IDENTIFYING PARENT-DAUGHTER RELATIONSHIPS AMONG DUPLICATED GENES[1]

MIRA V. HAN AND MATTHEW W. HAHN

Department of Biology and School of Informatics, Indiana University,
Bloomington, IN 47405 USA

In this paper we use the length of the shared synteny between genes to identify "parent" orthologs among multiple lineage specific duplicated genes. Genes in the region around each duplicated paralog are compared with the genes flanking an outgroup ortholog to estimate the probability of observing homologs in syntenic vs. non-syntenic regions. The length of the shared synteny is introduced as a hidden variable and is estimated using Expectation-Maximization for each lineage specific paralog. Assuming that the original, parental gene will preserve the longest synteny with the outgroup gene, and that any daughter genes will have a shorter syntenic block, we are able to determine parent-daughter relationships. We apply this method to lineage specific duplications in the human genome, and show that we are able to determine the direction and size of the duplication events that have created hundreds of genes.

1. Introduction

Gene families are groups of genes with high sequence similarity that are derived from a common ancestor. The relationships between members of a gene family can be classified into orthology and paralogy based on their evolutionary history. Orthologs are pairs of genes that are diverged from a common ancestor by speciation, while paralogs are pairs related through duplication events [1]. Paralogs can further be divided into out-paralogs and in-paralogs [2], generally defined as relationships between paralogs either between or within species, respectively (Fig. 1a). In-paralogs are considered to be co-orthologous to a single-copy gene in an outgroup, as they are all related to this gene by a speciation event (Fig. 1a).

We can define yet another relationship based on the direction of the duplication event. We call the original copy the parent (or "primary" ortholog) and any derived copy the daughter (or "secondary" ortholog). Ortholog-paralog relationships are different from parent-daughter relationships, and identifying orthologs does not necessarily determine the direction of the duplication event. The parent-daughter relationship can be determined unambiguously only when

[1] This work is supported by grant DBI-0543586 from the U.S. National Science Foundation.

you have a gene from an outgroup species that can be used to identify the ancestral position of the gene, assuming that the outgroup gene has maintained synteny (Fig. 1b). Identifying the parent-daughter relationships among lineage specific duplicates (i.e. in-paralogs) is the problem addressed in this paper. Lineage specific duplications are interesting because they are a potential source of lineage specific phenotypes [3]. Finding the parent-daughter relationships among lineage specific duplicates can provide information on the evolutionary forces governing the origin and maintenance of gene duplicates (e.g. [4])

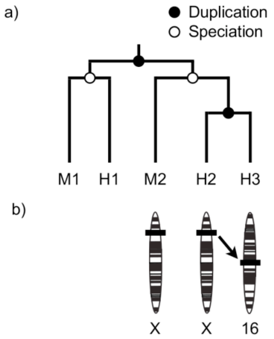

Figure 1. a) Examples of multiple types of homologous relationships. Genes M1 and H1 are orthologs, while H2 and H3 are in-paralogs of each other and are both co-orthologs of gene M2. Genes H1 and M2 are out-paralogs. b) Determining parent-daughter relationships. Genes M2 and H2 both reside on the X chromosome of their respective species, while gene H3 is found on chromosome 16. Assuming that gene H2 is in the ancestral location, H2 is the parental paralog ("primary" ortholog to M2) and H3 is the daughter paralog ("secondary" ortholog to M2).

Many computational methods have been introduced for distinguishing orthologs from out-paralogs based on synteny [5, 6], but none have addressed the problem of distinguishing ancestral and derived loci until the recent study by Jiang et al. [7]. The conceptual approach used in the current paper is similar to that introduced in this previous study [7], that is we assume that the parental duplicate shares greater synteny with the outgroup gene. The primary concern is thus how to identify the blocks of shared synteny, which is also a well studied problem [8, 9]. Many methods have also been developed to accommodate rearrangements within blocks, the general idea being to cluster or chain the pair-wise alignments by allowing a certain level of gaps between the alignments [10, 11].

