Tests for gene clustering*

Dannie Durand
Department of Biological Sciences,
Carnegie Mellon University,
Pittsburgh, PA 15213,
USA
email: durand@cmu.edu

David Sankoff
Department of Mathematics and Statistics,
University of Ottawa,
Ottawa, ONT K1N 6N5,
Canada
email: sankoff@uottawa.ca

November 5, 2003

Abstract

Comparing chromosomal gene order in two or more related species is an important approach to studying the forces that guide genome organization and evolution. Linked clusters of similar genes found in related genomes are often used to support arguments of evolutionary relatedness or functional selection. However, as the gene order and the gene complement of sister genomes diverge progressively due to large scale rearrangements, horizontal gene transfer, gene duplication and gene loss, it becomes increasingly difficult to determine whether observed similarities in local genomic structure are indeed remnants of common ancestral gene order, or are merely coincidences.

A rigorous comparative genomics requires principled methods for distinguishing chance commonalities, within or between genomes, from genuine historical or functional relationships. In this paper, we construct tests for significant groupings against null hypotheses of random gene order, taking incomplete clusters, multiple genomes and gene families into account. We consider both the significance of individual clusters of pre-specified genes, and the overall degree of clustering in whole genomes.

*To appear: J. Comput. Biol., 10(3-4):453-482.

1 Introduction

Comparison of gene order and content in related genomes is a rich source of information concerning genome evolution and function. Already an established approach in linkage genetics, comparative mapping has taken on new significance with the advent of whole genome sequencing. The biology literature contains an increasing number of articles in which local similarities in two or more genomes are presented as evidence of evolutionary relatedness or functional selection on gene order. To be convincing, such reports should reject the hypothesis that the observed similarities could have occurred by chance, yet many of those reports present no statistical analysis and those that do usually rely on intuitive criteria, *ad hoc* tests or, at best, randomization simulations. Very few formal probabilistic analyses of gene clustering have been presented and there is no consensus among them on what criteria best reflect biologically important features of gene clusters.

Biological background and significance: Speciation results in offspring genomes that initially have identical gene content and order. Similarly, whole genome duplication creates a new genome with two identical copies of the ancestral genome embedded in it. In both cases, the gene complement and gene order of the offspring genomes will diverge over time. Gene duplication and loss and horizontal gene transfer result in changes in gene complement, while gene order is disrupted by large scale rearrangements, including translocation, transposition, inversion and chromosome fission and fusion.

In the absence of selective pressure on gene order, successive rearrangement will lead to randomization of gene order. Therefore, similarity in genomic organization is a source of evidence for inferring evolutionary relationships and/or for predicting the functional roles of gene clusters. The availability of comprehensive linkage maps for numerous plant and animal species, as well as a rapidly growing number of whole genome sequences, has stimulated many lines of inquiry based on this evidence. For example, comparison of genetic maps of two or more species has been used to infer patterns of chromosomal rearrangement (Coghlan and Wolfe 2002; Ehrlich et al. 1997; Nadeau and Sankoff 1998b; Nadeau and Taylor 1984; Seoighe and Wolfe 1998) and as a basis for alternative phylogenetic approaches (Blanchette et al. 1999; Cosner et al. 2000; Hannenhalli et al. 1995; Sankoff et al. 2000a; Sankoff et al. 2000b; Tamames et al. 2001). Comparison of a genetic map from a single species with itself has been used to analyze patterns of gene duplication in genome evolution (Arabidopsis Genome Initiative 2000; El-Mabrouk and Sankoff 2003; El-Mabrouk et al. 1998; McLysaght et al. 2002; Semple and Wolfe 1999; Seoighe and Wolfe 1999; Venter and others 2001; Vision et al. 2000; Wolfe and Shields 1997). Interest in such questions has spawned a growing body of research in algorithms for inferring the history of rearrangements (see, for example, Pevzner 2000; Sankoff and El-Mabrouk 2002 for surveys.) In microbial genomics, comparisons of gene content and order have also been used to study the importance of spatial organization in genome function including functional selection (Huynen and Bork 1998; Kolsto 1997; Overbeek et al. 1999; Tamames 2001; Tamames et al. 1997), operon formation (Bork et al. 2000; Ermolaeva et al. 2001) and horizontal transfer (Lawrence and Roth 1996).

Identification of conserved chromosomal segments is a major problem for many of these analyses. Intuitively, rearrangement processes should result in a pattern of *conserved segments*, pairs of chromosomal regions, one in each genome, that are descended from a single, contiguous region

in the ancestral genome. Closely related species should manifest a few, long conserved segments, while distantly related species should have many short segments, since each rearrangement cuts one or more segments into shorter pieces. Because rearrangements may involve transfers of arbitrarily long chromosomal fragments to arbitrary locations within the genome, conserved segments that are adjacent in one genome will not necessarily be close to each other in the sister genomes.

According to the most stringent definitions, conserved segments are defined to be two or more contiguous regions that contain the same genes in the same order (Nadeau and Sankoff 1998a; O'Brien et al. 1997) and, in some cases, in the same orientation (Overbeek et al. 1999; Tamames 2001; Wolfe and Shields 1997). However, it is common practice in indicating conserved segments in comparative genomic maps to disregard small deviations from strict conservation of gene order (Goldberg et al. 2000; Sankoff et al. 1997). For example, the human-mouse comparisons in recent genome sequencing reports (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002) indicate only around 200 segments, many of which are known to contain small inversions and other inconsistencies.

In studies that focus on large scale genome organization and rearrangements, less strictly defined *gene clusters* are the units of interest. Under some rearrangement regimes (e.g. short inversions, single gene insertion, loss or duplication), a high degree of gene proximity is conserved, even while gene order is rapidly scrambled (Sankoff 2002). Strict notions of conserved segment lead to unstable estimates of the number of segments if these may be as small as one or two genes (e.g., Kumar et al. 2001). Strict definitions are also inappropriate in the presence of positional errors. These observations have led to formalizations of a more flexible concept of gene cluster, as well as algorithms for finding gene clusters given these more relaxed definitions (Arabidopsis Genome Initiative 2000; Bansal 1999; Bergeron et al. 2002; Bork et al. 2000; Goldberg et al. 2001; Heber and Stoye 2001a; Heber and Stoye 2001b; Nadeau and Sankoff 1998a; O'Brien et al. 1997; Overbeek et al. 1999; Tamames 2001; Venter and others 2001).

Our Results: Given a method for identifying gene clusters, how can we assess whether they are statistically meaningful? The issue of cluster significance arises in two types of analysis: detailed study of the history or function of a particular set of genes and large scale studies of the selective forces acting on the genome as a whole. This leads to two statistical questions:

> **Individual clusters:** Is it significant to find a *particular* set of genes in close proximity in two or more distinct, genomic regions?
>
> **Whole genome clustering:** Given two genomes, is the observation that a certain number of gene clusters appear in both genomes significant?

The problem of significance testing for gene clusters has been introduced by previous authors (Trachtulec and Forejt 2001; Venter and others 2001). Their results, described in Section 5, use combinatorial analysis to model clusters found under a limited set of conditions. In the current paper, we model a broad range of scenarios, taking into account the size of the gene families in the genomes where the cluster is found, the completeness of the cluster, the number of instances of the cluster observed, how the cluster was found (i.e., the size of the search space considered.) The number of factors that contribute to cluster significance, as well as the complexity of the interactions between these factors, makes it difficult to predict how significance varies as a function

of basic parameters. The goal of the current paper is to address this problem through detailed probabilistic models. In Section 2, we focus on a single cluster of pre-specified genes, providing exact expressions for the probability of finding a given set of m genes in a window of size r. In Section 3, we extend these results to derive significance tests for individual clusters, including incomplete clusters found in two or more genomes with gene families. Probabilistic clustering models of whole genome comparison, including both genome self-comparison and comparison of genomes from different species, are presented in Section 4. The application of these results to specific biological problems is discussed in Section 6.

2 Simple cluster probabilities

We begin by introducing a simple definition of a gene cluster and calculate the probability of observing such a cluster in a genome with uniform random gene order (a "random genome"). Let genome, $G = (1, \ldots, n)$, be an ordered set of n genes and let M be a pre-selected set of m genes of interest. These m genes may be of interest because they are contiguous in some other genome ("the reference genome") or because they share a functional property. In any case, the spatial organization of the genes on the reference genome does not enter into the analysis at this point.

Consider the case where the genes in M are found in any order in a window of *exactly* r slots in G. In this case, the first and last of the r slots contain two of the m genes and the remaining $r-2$ slots contain the remaining $m-2$ genes plus $r-m$ intruders. The probability[1] of this event is

$$\frac{(n-r+1)\binom{r-2}{m-2}}{\binom{n}{m}} . \tag{1}$$

In the event these m genes span *at most* r slots in G, it suffices that one of the end points of the window be occupied by one of the m genes. In this case, the probability is

$$q(n,m,r) = \frac{\left[(n-r)\binom{r-1}{m-1}\right] + \binom{r}{m}}{\binom{n}{m}} , \tag{2}$$

where the second term in the numerator addresses edge effects. If we require that the genes in M appear in a given order, then the probability of observing the cluster is $q(n,m,r)/m!$.

Intuitively, we expect the probability of finding a cluster by chance will depend on the size of the window, relative to the genome size, and the fraction of slots in the window that are occupied by intruders. For large n and $m < r$, we can make this intuition explicit by applying Stirling's

[1] Equations 1–3 correct the formulations given in a preliminary version of this paper (Durand and Sankoff 2002).

approximation to Equation 2 to obtain

$$q(n, m, r) \approx \left(\frac{w\theta}{e}\right)^m \theta^{-(r-\frac{1}{2})}, \tag{3}$$

where $w = \frac{r}{n}$ and $\theta = 1 - \frac{m}{r}$ are two parameters introduced to represent *window proportion* and *window sparsity*, respectively.

