	0	0	0
DS	chr18:27160523-27356213	0	0	0	0	0	0
	chr18:41459658-41573640	0	0	0	0	0	0
SERPINB	chr18:59406881-59805500	0	1	2	2	3	0
	chr19:230508-1050902	0	0	0	0	1	0
	chr19:6377406-7037708	1	4	5	6	8	0
ZNF; OR	chr19:8569586-9765797	3	5	15	24	34	1
OR	chr19:14771021-15113863	0	0	0	3	11	0
CYP4F	chr19:15508827-15669145	0	0	0	1	9	0
CYP4F; OR10H	chr19:15699700-15970865	0	1	2	7	26	0
	chr19:39695418-40633289	0	2	13	20	25	0
ZNF	chr19:40976726-43450858	0	7	13	18	27	0
CYP2	chr19:46016475-46404199	0	5	14	17	19	0
ZNF	chr19:49031476-49676451	0	1	5	9	33	0
	chr19:49840790-50069615	0	0	0	0	0	0
	chr19:55457577-55842758	0	0	0	2	4	0
KLK	chr19:56014236-56276734	0	0	1	1	3	0
KIR; LILR	chr19:59404199-60117280	0	16	30	43	65	0
CST	chr20:23560786-23885538	0	12	19	26	35	0
C20orf	chr20:31084573-31402526	0	1	1	1	1	0
WFDC	chr20:43531807-43853954	0	0	0	1	2	0
	chr20:44190604-44564928	0	0	0	0	0	0
KRTAP	chr21:30642250-30735038	0	0	0	1	2	0
KRTAP	chr21:30774233-30910843	0	0	0	1	1	0
KRTAP1	chr21:44783567-44947268	0	0	4	9	15	0
	chr22:18594272-19312230	3	4	6	6	6	1
	chr22:20705392-23410020	3	26	52	74	118	0
	chr22:30379202-31096691	0	4	5	7	8	0
APOBEC3	chr22:37674922-37828933	0	4	12	19	26	0