In this paper we use a novel probabilistic approach to defining synteny. Instead of chaining nucleotide alignments by defining appropriate gap penalties, we allow randomness in the sharing of gene content with a probability that is estimated from the data. This approach therefore takes into account duplications, deletions, and micro-rearrangements within the syntenic block, and obviates the need to *a priori* determine gap penalties. We define pairs of genes as the experimental unit, and present a simple model based on the presence/absence of common genes in the flanking regions to estimate the lengths of duplicated blocks. The probabilities of observing genes in common in homologous vs. non-homologous regions are estimated as parameters using maximum likelihood, and the lengths of the homologous regions are introduced as hidden variables. By comparing the lengths of shared synteny, this method allows us to identify parent-daughter relationships among paralogs.

2. Methods

2.1. *Model*

We consider M families with lineage specific expansions in a species S (in this case, human). Each family m has g_m number of genes from species S and o_m number of genes from the outgroup species S_{out} (in this case, macaque). The g_m genes are restricted to duplicates whose distances between any pair are less than twice the time since speciation with the outgroup. This restriction guarantees that the duplicated genes truly are lineage specific. The outgroup genes are found by traversing the gene family tree to the closest speciation node above the lineage specific duplication and collecting all genes from the outgroup species under that node. Sometimes there is more than one outgroup gene due to lineage specific expansion within the outgroup lineage. We can therefore make $g_m \cdot o_m$ pairs between species S and S_{out} for each family. Each pair is the unit of experiment. The final dataset has a total of N experimental units consisting of pairs of genes from human and macaque where,

$$N = \sum_{m}^{M} g_m \cdot o_m$$

Note that in the end we will assign each gene to only one category (parent or daughter) by a simple rule that if any pair containing g_i is designated as a parent, then the gene g_i is a parent. For each gene we collected the gene order data that spans ±L megabases (MB) from each gene. Gene order is a series of numbers encoding the genes flanking our gene of interest. Each gene order has a different number of genes depending on the gene density of the region, and the total length of the contig available in the region. We assume that each flanking gene has been assigned to a gene family through some alignment and clustering procedure. The gene order is then encoded by family IDs, so that the same numbers correspond to homologous genes in the sequence (an example is shown in Fig. 2). This numbering scheme precludes the need for one-to-one ortholog assignment of flanking genes, and allows for duplication and rearrangement of homologous flanking genes. Each pair of focal genes has a corresponding pair of gene order data, and we will use the comparison of gene orders to extract features of synteny.

2.2. *Formulation*

To find the in-paralog that shares greater synteny with the outgroup gene, we must first mathematically define shared synteny. In this paper we define shared synteny as the higher probability of co-occurrence of homologous genes in a continuous genomic region of each species. There are two variables in this definition, the "probability of co-occurrence" and the "length of the region harboring the higher probability." In our simple model, we assume that the probability of co-occurrence of a homologous gene between two genomes is different between regions of shared synteny (p_{syn}) and regions outside shared syteny (p_{nonsyn}), but is equal within each class for all pairs. By definition p_{syn} is larger than p_{nonsyn}. The actual value will be estimated as the maximum-likelihood estimate using all pairs of gene order comparisons as data. Note that co-occurrence means homologs occur in both species within the flanking region, but does not consider the order of the genes. The length of the syntenic region has larger variance depending on the type and size of duplications and rearrangements that happen in a genome, and can be quite different from event to event. It is therefore inappropriate to model the length as one random variable for the whole dataset. We will introduce two variables l_len_i and r_len_i for each pair i $(i = 1..N)$. l_len_i denotes the length of the syntenic region to the left of gene g_i, r_len_i denotes the length to the right.