We have introduced a very simple gene cluster model. Using it to address concrete questions of significance in comparative genomics depends on the biological question at hand, specific properties of the genomic data, whether the clusters in question are paralogous or orthologous and how those clusters were found. In the next two sections, we build on this model to construct significance tests that take these issues into account.

3 Significance of individual gene clusters

A *conserved* cluster is a set of two or more distinct (non-overlapping) chromosomal regions that have m genes in common; that is, there are m genes in one region for which a homolog may be found in the other. The cluster is *paralogous* if both regions are in the same genome, *orthologous* if the regions are in different genomes. Many conserved gene clusters, both paralogous and orthologous, have been reported in eukaryotes (see Table 2 in Section 6 for a survey of this literature), usually in the context of evolutionary studies. There is also a rich literature describing gene clusters in prokaryotes (Bork et al. 2000; Overbeek et al. 1999; Snel et al. 2000; Suyama and Bork 2001; Tamames 2001; Tamames et al. 1997), mostly concerned with the functional constraints that account for similarities in gene order.

Gene families: Identification of homology relationships between genes is a prerequisite to finding and testing gene clusters. Virtually all genomes contain gene families, sets of genes with similar sequence and function, that arose through duplication of genetic material. The problem of identifying true homologs has been much debated and a variety of solutions, typically using sequence analysis, have been proposed (e.g., Adams et al. 2000; Venter et al. 2001; Huynen and Bork 1998; Overbeek et al. 1999; Tatusov et al. 1997). In this paper, we assume that homology relationships have already been established in a pre-processing step and that the set of genes in G can be partitioned into non-intersecting gene families. Let gene family f_{ij} be a set of genes in genome G_i, such that each gene in f_{ij} is homologous to all other genes in f_{ij} and only those genes. (If a single genome, G, is under consideration, then jth gene family may simply be denoted f_j.) There are $\phi_{ij} = |f_{ij}|$ genes in family, f_{ij}. Let $\mathcal{F} = \{f_j\}$ be the set of all gene families in all of the genomes under consideration. By convention, we require that the number of gene families, $n_f = |\mathcal{F}|$, be the same in all genomes, but allow gene families to have zero members in some genomes.

The existence of gene families implies that, in general, a given gene $g \in M$ will be homologous to more than one gene in G. When it is possible to identify which gene in the homologous gene family is most closely related to g, then g has a unique match in G. This may be possible after a recent speciation or polyploidization event. However, because of convergent evolution, non-homologous gene displacement and multi-domain proteins generated by exon shuffling, it is usually

not possible to identify a unique match. Thus, a general model of cluster significance must allow for the possibility that more one gene in G matches each gene in M. As the size of gene families increases, so do chance occurrences of gene clusters. For example, if just one of the genes in M has two homologs, then there are two different sets of m genes that qualify as clusters and the probability of finding the cluster by chance almost doubles. Since the probability of observing a cluster depends on the number of genes in G that match each of the pre-specified genes, we derive test statistics that take gene family size as a parameter. Our test statistics model the most general case in which there are no constraints on the distribution of gene family sizes. However, a complete catalog of all gene families and their sizes for the genome in question is required to compute these statistics. This is currently possible for a few fully sequenced species, but requires consideration of all possible sets of m gene families, an onerous task. Distributions of gene family sizes, under various assumptions, have been published for a number of species (e.g., Friedman and Hughes 2001; Huynen and van Nimwegen 1998; Li et al. 2001; Nadeau and Sankoff 1997; Rubin et al. 2000; Tiuryn et al. 2000; Yanai et al. 2000). In future work, we plan to derive approximations that are more easily calculated based on parameterized models of such distributions. A first approach is to assume that each gene family has the same number of paralogs, ϕ. This assumption is only true under a limited set of circumstances, such as after recent speciation or polyploidization, when immediate identification of a unique homolog for each gene is possible. Nevertheless, it allows us to derive test statistics that are easy to calculate, providing a useful tool for exploring how cluster significance varies with gene family size. In the following analysis, in addition to general test statistics for any distribution of gene family sizes, we give tests based on simplified expressions assuming uniform gene family sizes.

The significance of a putative cluster depends on how it was found. The observation of a cluster may be a serendipitous finding. Alternatively, the authors may have been interested in a particular region and have searched one or more genomes for similar regions. Or, gene clusters may be found in "fishing expeditions" for clusters in whole genome comparisons. The significance of the cluster will depend on the number of possibilities considered during the search, yet the circumstances of the discovery are frequently not reported, resulting in presentations that are misleading as to the significance of the clusters. In order to address this issue, we model three common approaches to cluster analysis:

- **Reference region:** In some cases, an investigator is interested in a particular genomic region and searches for additional regions containing the same genes. In this case, we redefine the set of pre-specified genes, M, to be the set of genes found in the first region (the "reference region"). In a genome with no gene families, ($\phi_j = 1$, for all j), the simple cluster model developed in Section 2 can be used to test significance. In Section 3.1, we extend the model to derive tests for genomes with gene families ($\phi_j \geq 1$), as well as the case where the second region contains only a subset of the genes in the reference region. We also discuss the significance of clusters that appear in more than two genomes.

- **Window sampling:** Alternatively, a cluster may be found by selecting a pair of windows and comparing their gene content for shared homologs. This situation arises, for example, when an investigator searches in the vicinity of a pair previously known homologous genes

for additional evidence of conservation. We derive significance tests for clusters found by window sampling in Section 3.2. These tests also serve as the foundation for tests of the significance of aggregate clustering properties developed in Section 4.

- **Whole genome comparison:** Finally, individual gene clusters are often found through whole genome scans, although the description of the cluster may read as if the genes involved were the original focus of interest; i.e., were pre-specified as in Section 2. In this case, tests based on whole genome comparison models must be used to avoid underestimation of cluster significance due to a much larger search space. There are several natural whole genome comparison models. In Section 3.3, we present an approach in which one genome is searched for localized occurrences of every set m genes that are contiguous in a second, "reference" genome. Tests based on other models of whole genome comparison are presented in Section 4.

In the following sections, we derive a variety of statistical tests for determining the significance of gene clusters by rejecting the hypothesis that such a cluster could have occurred by chance in a genome with uniform random gene order, the most basic null hypothesis we can consider. If we cannot reject that null hypothesis, no more complex, biologically motivated null hypothesis need be considered.

Test statistics: The probability of observing a single cluster of pre-specified genes (Equation 2) is a measure of its statistical significance. However, for the more complex biological scenarios described above, there will typically be more than one cluster of genes that meet the criterion under consideration. In this case, the probability of observing at least one such cluster may be used to test significance. Unfortunately, in many instances this probability is difficult to calculate because some sets of genes that meet the criterion intersect, so that the events under consideration are not independent.

It is generally easier to calculate the expected number of clusters of a given type. Such a result can be used as a benchmark or informal test; if the number of observed clusters, ν, is much greater than the expected number, S, we can assume that about $\nu - S$ of them represent evolutionary or functionally derived clusters. Markov's inequality provides a formal, albeit weak, test: if $\nu > S/\alpha$, then the number of observed clusters exceeds the null hypothesis at a significance level of α.

The above approach assumes that it is possible to calculate the number of observed clusters, ν, from experimental data. In some cases, enumerating all observed clusters may be difficult and it is more convenient to use an approach based on sampling windows from the genome. In this case, significance tests focus on the expected number of windows in the sample that contain a cluster of interest and on the probability that the sample contains at least one such window.

We derive test statistics for both orthologous and paralogous clusters. The derivation is similar in both cases, but statistics for paralogous clusters are complicated by the requirement that paralogous clusters not overlap, nor share the same genes. In each section, the analysis for orthologous clusters is introduced first, followed by a discussion of how the test statistics may be adapted to the paralogous case. Tables summarizing all of these test statistics appear in the appendix.

Notation: For any given biological scenario, the probability of observing a cluster in a random genome is denoted by $q()$, the expected number of clusters in a random genome by $S()$ and the probability of observing at least one cluster in a random genome by $P()$. The situation in which

the statistic is applicable is indicated by sub- and superscripting. Test statistics based on window sampling are subscripted with a W. The superscripts o and p indicate orthologous and paralogous clusters, respectively. The superscripts Φ and F refer to gene families, H to incomplete clusters, N to multiple genomes, C to clusters found by whole genome comparison and R to clusters found by comparisons to a reference genome.

3.1 Reference regions

Suppose a chromosomal region in one genome is of particular interest (the "reference region") and a second region is found by scanning a different genome, G, for genes found in the first region. Define M to be the genes found in the reference region. We can then use the probability that m pre-specified genes are found in a window of size r to test the significance of a cluster found in this way. If G does not contain gene families and all m genes in M are found in the window in G, then Equations 2 and 3 can be used to test whether a specific set of m genes is more highly clustered than by chance.