If we assume that the co-occurrence of a homologous gene in the region follows a Bernoulli distribution with p as the probability of co-occurrence, then the probability of observing a certain gene order of species S for pair x_i can be expressed as,

$$P(x_i \mid \theta) = \prod_{p\_left<j<p\_right} p_{syn}{}^{m_j}(1-p_{syn})^{1-m_j} \cdot \prod_{j<p\_left \vee j>p\_right} p_{nonsyn}{}^{m_j}(1-p_{nonsyn})^{1-m_j},\ j = 1..\ G_i$$

$\theta :\ p_{syn}, p_{nonsyn}, l\_len_i, r\_len_i$

p_{syn} : probability of co-occurrence of homologs in a syntenic region

p_{nonsyn} : probability of co-occurrence of homologs in a non-syntenic region

l_len_i : length of the syntenic region to the left of the focal gene g_i in S

r_len_i : length of the syntenic region to the right of the focal gene g_i in S

$$m_j :\ \begin{cases} 1 & \text{if gene } j \text{ in species } S \text{ occurs in gene order of species } S_{out} \\ 0 & \text{otherwise} \end{cases}$$

G_i : length of the total gene order of S

$pos(x_i)$: index of gene g_i

$p\_left = pos(x_i) - l\_len_i$: index of the left border of the syntenic region

$p\_right = pos(x_i) + r\_len_i$: index of the right border of the syntenic region

Note that there is no explicit variable denoting the length of the synteny for the outgroup genome. This is to limit the dimension of the search space. When looking for matches we will consider the whole gene order of S_{out}. But since a match is a pairwise observation, the length variables for species S implicitly determines the length of the synteny in the outgroup S_{out}.

The log-likelihood of the whole data X can be expressed as the sum of the likelihood of all x_i's.

$$\ln L = \sum_{i=1}^{N} \left(s_i \ln p_{syn} + (l_i - s_i)\ln(1-p_{syn}) + r_i \ln p_{nonsyn} + (G_i - l_i - r_i)\ln(1-p_{nonsyn}) \right)$$

$l_i = l\_len_i + r\_len_i + 1$: total length of the syntenic region of gene order x_i

s_i : total number of hits in the syntenic region of gene order x_i

r_i : total number of hits in the non-syntenic region of gene order x_i

The maximum likelihood estimate of p_{syn} and p_{nonsyn} are,

$$\hat{p}_{syn} = \underset{p_{syn}}{\operatorname{argmax}} \left(\ln p_{syn} \sum_{i=1}^{N} s_i + \ln(1-p_{syn}) \sum_{i=1}^{N} (l_i - s_i) \right)$$

$$\hat{p}_{nonsyn} = \underset{p_{nonsyn}}{\operatorname{argmax}} \left(\ln p_{nonsyn} \sum_{i=1}^{N} r_i + \ln(1-p_{nonsyn}) \sum_{i=1}^{N} (G_i - l_i - r_i) \right)$$

Implicit in the MLE estimate are the hidden variables l_len_i and r_len_i for each x_i. We use Expectation-Maximization to calculate the likelihood according to the inferred expectation of each hidden variable. The software that implements the model, PRIUS (Parental Relationship Inference Using Synteny) can be found at http://www.bio.indiana.edu/~hahnlab/Software.html.

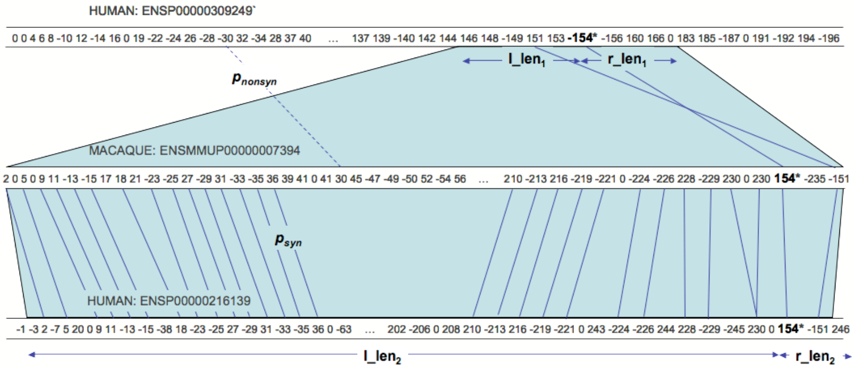

Figure 2. Description of the model. Two lineage specific human paralogs, denoted in bold, are paired with their macaque ortholog (center). The gene order of each region is encoded by the family IDs for each flanking gene, with the same family ID indicating a match; positive or negative values indicate the strand for each gene, though the strand is irrelevant to the match or mismatch. The length of the syntenic region estimated for each paralog is shown (shadowed), as are the variables representing the probabilities of finding homologous genes in syntenic and non-syntenic regions.

2.3. *Clustering and Assignment*

Once we have the length of the shared synteny for each pair of genes, we can cluster the genes of species S into parent and daughter based on the degree of shared synteny with the outgroup genome. Both the synteny length and the ratio of the synteny length to the total length are considered in distinguishing the two populations. Because the distribution of the log-transformed length of synteny shows better separation than the absolute length, the clustering used two variables: the log-transformed synteny length ($V_{loglen}=\log l_i$), and the ratio of the shared synteny length to the total gene-order length ($V_{ratio}=l_i/G_i$). Total gene-order length varies because there are a variable number of genes contained within the regions being compared for any pair of genes. If the total gene-order length is less than 30 genes, we consider the case uninformative and exclude it. We assume that parents have longer shared synteny with the outgroup than daughters, and should therefore have both higher absolute synteny lengths and higher ratios of synteny length to gene-order length.

We cluster the genes into two groups, parents or daughters, using a simple 2-means clustering on the two variables. Since the two variables use different scales, we first scale and center them. To do this we carry out a principal component analysis, and run the clustering on the two principal components. For all data points we calculate the difference between the distances to the two cluster means, such that points close to one center has high differences. If the difference of distances is smaller than 2 S.D. from the mean, the data point is labeled ambiguous. Since tandem duplications all share the same synteny, we cannot distinguish parents and daughters among tandem duplicates with this method. This means that there can be multiple genes assigned as 'parent' due to tandem duplication of the original parent gene. Likewise, tandem duplicates of a daughter gene are also assigned as 'daughter'. The clustering of each (o_i, g_i) pair is translated into the assignment for each gene g_i following a simple rule, such that if any pair containing g_i is clustered as a parent, then the gene g_i is a parent.

3. Results

We used lineage specific duplications from the human genome for our analysis, with macaque as the outgroup species. The dataset of gene families constructed from six mammalian genomes, each with a phylogenetic tree, is described in Hahn et al. [3]. To find human specific duplicates, we collected the genes under human specific duplication nodes from the reconciled tree. The outgroup genes were collected by selecting all macaque genes that are sibling to the duplicated node. We paired each human gene and each macaque gene within a family to assemble the dataset. For each gene, we downloaded the ±10 MB flanking region from Ensembl and extracted the gene order by encoding the genes according to the gene family IDs defined in Ensembl. We chose the length of the flanking region (±10 MB) to be long enough to contain duplications of all sizes and to have an adequate number of genes outside the duplicated segment so that we can confidently assign duplication breakpoints. The sizes of duplications in primates are mostly less than 300 KB and up to about 1 MB at most [12].