Clusters in genomes with gene families: We now extend this model to orthologous clusters in genomes with gene families, defining M to be a set of m pre-specified *gene families*, $M = \{f_1 \ldots f_m\}$. (We assume that no two genes in the reference region are members of the same gene family.) At the end of this section, we give test statistics for paralogous clusters, where the reference region and the matching region are found in the same genome. Let $\mathcal{M}$ be the set of distinct sets of genes that are homologous to M in G_i. There are

$$\Phi_i(M) = \prod_{f_j \in M} \phi_{ij} \tag{4}$$

sets in $\mathcal{M}$[2]. For each of these, the probability that it spans at most r slots is $q(n, m, r)$. Thus, the expected number of homologous clusters is

$$S_\phi(n, m, r) = \Phi(M) q(n, m, r). \tag{5}$$

Notice that Equation 5 gives the expected number of sets found in windows of size r but does not require that these windows be non-overlapping. Thus, Equation 5 does not capture the expected number of *distinct* clusters. For this reason, significance tests based on the probability of observing *at least* one homologous cluster are easier to interpret. This probability is

$$P_\Phi(n, m, r) = \text{Prob}(\cup_{i=1}^{\Phi(M)} E_i), \tag{6}$$

where E_i is the event that the ith set in $\mathcal{M}$ is found in a window of size r. The approximation $P_\Phi(n, m, r) \approx 1 - [1 - q(n, m, r)]^{\Phi(M)}$ can be used for rough tests, but it is based on an unwarranted assumption of independence of occurrence among the $\Phi(M)$ possible clusters.

A better approximation of $P_\Phi(n, m, r)$ can be estimated using the inclusion-exclusion rule to correct for overlapping clusters. Let $E_{i_1, \cdots, i_g}$ be the event that each of the sets $i_l \in \mathcal{M}$,

[2]When the identity of the genome under consideration is unambiguous, we refer simply to ϕ_j and $\Phi(M)$.

$l \in (1, \cdots, g)$, appears in a window of size at most r in G. Then, by the inclusion-exclusion rule,

$$P_\Phi(n, m, r) = \sum_{i=1}^{\Phi(M)} \text{Prob}(E_i) - \sum_{i_1 \neq i_2}^{\Phi(M)} \text{Prob}(E_{i1,i2}) + \sum_{i_1 \neq i_2 \neq i_3}^{\Phi(M)} \text{Prob}(E_{i1,i2,i3}) - \cdots \quad (7)$$

The first term of this equation is $S_\Phi(n, m, r)$ and the remaining terms correct for intersecting sets. In the genomic context where n is large, the dominant term of this correction will be due to pairs of clusters that share identical genes in all but one of the m families. The windows containing such a pair must overlap by at least $m - 1$ positions. Thus we can estimate the second term of Equation 7 by calculating

$$S'_\Phi(n, m, r) = q(n, m+1, 2r-m+1) \sum_{j \in M} \frac{\Phi(M)}{\phi_j} \binom{\phi_j}{2}, \quad (8)$$

the expected number of windows of size $2r - m + 1$ containing a cluster plus an extra member of one of the m families. This is only an estimate of the second term because not every such window will be the union of two windows of size at most r each containing a complete cluster. Then

$$P_\Phi(n, m, r) \approx S_\Phi(n, m, r) - S'_\Phi(n, m, r) \quad (9)$$

represents a first order approximation to the probability that at least one cluster appears.

A simplified model can be obtained under the assumption that the gene family size is uniform over all gene families. In this case, the quantity $\Phi(M)$ in Equations 5 and 8 can be replaced by ϕ^m, where ϕ is the fixed gene family size, yielding

$$P_\Phi(n, m, r, \phi) \approx \phi^m \left[q(n, m, r) - \frac{m(\phi - 1)}{2} q(n, m+1, 2r-m+1) \right]. \quad (10)$$

The conditions under which Equation 10 yields a good approximation have not yet been investigated.

Incomplete clusters: Frequently, only a subset of the m genes of interest is found in close proximity in the genome. When is this event significant? To model this scenario, let $\mathcal{H}$ be the set of all subsets of M of size $h < m$. In the absence of gene families, the probability that a specific subset in $\mathcal{H}$ appears in a window spanning at most r slots is $q(n, h, r)$ and the expected number of such subsets is

$$S_H(n, h, m, r) = \binom{m}{h} q(n, h, r). \quad (11)$$

As above, the subsets in $\mathcal{H}$ may intersect. For example, if all m genes are found in a single window of length r, then G contains all the incomplete clusters in $\mathcal{H}$ but only one biologically interesting cluster.

The probability of observing at least one incomplete cluster can again be estimated by using the inclusion-exclusion rule (Equation 7) to correct for overlapping clusters. In this case, the first term of that equation is $S_H(n,h,m,r)$ and the remaining terms correct for intersecting subsets. For large n, the dominant term of this correction will be due to pairs of subsets whose intersections are as large as possible, namely of size $h-1$. The windows containing such a pair must overlap by at least $h-1$ positions. Thus we can estimate the dominant term of Equation 7 by calculating

$$S'_H(n,h,m,r) = \binom{m}{h+1} q(n,h+1,2r-h+1),$$

the expected number of windows of size $2r-h+1$ containing $h+1$ of the m genes. As above, this is not exact because not every such window will be the union of two windows of size at most r each containing h members of M. Then

$$P_H(n,h,m,r) \approx S_H(n,h,m,r) - S'_H(n,h,m,r) \tag{12}$$

represents a first order approximation to the probability that at least one incomplete cluster of size h appears in G.

An upper bound on the probability of finding at least one incomplete cluster can be derived by observing that, given a *particular* window of size $r \geq h$ of G, the probability that *exactly* h of the m genes fall into that window is given by the hypergeometric distribution. The probability that *at least* h of the m genes fall into that window can be calculated by summing over that hypergeometric probability so that

$$q_{HW}(n,h,m,r) = \sum_{i=h}^{\min(r,m)} \frac{\binom{m}{i}\binom{n-m}{r-i}}{\binom{n}{r}}. \tag{13}$$

The probability of finding at least one incomplete cluster from $\mathcal{H}$ anywhere in the genome can now be bounded above by sampling all windows of size r in G that have a gene from M in the first position:

$$P_H(n,h,m,r) \leq m\, q_{HW}(n,h\text{–}1,m\text{–}1,r\text{–}1). \tag{14}$$

In the usual case where n is very large and either r or m is small, the combinatorial terms involving n may be approximated, and q_{HW} rapidly calculated. In the case of larger m and r, we may use the binomial approximation to the hypergeometric:

$$P_H(n,h,m,r) \lessapprox m \sum_{i=h}^{\min(r,m)} \binom{r}{i} \left(\frac{m}{n}\right)^i \left(1-\frac{m}{n}\right)^{r-i}. \tag{15}$$

Alternatively, one may numerically integrate a normal approximation with mean $\frac{rm}{n}$ and variance

$$\left(\frac{n-r}{n-1}\right)\left(\frac{rm}{n}\right)\left(1-\frac{m}{n}\right). \tag{16}$$

These approximations improve as m increases with respect to r.

The expected number of incomplete clusters in a genome of gene families can be derived by combining Equations 5 and 11, yielding

$$S_{\Phi H}(n, h, m, r) = \left(\sum_{H \in \mathcal{H}} \Phi(H) \right) q(n, h, r). \tag{17}$$

For uniform gene families of size ϕ, this expression simplifies to

$$S_{\Phi H}(n, h, m, r, \phi) = \binom{m}{h} \phi^h q(n, h, r). \tag{18}$$

Clusters in multiple genomes: The probability that a gene cluster is a chance occurrence decreases, and its statistical significance increases, if this cluster is found in more than one genome. For N genomes of same gene content with no gene families, the probability that a specific set of m genes appear in all these genomes in windows spanning at most r slots is $q^N = q(n, m, r)^N$, which becomes very small as N increases, even if q itself is only moderately small.

The announcement of a cluster validated by occurrences in multiple genomes should, however, be accompanied by a report of

1. An account of how many genomes were searched for the cluster, not just the number of genomes in which it was found.

2. How complete the cluster is in each of the genomes in which it is found.

3. The values of r and n for each genome considered.

These considerations all have an impact on the statistical significance of a cluster, in many cases weakening it.

For example, suppose N genomes were examined and a cluster was found in only $N' < N$ of them. The probability that the cluster appears in at least N' genomes, spanning at most r slots in each case, is

$$P_N = \sum_{j=N'}^{N} \binom{N}{j} q^j (1-q)^{N-j}, \tag{19}$$

which should be used for testing purposes instead of $q^{N'}$. The greater the difference between N and N', the lower the significance of the cluster.

Furthermore, the typical case reported in the literature is one where different subsets of M are found in different genomes. Consider N random genomes of size $n_1, \cdots, n_N$ containing subsets of M of sizes $m_1, \cdots, m_N$, respectively. The probability that, for each genome G_i, at least one subset of M of size h_i appears in G_i in a window spanning at most r_i slots is $\Pi_{i=1}^{N} P_H(n_i, h_i, m_i, r_i)$, where each $h_i \leq \min[m_i, r_i]$. An *ad hoc* test based on all the h_i and r_i is not rigorous, however, since in general these criteria are not rigidly fixed before the search, but are the result of it. For fairness,

then, the test should be based on fixed parameters that are the least favorable to rejecting the null hypothesis, namely $h = \min[h_1, \cdots, h_N]$ and $r = \max[r_1, \cdots, r_N]$ and the test distribution becomes

$$P_N = \prod_{i=1}^{N} P_H(n_i, h, m_i, r), \tag{20}$$

where P_H is defined in Equations 12 and 14.

In the general case of incomplete clusters in some genomes and missing clusters from others, to get an expression analogous to Equation 19, we need to substitute some uniform value of $P_H(n, h, m, r)$ for q. The appropriate values of n and m, those least favorable to rejecting the null hypothesis, are $n = \min[n_1, \cdots, n_N]$ and $m = \max[m_1, \cdots, m_N]$. Then the probability that subsets of M of size h appear in at least N' of the N genomes, spanning at most r slots in each case, is obtained by substituting $P_H(n, h, m, r)$ for q in Equation 19, yielding

$$P_N = \sum_{j=N'}^{N} \binom{N}{j} P_H^j (1 - P_H)^{N-j}. \tag{21}$$

In a multigenome analysis, the researcher must trade off the reduced statistical significance of incomplete instances of a cluster against the increased significance of its widespread occurrence. As Equation 12 shows, the significance of an incomplete cluster decreases as the window size r increases and h decreases. Decreasing the minimum ratio of h to r required for inclusion in the analysis diminishes the significance of the clusters included by increasing $P_H()$. If, on the other hand, only compact clusters (with a higher h to r ratio) are considered, the number of genomes in which the cluster is found will be reduced, also leading to diminished significance.