In total, we found 337 families with human specific duplications, containing 871 human genes and 385 macaque genes, and we constructed 1075 pairs with these genes. We ran the EM algorithm twice, once with initial values of p_{syn}=0.99 and p_{nonsyn}=0.01 and once with p_{syn}=0.51 and p_{nonsyn}=0.49. Both runs converged within 10 iterations. The estimated probability of finding a match in a syntenic region was p_{syn} = 0.829, and the probability of finding a match in a non-syntenic region was p_{nonsyn}= 0.050 (note that the two values do not have to add up to one). The average for the estimated lengths of shared synteny was 141 genes (SD=154). The distribution of the lengths showed two separate peaks, indicating that the population is a mixture of two distributions. (Fig. 3a)

We clustered the pairs using the variables V_{loglen} and V_{ratio}, as described above. To see how these two variables contribute to variance in the data, we conducted a principal components analysis. Almost 97% of the variance was explained by the equal contribution of both variables after they were scaled and centered. We carried out a 2-means clustering with the two principal components and obtained one cluster with low syntenic lengths and low synteny ratios (syntenic length/total length) and one with high syntenic lengths and high ratios. We were able to confidently assign 425 pairs as daughters and 603 pairs as parents. We classified 25 and 22 pairs that clustered with the daughters and parents, respectively, as ambiguous. (Fig. 3b) Using the clustered human-macaque pairs, we assigned each human gene as either a parent or daughter. This resulted in 826 unambiguous calls for the human proteins, 493 of them parents and 333 of them daughters.

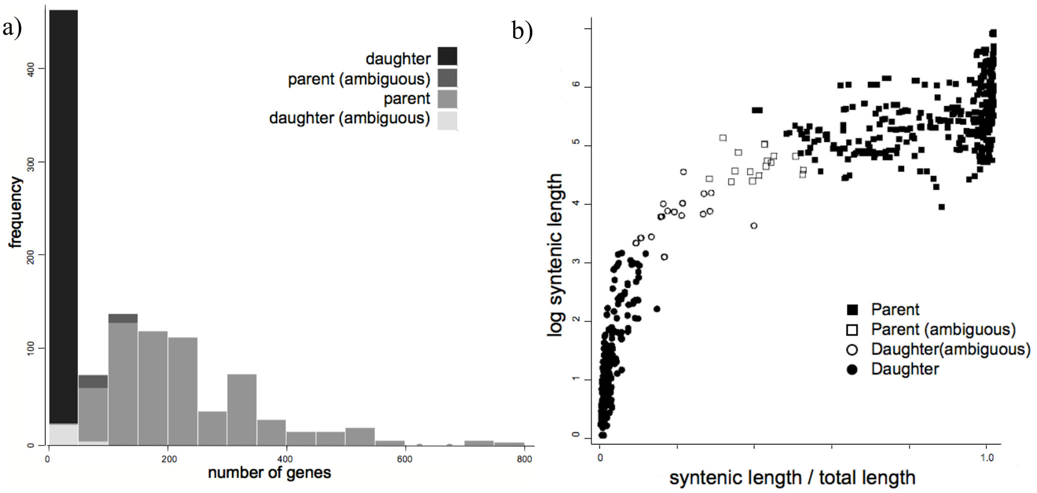

Figure 3. a) Distribution of estimated lengths of shared synteny for all pairs, measured in the number of genes in syntenic regions. b) Groups obtained by 2-means clustering on the principle components of V_{loglen} and V_{ratio} . Parent clusters have longer synteny and a higher ratio of synteny to total length. Daughter clusters have short synteny and small ratio of synteny to total length. 47 pairs have ambiguous assignment to either the parent or daughter cluster.

Based on the type and number of assigned genes, we classified the families into four classes (Table 1): 1) Families can have daughters only, likely because the original parent copy has been lost. 2) Families can have parents only, where no clear daughter can be identified because duplicate genes have been created by tandem duplication. 3) Families can have a single parent and a single daughter gene. 4) Families may contain multiple copies of both parent and daughter genes due to tandem duplication of either paralog.

Table 1. Counts of human genes in different classes based on parent-daughter assignments and the type and number of genes in a family. Ambiguous assignments are excluded from results. Counts of families are in parentheses.

Family Class	Daughter	Parent	Total
Daughter Only	201 (79)		201 (79)
Parent Only		357 (142)	357 (142)
Parent-Daughter (1:1)	85 (85)	85 (85)	170 (85)
Parent-Daughter (>1)	47 (24)	51 (24)	98 (24)
Total	333 (188)	493 (251)	826 (330)

We verified parent-daughter assignments by checking the chromosomal location of the genes. As expected, we found that all the human duplicates in the parent-only class were on the same chromosome, with two exceptions. The first case has a gene on each of chromosomes 14 and 15. These chromosomes are both homologous to the macaque chromosome 7, due to a chromosomal fission event along the human lineage [13]. The second case has one gene on chromosome 17 and four genes on chromosome X. It appears that the human gene located on chromosome 17 was misassembled with the X in our dataset, and has since been updated in the newest human genome assembly. In the parent-daughter 1:1 class, all but six pairs were on different chromosomes, as expected. The six pairs of genes that were on the same chromosome were all at least 19 MB and up to 91 MB apart, so that there were striking differences in the degree of shared synteny. The other classes do not have obvious predictions to measure against.

4. Discussion

In order to uncover the directionality of gene duplication events, we have introduced a probabilistic method for estimating shared syntenic blocks between genes. Using this method we are able to classify lineage specific paralogs as either parent or daughter copies, and thus to infer the direction of the duplication event that created the new gene. The directionality of duplications is revealing for a number of reasons. One of the most important distinctions between parents and daughters is that, though they are both technically orthologous with single-copy genes in the outgroup species, only the parent shares the same genomic environment with the outgroup (thus the term "primary ortholog") and the daughter gene is likely to experience a new genomic environment (thus "secondary ortholog"). The difference in the local environment may range from the presence of different regulatory signals to their epigenomic properties, such as different levels of histone modifications [15]. Under the model for the evolution of novel functions in gene duplicates introduced by Ohno [14], it seems probable that these functions are most likely to arise in the daughter genes because they are less likely to be able to maintain the complete ancestral function in their new location. Data on the prevalence of adaptive natural selection among paralogs (Han et al. *in review*) seems to bear this prediction out.

Because duplication via retrotransposition results in daughter copies that do not have any introns, previous research has been able to infer the directionality of these duplication events [4, 16]. While our method is able to identify parent-daughter relationships due to any mechanism of duplication, we can compare our results to these previous studies. In particular, Emerson et al. [16] found an excess of interchromosomal gene duplication involving the X chromosome. We therefore asked whether the same pattern could be found among all duplicates. We found no excess of parents or daughters on the X chromosome, though we did find an excess of tandem duplications on the X chromosome. This discordance with previous reports may be because we are looking at much younger genes than previous studies, namely only those duplicated since the human-macaque split.

Identification of parent-daughter relationships also provides insight into the molecular mechanisms of gene duplication. For instance, our results show that we can confidently identify at least 215 genes created by tandem duplication in the human lineage (357 parental genes in 142 parent-only families), vs. at least 164 genes created by duplication into a different location (79 daughter-only families + 85 parent-daughter pairs; Table 1). This high proportion of dispersed duplication is consistent with previous results [7]. Our results also provide information about the size of these interchromosomal duplications, with some duplications copying just one gene (largely via retrotransposition) and a number copying blocks up to 22 genes long between chromosomes. While the parent copies had an average synteny of 243 genes long, the daughter copies had an average duplication length of 4.5 genes. Approximately 57% of the duplications were less than 3 genes long, with the frequencies decreasing exponentially. We identified 23 unambiguous retrotransposition events—multiple exons in one copy and a single exon in the other—that could be used to verify our results. Of these, 8 out of 23 parents were misclassified into daughter-only classes due to a lack of synteny, 4 daughters were misclassified into parent-only classes because they were retrotransposed into a proximal location, and 11 out of 23 cases were correctly classified into parent-daughter relationships. Our method seems to suffer from low sensitivity when synteny with the outgroup is short due to disruptions in either lineage. The accuracy could be improved by incorporating additional variables such as exon number or a gene's rank of synteny among all family members. For instance, if we observe another member of a family assigned as parent with higher confidence, this information can add support to the assignment of an ambiguous gene as the daughter.