In this section, we derived tests for orthologous gene clusters; that is, clusters observed in G that are orthologous to a reference region in a different genome, G_1. When G contains gene families, significance tests for such clusters can be performed using either the expected number of clusters (Equation 5) or the probability of observing at least one cluster (approximated by Equation 9). The significance of incomplete, orthologous clusters in genomes with no gene families can be tested using Equation 11, when $S_H() \ll 1$. Alternatively, the approximations in Equations 12 and 14 can be used to show that the probability of observing at least one such cluster is small. Equations 17 and 18 give test statistics for incomplete clusters in genomes with gene families. The significance of clusters found in multiple genomes may be tested using Equations 19 – 21

Tests for paralogous clusters, where both regions appear in the same genome, G, are also needed. The tests derived here may be adapted to the paralogous case by adjusting for the fact that genes that appear in the reference region cannot also appear in the matching region. In the paralogous case, there are only $\phi_j - 1$ possible matches for each family, $f_j \in M$. Furthermore, since the reference region is excluded from the search for the cluster, we must replace n with $n - m$. Test statistics that take these factors into account are given in Table 4 in the appendix.

3.2 Clusters found through window sampling

A cluster may also be discovered by searching the genomic neighborhoods surrounding a pair of known homologous genes. In this case, significance tests can be developed based on a window sampling approach. We first consider pairs of windows sampled from two different genomes and subsequently discuss how our model must be modified for the case where both windows are sampled from a single genome.

Two genomes, no gene families: Consider two genomes, G_1 and G_2, with no gene families and a pair of windows, W_1 and W_2, of length r selected from G_1 and G_2, respectively. The probability that they have at least m genes in common is

$$q^o_W(n,r,m) = \sum_{i=m}^{r} \frac{\binom{r}{i}\binom{n-r}{r-i}}{\binom{n}{r}}, \tag{22}$$

where the expression inside the sum is simply the probability that *exactly* i of the r genes in W_1 also appear in W_2.

Two genomes with gene families: Now[3], suppose genomes G_1 and G_2 of length n_1 and n_2 each contain the same set of gene families $\mathcal{F} = \{f\}$, including singletons (families consisting of only one copy of a gene). Let $\mathcal{F}^k = \{F\}$ be the set of all subsets of $\mathcal{F}$ containing k gene families, written $F = \{f_{i_1}, \ldots, f_{i_k}\}$. Note that the restriction that all genomes have the same gene families is in no way essential to the discussion in this section. It may be completely abandoned without changing notation, under the interpretation that $p_i(F) = 0$ if genome G_i is missing any gene family $f \in F$, where $p_i()$ refers to $p_1()$ and $p_2()$ defined below.

The probability that two arbitrarily chosen windows W_1 and W_2 share at least m gene families is

$$q^o_{WF}(m) = \sum_{k=m}^{r} \left[\sum_{F \in \mathcal{F}^k} \left(p_1(F) \sum_{l=m}^{k} \sum_{\substack{F' \in \mathcal{F}^l \\ F' \subseteq F}} p^o_2(F') \right) \right], \tag{23}$$

where $p_1(F)$ is the probability that the gene families in W_1 are just the k families in F and $p^o_2(F')$ is the probability that a subset $F' \subseteq F$ containing l of those k families appears in W_2. To determine $p_1(F)$, we must consider all ways that genes from the k distinct gene families in F can fill a window of size r. Let x_{1j} be the number of genes in W_1 from the jth family in F, out of a total of ϕ_{1j}

[3]The following material corrects the derivation of $p_2()$ given in a preliminary version of this paper (Durand and Sankoff 2002).

genes in this family in G_1. Then

$$p_1(F) = \frac{\sum_{\mathcal{S}(F)} \binom{\phi_{11}}{x_{11}} \binom{\phi_{12}}{x_{12}} \cdots \binom{\phi_{1k}}{x_{1k}}}{\binom{n_1}{r}}, \tag{24}$$

where $\mathcal{S}(F)$ is the set of all k-tuples $(x_{11}, \ldots, x_{1k})$ such that

$$\sum_{h=1}^{k} x_{1h} = r,$$
$$0 < x_{1h} \leq \phi_{1h}.$$

To determine $p_2^o(F')$, we must consider all ways that genes from the l distinct gene families in F' can partially or completely fill a window of size r, without any of the $k - l$ families in $F \setminus F'$ being represented. Let x_{2j} be the number of genes in W_2 from the jth family in F', out of a total of ϕ_{2j} genes in this family in G_2. Then

$$p_2^o(F') = \frac{\sum_{\mathcal{S}(F')} \binom{\phi_{21}}{x_{21}} \cdot \binom{\phi_{22}}{x_{22}} \cdots \binom{\phi_{2l}}{x_{2l}} \binom{n_2 - \sum_{h=1}^{k} \phi_{2h}}{r - \sum_{h=1}^{l} x_{2h}}}{\binom{n_2}{r}}, \tag{25}$$

where $\mathcal{S}(F')$ is the set of all l-tuples $(x_{21}, \ldots, x_{2l})$ such that

$$\sum_{h=1}^{l} x_{2h} \leq r,$$
$$0 < x_{2h} \leq \phi_{2h}.$$

Note that the summation of the $\{\phi_{2h}\}$ in the last factor of the numerator in Equation 25, representing the reduction in the pool of genes from which W_2 must be filled after using the $\sum_{h=1}^{l} x_{2h}$ genes from the families in F', excludes all k families in F, not only the l families in F'.

More than two genomes with gene families: Can we generalize these considerations to the case of windows $W_1, \cdots, W_N$ in the N genomes $G_1, \cdots, G_N$? When $N = 2$, the notion that two windows share m gene families is well defined. However, for $N > 2$, there is more than one natural interpretation of this concept, each corresponding to a different biological situation and requiring a different mathematical realization. We consider just one of these here. Suppose we are given F, some set of k gene families of interest, where $k \leq n_f$. The probability that each window W_i contains elements of at least $m \leq k$ of these families is

$$q_{W_1 \cdots W_N}(m) = \prod_{i=1}^{N} \left(\sum_{l=m}^{k} \sum_{\substack{F' \in \mathcal{F}^l \\ F' \subseteq F}} p_i(F') \right), \tag{26}$$

where each $p_i()$ is an expression of the form of Equation 25, substituting family sizes ϕ_{ij} and genome length n_i for genome, G_i. Note that this expression differs from Equation 23 in that the set of shared gene families, F, is pre-specified. In contrast, when $N = 2$, F is the set of distinct gene families found in one of the two windows. Either W_1 or W_2 may be treated as a reference window, without loss of generality. However, for $N > 2$, selecting one of the N windows as a reference introduces an undesirable asymmetry.

Two genomes with gene families of equal size: Calculating test statistics based on window sampling is simpler under the assumption that all gene families are of equal size. In the comparison of two genomes, the probability that a given pair of windows, W_1 and W_2, of length r, selected from G_1 and G_2, respectively, share at least m gene families simplifies to

$$q^o_{WF}(m) = \sum_{k=m}^{r} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p^o_2(l) \right]. \tag{27}$$

The first term, the probability that k given gene families appear in W_1, is given by

$$p_1(k) = \frac{\sum_{\mathcal{S}} \binom{\phi}{x_{11}} \binom{\phi}{x_{12}} \cdots \binom{\phi}{x_{1k}}}{\binom{n_1}{r}}, \tag{28}$$

where $\mathcal{S}$ is the set of all k-tuples $(x_{11}, \ldots, x_{1k})$ such that

$$\sum_{h=1}^{k} x_{1h} = r,$$
$$0 < x_{1h} \leq \phi.$$

Table 1 shows an example of $\mathcal{S}$ when $r = 5$, $k = 3$ and $\phi = 4$. For the purposes of this calculation, it is not necessary to distinguish between individual members of a gene family. Nor does the order of the genes in the window matter. Thus it is sufficient to give the elements from each family in lexicographic order. In this case, $\mathcal{S}$ contains six elements. For example, the first row corresponds

	x_1, x_2, x_3	z_1, z_2, z_3, z_4
(a, b, c, c, c)	*1,1,3*	2,0,1,0
(a, b, b, b, c)	1, 3, 1	2,0,1,0
(a, a, a, b, c)	3, 1, 1	2,0,1,0
(a, b, b, c, c)	*1,2,2*	1,2,0,0
(a, a, b, c, c)	2, 1, 2	1,2,0,0
(a, a, b, b, c)	2, 2, 1	1,2,0,0

Table 1: All possible ways of selecting $r = 5$ genes from $k = 3$ gene families (a, b and c) of size $\phi = 4$.

to the case where the window is filled with one gene each from families a and b and three genes from family c.

In order to calculate the probability in Equation 28, it is necessary to enumerate all k-tuples in S. This can be simplified by observing that S can be partitioned into sets of k-tuples that are closed under permutation and such that the summand in the numerator in Equation 28 has the same value for all k-tuples in the same partition. Thus, in order to calculate $p_1()$ it is sufficient to enumerate only one k-tuple from each partition; for example, the set of k-tuples that are non-decreasing. We define the k-tuple $(x_{11}, \ldots, x_{1k})$, such that $x_{11} \leq x_{12} \leq \ldots \leq x_{1k}$, to be the *canonical* k-tuple for the partition that contains it. In the example in Table 1, $\mathcal{S}$ has six elements in two partitions represented by the canonical k-tuples (1,1,3) and (1,2,2), shown in italics. The recursive algorithm in Fig. 1 will generate all such canonical k-tuples for a given windows size r and fixed family size, ϕ. It returns a list of lists of length k, each representing one non-decreasing k-tuple.

The calculation of Equation 28 can be simplified further by observing that, for fixed family sizes, the gene families are indistinguishable. Let z_{1i} be the number of families that contribute i members to W_1; that is, $z_{1i} = |\{x_{1j} \ni x_{1j} = i\}|$. Note that all k-tuples in the same partition contribute z_{11} families with one member to W_1, z_{12} families with two members to W_1 and so on. Then each partition in S can also be represented by a vector, $(z_{11}, \ldots, z_{1\phi})$, where

$$z_{1\phi} + z_{1\phi-1} + \cdots + z_{11} = k \tag{29}$$

and

$$\phi \cdot z_{1\phi} + (\phi\text{–}1) \cdot z_{1\phi-1} + \cdots + 1 \cdot z_{11} = r. \tag{30}$$

There are

$$\frac{k!}{z_{11}! \cdots z_{1\phi}!} \tag{31}$$

k-tuples in that partition. In the example in Table 1, the first and second partitions are represented by the vectors (2,0,1,0) and (1,2,0,0) respectively. There are $3!/(2!1!) = 3$ elements in each partition. Observing that all of the k-tuples in the same partition make the same contribution to the sum in the numerator, Equation 28 can be rewritten as

$$p_1(k) = \binom{n}{r}^{-1} \sum_{\mathcal{Z}} \left[\frac{k!}{z_{11}! \cdots z_{1\phi}!} \binom{\phi}{1}^{z_{11}} \binom{\phi}{2}^{z_{12}} \cdots \binom{\phi}{\phi}^{z_{1\phi}} \right], \tag{32}$$

where $\mathcal{Z}$ is the set of all vectors $(z_{11}, \ldots, z_{1\phi})$, satisfying Equations 29 and 30. The set $\mathcal{Z}$ can be calculated from the canonical k-tuples in $\mathcal{S}$ using the algorithm in Fig. 1 or by finding all solutions to Equations 29 and 30 over the non-negative integers.

The probability that exactly l of those k gene families appear in W_2, simplifies to

$$p_2^o(l) = \frac{\sum_{\zeta=0}^{r-l} \sum_{\mathcal{T}(\zeta)} \binom{\phi}{x_{21}} \binom{\phi}{x_{22}} \cdots \binom{\phi}{x_{2l}} \binom{n_2 - k\phi}{r - \zeta}}{\binom{n_2}{r}}, \tag{33}$$

```
enumerateX(k, r, φ)
{

thislist = an empty list; (* list of partial solutions *)
for (x = min(φ, r-(k-1)) ; x > 0; x--)
{   nextK = k - 1;
    nextR = r - x;
    (* window successfully filled - end recursion
    create new partial solution and return *)
    if (nextR == 0 && nextK == 0) {
       let new = (x) be a new list;
       add new to thislist;
       return(thislist);
    }
    (* if solution still possible, try to fill rest of window *)
    if ( (nextR ≥ nextK) && ((nextK >= 0) && (nextR >= 0)) ) {
       newlist = enumerateX(nextK, nextR, x);
       foreach list l in newlist {
          append x to l;
          append l to thislist;
       }
    }
    (* else, try another value of x *)
}
return(thislist);
}
```

Figure 1: Algorithm to enumerate all ways of filling a window of size r with genes from exactly k families of size ϕ. Returns a list of lists of length k.

where

$$\zeta = \sum_{h=1}^{l} x_{2h}$$

and $\mathcal{T}(\zeta)$ is the set of all l-tuples $(x_{21}, \ldots, x_{2l})$ such that

$$\sum_{h=1}^{l} x_{2h} = r - \zeta,$$
$$0 < x_{2h} \leq \phi.$$

As above, each set $\mathcal{T}(\zeta)$ can be partitioned into subsets that are closed under permutation and the (nondecreasing) canonical l-tuples for each partition may be calculated using the algorithm in Fig. 1, where r is replaced by $r - \zeta$.

Similarly, each partition in $\mathcal{T}(\zeta)$ may be represented by a vector, $(z_{21}, \ldots, z_{2\phi})$, where

$$z_{2i} = |\{x_{2j} \ni x_{2j} = i\}|, \tag{34}$$

$$z_{2\phi} + z_{2(\phi-1)} + \cdots + z_{21} = l \tag{35}$$

and

$$\phi \cdot z_{2\phi} + (\phi\text{–}1) \cdot z_{2\phi-1} + \cdots + 1 \cdot z_{21} = r - \zeta. \tag{36}$$

Since there are

$$\frac{l!}{z_{21}! \cdots z_{2\phi}!} \tag{37}$$

l-tuples in that partition, all making the same contribution to the sum in the numerator, Equation 33 can be rewritten as

$$p_2^o(l) = \frac{\sum_{\zeta=0}^{r-l} \sum_{\mathcal{Z}(\zeta)} \left[\frac{l!}{z_{21}! \cdots z_{2\phi}!} \binom{\phi}{1}^{z_{21}} \binom{\phi}{2}^{z_{22}} \cdots \binom{\phi}{\phi}^{z_{2\phi}} \binom{n_2 - k\phi}{\zeta} \right]}{\binom{n}{r}} \tag{38}$$

where $\mathcal{Z}(\zeta)$, the set of all vectors $(z_{21}, \ldots, z_{2\phi})$, is equivalent to the set of all solutions to Equations 35 and 36 over the positive integers.

Two windows sampled from the same genome: In the paralogous case, the above analysis must be modified to take into account the fact that W_1 and W_2 are two non-overlapping windows sampled from the same genome G. The probability, $p_1(F)$, that W_1 is filled by elements from the k distinct gene families in F is again given by Equation 24. However, the calculation of the probability that at least m of those families appear in W_2, must take into account the fact that gene

family members that appear in W_1 cannot also appear in W_2. In this case, a general expression for $p_2^p()$ may be obtained by renumbering the families in F' to correspond to their order in F:

$$p_2^p(F') = \frac{\sum_{\mathcal{S}(F')} \binom{\phi_{21} - x_{11}}{y_{21}} \binom{\phi_{22} - x_{12}}{y_{22}} \cdots \binom{\phi_{2l} - x_{1l}}{y_{2l}} \binom{n - \sum_{h=1}^{k} \phi_{2h}}{r - \sum_{h=1}^{l} y_{2h}}}{\binom{n-r}{r}}, \tag{39}$$

where $\mathcal{S}(F')$ is the set of all l-tuples $(y_{21}, \ldots, y_{2l})$ such that

$$\sum_{h=1}^{l} y_{2h} \leq r,$$
$$0 < y_{2h} \leq \phi_{2h} - x_{1h}.$$

For gene families of uniform size, the expression for $p_2^p()$ reduces to

$$p_2^p(l) = \frac{\sum_{\mathcal{T}} \binom{\phi - x_{11}}{y_{21}} \binom{\phi - x_{12}}{y_{22}} \cdots \binom{\phi - x_{1l}}{y_{2l}} \binom{n - k\phi}{r - \sum_{h=1}^{l} y_{2h}}}{\binom{n-r}{r}}, \tag{40}$$

yielding

$$q_{WF}^p(m) = \sum_{k=m}^{r} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2^p(l) \right]. \tag{41}$$

Unlike the orthologous case, since the terms in the numerator depend on x_{2j} as well as ϕ and y_{2j}, it is not true that all families that contribute the same number of genes to the window also make the same contribution to the sum in the numerator. Therefore, we cannot pursue the same simplification used in the orthologous case.

3.3 Individual clusters found through whole genome comparison

Orthologous gene clusters can also be found through comparison of whole genomes by designating one genome as the reference genome (without loss of generality, G_1) and considering the set of n–m+1 contiguous runs of m genes in that genome. We define M to be each of these sets of m gene families in turn and search for a cluster in the second genome, G_2. Since this involves $O(n)$ whole genome searches, the significance of a paired cluster found in this fashion is much lower. Here we derive a test based on the probability of observing at least one cluster in such a search, assuming no gene families. Tests for whole genome comparison with gene families are presented in Section 4.

The expected number of those runs that will appear in a window of length r in G_2 is

$$S_R(n, r, m) = (n - m + 1)\, q(n, m, r). \tag{42}$$

In a genome with uniform gene families of size ϕ,

$$S_{RF}(n, r, m) = (n-m+1)\, \phi^m\, q(n, m, r). \tag{43}$$

What is the probability that at least one of those runs will be clustered in the second genome? Let $E_i(m, n, r)$ be the event that the m consecutive genes starting at gene i in G_1 appear in a window of size at most r in the second genome. (Note that $\text{Prob}(E_i) = 0$ if $i > n-m+1$, since there are only n genes in the genome.) Let $E_{i_1,\cdots,i_g}(m, n, r)$ be the event that *all* of the g runs of m consecutive genes in G_1 starting at genes $i_1, \cdots, i_g$, respectively, appear in windows of size at most r in G_2. Note that some of the runs may overlap.