The major limitation of our approach is the method's dependence on shared synteny between species. Increased divergence will more likely disrupt the synteny between the ancestral regions, causing more families to be misclassified as daughter-only classes. We are planning to explore the effect of divergence with different species pairs in the future, as well as the possibility that incorporating multiple species could alleviate this problem. A major advantage of our approach is that we are able to estimate the probability of finding homologous genes in homologous genomic regions directly from the data. Considering that this value is dependent on both the time since divergence of two species and the rate of genomic rearrangement (including insertion, deletion, and inversion, among other factors), our results show that between homologous regions of the human and macaque genome there is an ~83% probability of finding homologous genes. While this represents an average probability across the genome, more complex models could incorporate regional variation. Our model for defining synteny can be applied to a number of problems in comparative genomics—such as distinguishing between orthologs and out-paralogs (Figure 1a)—and is not necessarily associated with the biological problem addressed here. However, direct comparison with alternative methods that use alignment of whole chromosomal DNA [11] will need to await further study.

References

1. Fitch, W.M., *Distinguishing homologous from analogous proteins.* Systematic Zoology, 1970. **19**(2): p. 99-113.
2. Sonnhammer, E.L.L. and E.V. Koonin, *Orthology, paralogy and proposed classification for paralog subtypes.* Trends in Genetics, 2002. **18**(12): p. 619-620.
3. Hahn, M.W., J.P. Demuth, and S.-G. Han, *Accelerated rate of gene gain and loss in primates.* Genetics, 2007. **177**: p. 1941-1949.

4. Betran, E., K. Thornton, and M. Long, *Retroposed New Genes Out of the X in Drosophila.* Genome Res., 2002: p. 1854-1859.
5. Fu, Z., et al., *MSOAR: A high-throughput ortholog assignment system based on genome rearrangement.* Journal of Computational Biology, 2007. **14**(9): p. 1160-1175.
6. Cannon, S. and N. Young, *OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies.* BMC Bioinformatics, 2003. **4**(1): p. 35.
7. Jiang, Z., et al., *Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.* Nat Genet, 2007. **39**(11): p. 1361-1368.
8. Nadeau, J.H. and B.A. Taylor, *Lengths of chromosomal segments conserved since divergence of man and mouse.* Proceedings of the National Academy of Sciences of the United States of America, 1984. **81**(3): p. 814-818.
9. Sankoff, D., V. Ferretti, and J.H. Nadeau, *Conserved segment identification*, in *Proceedings of the first annual international conference on Computational molecular biology*. 1997, ACM: Santa Fe, New Mexico, United States.
10. Pevzner, P. and G. Tesler, *Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes.* Genome Res., 2003. **13**(1): p. 37-45.
11. Kent, W.J., et al., *Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(20): p. 11484-11489.
12. Bailey, J. and E.E. Eichler, *Primate segmental duplications: crucibles of evolution, diversity and disease*, in *Nat Rev Genet*. 2006.
13. Rhesus Macaque Genome Sequencing and Analysis Consortium, *Evolutionary and biomedical insights from the Rhesus Macaque genome.* Science, 2007. **316**(5822): p. 222-234.
14. Ohno, S., *Evolution by gene duplication*. 1970, Berlin: Springer-Verlag.
15. Zheng, D., *Asymmetric histone modifications between the original and derived loci of human segmental duplications.* Genome Biology, 2008. **9**(7): p. R105.
16. Emerson, J.J., et al., *Extensive gene traffic on the mammalian X chromosome.* Science, 2004. **303**(5657): p. 537-540.