The probability that at least one run of m (or more) consecutive genes in G_1 appears in a window of size at most r in G_2 is $P_R(m, n, r) = \text{Prob}(\cup_{i=1}^{n} E_i)$ and can be calculated using the inclusion-exclusion rule (Equation 7). The dominant term is due to pairs of overlapping windows that share $m-1$ genes. For large n, we may neglect third and higher order terms and even those second-order terms where $i_2 > i_1+1$, yielding the approximation

$$P_R \approx (n-m+1) q(m, n, r) - \text{Prob}(E_{1,2}),$$

where

$$\text{Prob}(E_{1,2}) = \sum_{i=0}^{n-m} E_{i,i+1}.$$

We can calculate $\text{Prob}(E_{1,2})$ exactly, by considering all the ways in which genes $1 \ldots m+1$ can appear in two windows that overlap at $m-1$ positions. Stated formally, we compute the probability of the event that genes $1, ..., m$ appear in a window of size exactly $r_1 \geq m$ and that genes $2, ..., m+1$ appear in a window of size exactly $r_2 \geq m$, where the leftmost positions of the two windows are a_1 and a_2, respectively. Fig. 2 lists all possible configurations for two overlapping windows W_1 and W_2 with endpoints a_1+1, a_1+r_1 and a_2+1, a_2+r_2, respectively, that can satisfy these conditions. Note that the requirement that W_1 and W_2 overlap by $m-1$ positions rules out the two cases where the endpoints of W_1 is completely contained within W_2 and vice versa. Furthermore, an endpoint of W_1 is excluded from W_2 if and only if it is occupied by gene 1. Similarly, an endpoint of W_2 is excluded from W_1 if and only if it is occupied by gene $m+1$.

We will assume n is large enough so that we can neglect partial windows at the ends of the genome. Then we can use the approximation

$$q(n, m, r) \approx m \frac{\binom{r-1}{m-1}}{\binom{n-1}{m-1}}.$$

Let $P_a, P_b, \cdots, P_g$ be the probabilities that the seven configurations in Fig. 2 occur. Then

$$\text{Prob}(E_{1,2}) = \sum_{r_1, r_2 = m}^{r} \left[\sum_{a_1, a_2} P_a + P_b + P_c + P_d + P_e + P_f + P_g \right] \tag{44}$$

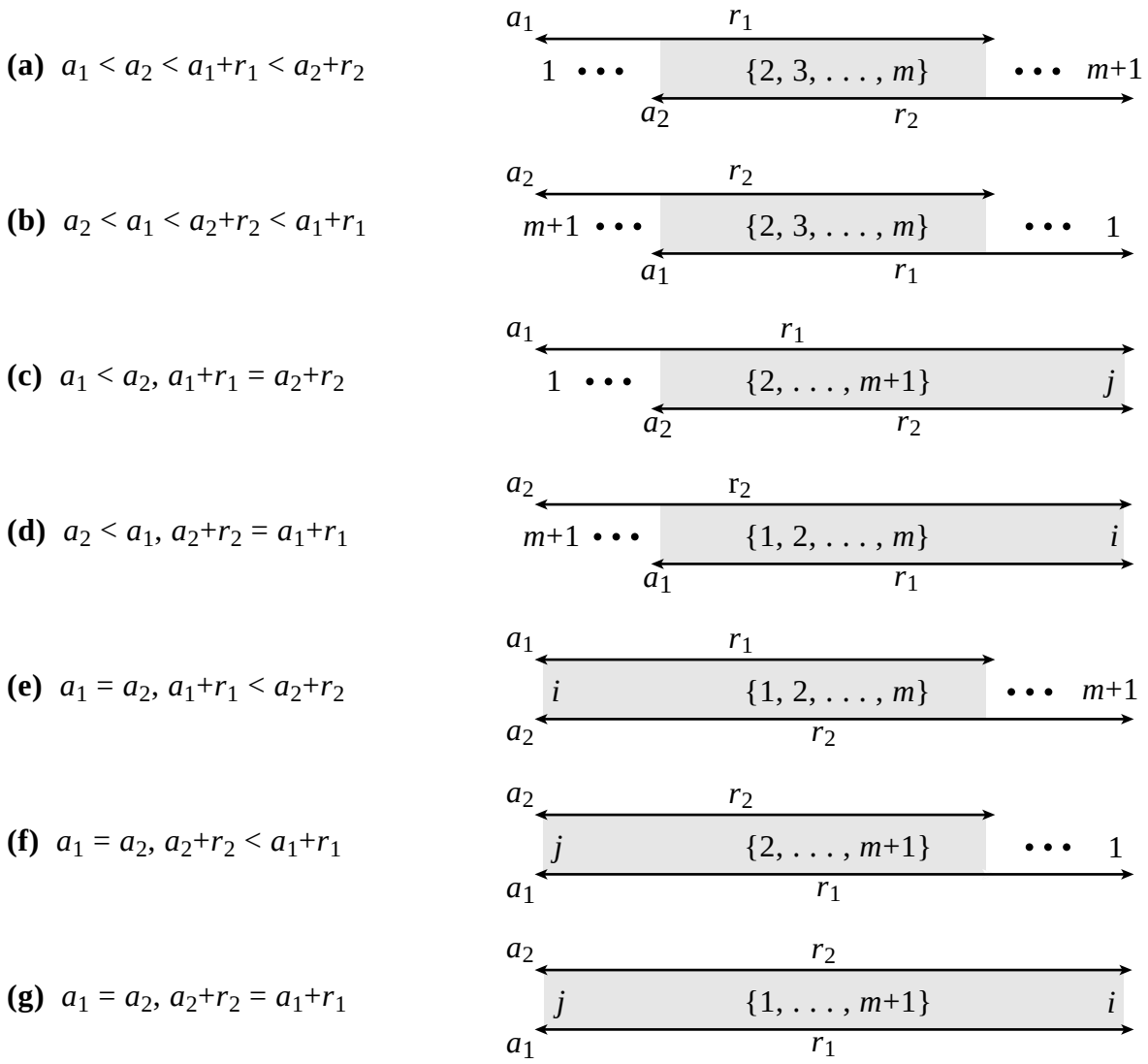


Figure 2: All possible configurations for two overlapping windows W_1 and W_2 such that genes $1, ..., m$ appear in W_1 and genes $2, ..., m+1$ appear in W_2. Genes 1 and $m+1$ must be present where indicated at the endpoints of windows. Genes i and j may be any of the genes within the braces, except gene 1 or gene $m+1$. Except for these constraints, genes within the braces may occur in any order within the overlap between the windows, and may be intermingled with other genes.

What is the probability P_a that the configuration in Fig. 2(a) occurs? Without loss of generality, we denote the position of gene 1 by $a_1 + 1$. Then the probability that gene $m + 1$ appears in position $a_2 + r_2$ is $\frac{1}{n-1}$, again neglecting end effects. The probability that the remaining genes

$\{2, \cdots, m\}$ are contained within the $a_1 + r_1 - a_2$ positions between $a_2 + 1$ and $a_1 + r_1$ is

$$\frac{\binom{a_1 + r_1 - a_2}{m-1}}{\binom{n-2}{m-1}},$$

so that

$$P_a = \frac{1}{n-1} \frac{\binom{a_1 + r_1 - a_2}{m-1}}{\binom{n-2}{m-1}}.$$

For fixed a_1, a_2 and r_2, the sum of terms of form P_a is:

$$\begin{aligned} \sum_{r_1} P_a &= \frac{1}{(n-1)\binom{n-2}{m-1}} \sum_{s=m-1}^{a_1+r-a_2} \binom{s}{m-1} \\ &= \frac{1}{n-1} \frac{\binom{a_1 + r - a_2 + 1}{m}}{\binom{n-2}{m-1}}, \end{aligned} \tag{45}$$

using the upper summation identity for binomial coefficients. The values of r_2 for which this is valid range from $m-1$ to r, so setting $\delta = a_2 - a_1$,

$$\begin{aligned} \sum_{r_1, r_2, \delta} P_a &= \frac{r-m+2}{n-1} \frac{\sum_{\delta=1}^{r-m+1} \binom{r-\delta+1}{m}}{\binom{n-2}{m-1}} \\ &= \frac{r-m+2}{n-1} \frac{\binom{r+1}{m+1}}{\binom{n-2}{m-1}}, \end{aligned} \tag{46}$$

and

$$\sum_{r_1, r_2, a_1, a_2} P_a = \frac{(r-m+2)r(r+1)}{m^2(m-1)} q(n, m, r), \tag{47}$$

assuming about $n - m$ possible positions for a_1.

By symmetry the expression in Equation 47 also holds for $\sum_{r_1,r_2,a_1,a_2} P_b$.

In Fig. 2 (c), given that gene 1 is in position a_1+1, the probability that position a_1+r_1 be occupied by some element j in $\{2,\cdots,m\}$ is $\frac{m-1}{n-1}$, and the probability that $m-1$ of the $a_1+r_1-a_2-1$ positions between a_2+1 and a_1+r_1-1 be occupied by the remaining genes in $\{2,\cdots,m+1\}-\{j\}$ is

$$\frac{\binom{a_1+r_1-a_2-1}{m-1}}{\binom{n-2}{m-1}}.$$

Thus

$$P_c = \frac{m-1}{n-1}\frac{\binom{r_2-1}{m-1}}{\binom{n-2}{m-1}}. \tag{48}$$

The range of r_2 is from m to r_1-1, so that for fixed r_1

$$\sum_{r_2} P_c = \frac{m-1}{n-1}\frac{\binom{r_1}{m}}{\binom{n-2}{m-1}}. \tag{49}$$

The range of r_1 is from $m+1$ to r, so that

$$\begin{aligned}\sum_{r_1,r_2,a_1} P_c &= (m-1)\frac{\left[\binom{r+1}{m+1}-1\right]}{\binom{n-1}{m-1}} \\ &= \frac{(m-1)}{m}\left[\frac{r(r+1)}{m(m+1)}-\frac{1}{\binom{r-1}{m-1}}\right] q(n,m,r),\end{aligned} \tag{50}$$

which also holds for the total contributions of P_d, P_e and P_f.

To calculate P_g, consider the $(m-1)(m-2)$ ordered pairs (j,i) in $\{2,\cdots,m\}$. Without loss of generality, denote the position of gene j by a_1+1. The probability that gene i will appear in position a_1+r_1 is $1/(n-1)$. The probability that all $m-1$ genes in the set $\{1,\cdots,m+1\}-\{i,j\}$ will appear in the r_1-2 positions between a_1+2 and a_1+r_1-1 is

$$\frac{\binom{r_1-2}{m-1}}{\binom{n-2}{m-1}},$$

so that

$$P_g = \frac{(m-1)(m-2)}{n-1} \frac{\binom{r_1-2}{m-1}}{\binom{n-2}{m-1}}. \quad (51)$$

Summing over the range of r_1 from $m+1$ to r, and all positions of a_1, we have

$$\sum_{r_1,a_1} P_g = \frac{(n-m)(m-1)(m-2)}{(n-1)} \frac{\binom{r-1}{m}}{\binom{n-2}{m-1}}$$
$$= \frac{(r-m)(m-1)(m-2)}{m^2} q(n,m,r). \quad (52)$$

Collecting terms, Equation 44 becomes

$$\text{Prob}(E_{1,2}) = 2 \sum_{r_1,r_2,a_1,a_2} P_a + 4 \sum_{r_1,r_2,a_1} P_c + \sum_{r_1,a_1} P_g,$$

which may be calculated rapidly with the help of the approximation in Equation 3.

4 Whole genome comparison

The advent of comparative maps also introduces the question of the significance of multiple shared clusters in the context of whole genome comparison. When comparing entire genomes, how many pairs of homologous clusters should we expect to find by chance alone? Aggregate clustering properties have been used to study the functional and evolutionary implications of large-scale genomic organization, including rates of rearrangement (Ehrlich et al. 1997; McLysaght et al. 2000; Seoighe and others 2000; Seoighe and Wolfe 1998), the distribution of breakpoints (Coghlan and Wolfe 2002; Nadeau and Taylor 1984; Nadeau and Sankoff 1998b), conservation of gene order (Tamames 2001), and the duplication processes (e.g., tandem duplication, whole genome duplication, duplication of subchromosomal segments) that dominate in a given lineage (Arabidopsis Genome Initiative 2000; Chervitz et al. 1998; Friedman and Hughes 2001; McLysaght et al. 2002; Semple and Wolfe 1999; Vision et al. 2000). In order to interpret such data correctly, the significance of observing a certain number of shared clusters must be determined. The quantities derived here can also be used to test the significance of individual clusters found through whole genome comparison.

4.1 Comparing two different genomes

Consider two genomes, G_1 and G_2, from different species that share a certain number of gene clusters. We define a paired cluster to be a set of m genes observed in two windows of length at

most r, one in G_1 and one in G_2. In a genome with no gene families ($\phi_j = 1, \forall j$), the expected number of such paired clusters is

$$S_C^o(n, m, r) = \binom{n}{m} q(n, m, r)^2, \tag{53}$$

where $q()$ is defined in Equation 2.

In the case of general gene families where $\phi_j \geq 1, \forall j$, the expected number of paired clusters found when comparing G_1 and G_2 is

$$S_F^o(n, m, r) = \left[\sum_{M \in \mathcal{F}^m} \Phi_1(M)\Phi_2(M) \right] q(n, m, r)^2, \tag{54}$$

where $\mathcal{F}^m$ is the set of all sets of m distinct gene families. For gene families of uniform size, ϕ, this expression simplifies to

$$S_F^o(n, m, r, \phi) = \binom{n_f}{m} [\phi^m q(n, m, r)]^2, \tag{55}$$

where n_f is the number of gene families in both G_1 and G_2.

Window sampling: While this expression provides a measure of the degree of shared clustering between G_1 and G_2, it is not a convenient basis for data analysis because it requires enumerating all paired clusters. An alternate approach, based on sampling windows from the genome at random, may be preferable. In genomes with no gene families, given a random sample of n_w pairs of windows, such that no window in the sample overlaps with any other window in the sample, the expected number of pairs that share at least m genes is

$$S_W^o = n_w q_W^o(n, m, r), \tag{56}$$

where $q_W^o(n, m, r)$ is the probability, given in Equation 22, that a pair of windows of length r, one from each genome, share at least m homologous gene pairs. Given a random sample of non-identical, but possibly overlapping, windows, the above expressions can be used to estimate the expected number of pairs that share m homologous pairs, since the fraction of overlapping pairs is $\mathcal{O}(n^{-1})$, when $r \ll n$. The probability of finding at least one pair of windows in the sample that share at least m genes can be approximated by the equation

$$P_W^o(n_w, n, m, r) \approx 1 - [1 - q_W^o(n, m, r)]^{n_w}. \tag{57}$$

but since it is based on an unwarranted assumption that the events of finding clusters in the various pairs of windows are independent, it provides only a rough estimate. If G_1 and G_2 have gene families, the expected number of window pairs that have m gene families in common can be obtained by substituting $q_{FW}^o(n, m, r)$, given in Equations 23 and 27, for $q_W^o(n, m, r)$ in Equation 56. Similarly, the probability of observing at least one such window pair in two random genomes with gene families can be estimated by substituting $q_{FW}^o(n, m, r)$ for $q_W^o(n, m, r)$ in Equation 57.

4.2 Genome self-comparison

Clusters of paralogs in the same genome are often presented as evidence of whole genome duplication or duplication of large sub-chromosomal segments. The goal in genome self-comparison is to determine the degree of clustering among duplicated genes. As above, we designate by $\mathcal{F}^m$ the set of all distinct sets of m gene families in G. Let $M \in \mathcal{F}^m$ be a particular set of m different gene families. The total number of pairs of non-intersecting sets of m genes, one from each family in M, is

$$\Psi(M) = \prod_{f_j \in M} \binom{\phi_j}{2}.$$

In the paralogous case, we define a paired cluster to be such a pair of non-intersecting sets found in two, possibly overlapping, windows of length at least r in the same genome. The expected number of paired clusters is then

$$S_F^p(n, m, r) = \left[\sum_{M \in \mathcal{M}} \Psi(M) \right] q(n, m, r)^2. \tag{58}$$

If all gene families have exactly ϕ members, then

$$S_F^p(n, m, r, \phi) = \binom{n_f}{m} \binom{\phi}{2}^m q(n, m, r)^2. \tag{59}$$

Window sampling: The degree of clustering of duplicated genes in a genome can also be estimated by counting the number of pairs of windows that share a given number of gene families. Given a random sample of n_w pairs of non-overlapping windows taken from G, the expected number of pairs that have m gene families in common is

$$S_{FW}^p(n_w, n, m, r) = n_w \, q_{FW}^p(n, m, r), \tag{60}$$

where $q_{FW}^o(n, m, r)$ is defined in Equation 23.

$$P_{FW}^p(n_w, n, m, r) \approx 1 - [1 - q_{FW}^p(n, m, r)]^{n_w} \tag{61}$$

yields a rough approximation for the probability of finding at least one such pair.

5 Previous Work

In their analysis of the significance of conserved synteny, Trachtulec and Forejt (2001) estimate the probability of finding m genes in a window of exactly r slots by chance to be $(r/n)^{(m-1)}$, where n is the number of genes in the genome. If $r \ll n$, for a range of values of m this formula approximates our exact expression in Equation 1.

As part of their analysis of gene duplication in the human genome, Venter *et al.* (2001) suggest that the probability of a fixed set of m genes occurring in a given order within an interval of r successive gene positions in a random genome of length n is

$$u_1(n,m,r) = \frac{\sum_{i=m-2}^{r-2} \binom{i}{m-2}}{n^{m-1}} \tag{62}$$

For a large genome, where $n^{m-1} \approx (n\text{–}1)!/(n\text{–}m)!$, and neglecting end effects (or assuming a circular genome), Equation 62 is essentially correct. An exact expression for this quantity is $u(n,m,r) = q(n,m,r)/m!$, where $q(n,m,r)$ is defined in Equation 2

They further consider the case of two sets of m genes that are pairwise paralogous and state a probability "allowing for" the two sets "to be spread across r positions" in two separate locations:

$$u_2(n,m,r) = \frac{\left[\sum_{i=m-2}^{r-2} \binom{i}{m-2}\right]^2}{n^{m-1}} \tag{63}$$

However, it is not clear what event has this probability, even approximately. Indeed, for $m = 3$ and $r = \frac{n}{2}$, for example, Equation 63 is $O(n^2)$ and thus cannot be a probability.

6 Application to biological data

There is a broad literature in which gene cluster analysis has been used to interpret the evolutionary or functional implications of gene order in species ranging from viruses and bacteria to mammals, based on data derived from both whole genome sequencing and linkage mapping. To show the utility of the models developed in the previous sections, we apply our results to a few examples from this literature. Our intent here is not to reanalyze the data or question the conclusions of the studies cited below, but rather to provide concrete examples of how our models can be put to practical use in real biological studies.

Individual clusters: Since Ohno (1970) first hypothesized two whole genome duplications in early vertebrates, the role of large scale duplication in vertebrate evolution has been much debated (Durand 2003; Hughes et al. 2001; Hughes 1999; McLysaght et al. 2002; Sankoff 2001; Skrabanek and Wolfe 1998; Wolfe 2001). One type of evidence that is offered in these debates is the presence of linkage groups that appear to be duplicated and also to be conserved across several species.

At least a dozen papers analyzing such regions, summarized in Table 2, have appeared in the last decade. These clusters typically contain five to fifteen genes spread over a window of 15 to 100 slots. Are conserved clusters of this sparsity truly significant? Fig. 3 shows the expected number of clusters found in the self-comparison of a random genome of size $n = 3000$, calculated using Equation 59. The parameter n refers to the number of genes in the data set not the number of genes in the organism. Since these studies were performed on linkage data, n is chosen to reflect the number of mapped, sequenced genes in the organism in the Jackson Laboratories Mouse Genome

Region	*Gene families found in region*	*References*
MHC	*Abc*, *C3/4/5*, *Col*, *Hsp*, *Notch*, *Pbx*, *Psmb*, *Rxr*, *Ten*	Endo et al. 1997; Hughes 1998; Kasahara 1997; Katsanis et al. 1996; Smith et al. 1999; Spring 2002; Trachtulec and Forejt 2001
HOX	*Achr*, *Ccnd*, *Cdc*, *Cdk*, *Dlx*, *En*, *Evx*, *Gli*, *Hh*, *Hox*, *If*, *Inhb*, *Nhr*, *Npy/Ppy*, *Wnt*	Amores et al. 1998; Hughes 1998; Spring 2002
FGR	*Adr*, *Ank*, *Egr*, *Fgfr*, *Lpl*, *Pa*, *Slc4A*, *Vmat*	Coulier et al. 1997; Lipovich et al. 2001; Lundin 1993; Pebusque et al. 1998; Spring 2002
TBOX	*Cryb*, *Lhx*, *Nos*, *Tbx*, *Tcf*, *Prkar*	Ruvinsky and Silver 1997
MATN	*Eya*, *Hck*, *Matn*, *Myb* , *Myc*, *Sdc*, *Src*	Gibson and Spring 2000

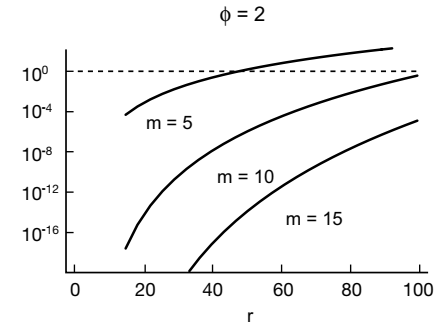
Table 2: Paralogous gene clusters in vertebrate genomes recently reported in the literature. Many of these clusters appear in several vertebrate species and have also been found in invertebrate genomes.

Database (MGD, *http://www.informatics.jax.org*) at the end of the twentieth century. The curves in Fig. 3(a) suggest that when gene family sizes are small ($\phi = 2$), a cluster larger than ten is significant even if spread over a large window. However, as ϕ increases, clusters found in larger windows are no longer significant (Fig. 3(b)).

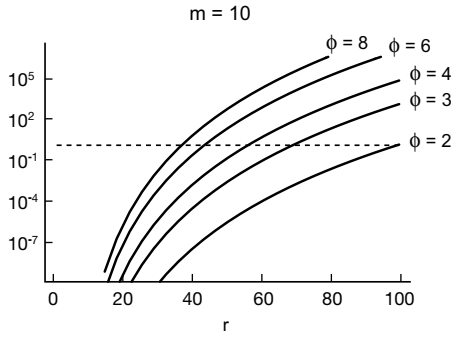
Let us consider one of these examples, the TBOX cluster, in detail. Ruvinsky and Silver (1997) observed paralogous gene clusters on mouse chromosomes 5 and 11, shown in Fig. 4, and explored the hypothesis that these genes were duplicated in a single event. The central cluster is quite compelling but it is more difficult to decide whether the more distant *Prkar* paralogs should be included in this candidate duplicated region. A statistical test using the reference region approach can help resolve this question. In a data set of 2888 mapped genes extracted from MGD, genes from the *Cryb*, *Lhx*, *Nos*, *Tbx* and *Tcf* families were found in a window of 15 slots on chromosome 5 and a window of 48 slots on chromosome 11. The inclusion of the *Prkar* genes, yields a cluster of seven genes in windows of 47 and 65 slots, respectively. Using the expression for $S_{\Phi H}()$ given in Table 4, we estimate the expected number of clusters with these values in a random genome with uniform gene family size, $\phi = 3$. If we take the six-gene cluster on chromosome 5 as the reference, assuming $\phi = 3$, the expected number of such clusters in a random genome is $S_{FH}(2888, 6, 15, 48, 3) = 2.0 \times 10^{-3}$, suggesting that the six-gene cluster is significant. Adding *Prkar1b* to the reference cluster, yields $h = 7$, $m = 47$ and $r = 65$, and the expected number of clusters becomes 5.7. In this case, it is no longer possible to reject chance as a possible explanation for the seven-gene cluster with confidence. Moreover, if we select chromosome 11 as the reference, then the expected numbers of comparable six- and seven-gene clusters in a random genome are 6.1×10^{-3} and 8.3, respectively, leading to the same conclusions.

Whole genome analysis: In large scale studies of conserved regions, the intent is to char-

acterize processes of duplication, rearrangement and conservation on a genome wide scale rather than detailed study of a particular region. In one example of such an analysis, Tamames (2001) compared pairs of bacterial genomes in a study of gene order conservation in prokaryotes. His approach uses a parameterized method for identifying pairs of runs of orthologs (one in each species), in which the user must specify two parameters: m_0, the minimum number of pairs of orthologs in the run and, g, the maximum number of intruders found between any pair of orthologs in the run. Thus, a run with m orthologs can be at most $(g+1)(m-1)+1$ slots long. For the purposes of the study (Tamames 2001), m_0 and g were both set to three. Our model provides a rational basis for selecting parameters of clustering algorithms such as this. In this example, we seek the minimum m_0 and maximum g such that the clusters obtained are still significant. Figs. 5 and 6 show the expected number of clusters with m gene families in common, calculated using Equation 55. Since bacterial genomes range from roughly 500 to 7000 genes, an intermediate size of $n = 3000$ was



(a)



(b)

Figure 3: Expected number of paralogous clusters in a random genome with $n = 3000$ genes as a function of the window size, r. The threshold, $S_F^p() = 1$, is shown as a dashed line. **(a)** $\phi = 2$. Cluster size ranges from $m = 5$ to $m = 15$. **(b)** $m = 10$. Gene family size ranges from $\phi = 2$ to $\phi = 8$.

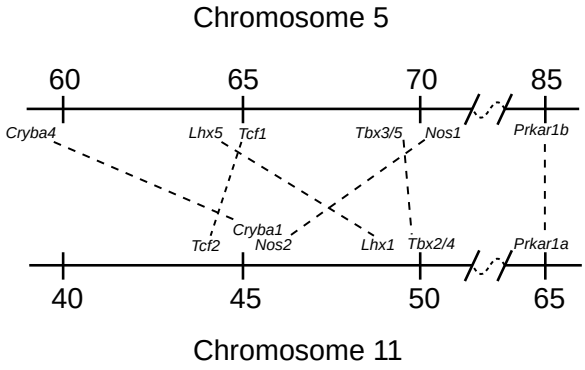


Figure 4: Clusters of paralogous mouse genes on chromosomes 5 and 11. Adapted from Ruvinsky and Silver (1997).

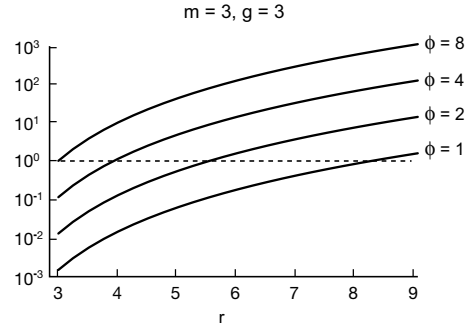
used. Fig. 5(a) shows that with a gap of three, clusters of three genes are significant when there are no gene families but cease to be for $\phi \geq 2$. When $m = 9$, however, most clusters are significant except when $\phi \geq 4$ and r approaches its maximum range (Fig. 5(b)). The dependence of cluster significance on gap size is demonstrated in Fig. 6. When $\phi = 4$, clusters in windows of maximum size with $g = 3$ are not significant for any value of m, no matter how large. However, if $g = 2$ clusters are significant for $m \geq 6$ and for $g = 1$ most clusters are significant. In the absence of additional biological information that can be used to determine cluster significance, such as gene orientation, these results suggest that a slightly higher value of m_0 and a gap size of $g \leq 2$ would guarantee the significance of clusters found with this algorithm. This example demonstrates that statistical models of gene clusters are useful not only in data analysis but also in algorithm design.

7 Discussion and Future Work

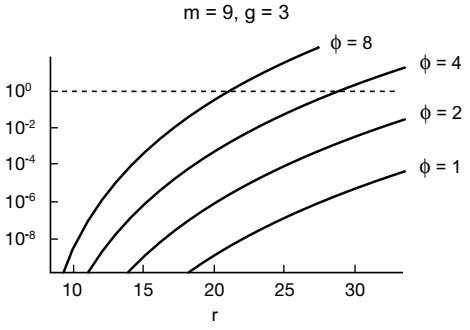
Gene clusters, local similarities in genome organization, are the basis of many studies of genome evolution and function in organisms ranging from bacteria to vertebrates. We have presented probabilistic models for determining the significance of gene clusters in both paralogous and orthologous settings. Under a model of uniform random gene order, we consider the probability of finding a cluster of a particular set of genes, as well as the expected number of clusters observed in whole genome comparison. Many factors have an impact on cluster significance. Clusters with conserved gene order are more significant than those with random gene order. Significance decreases in genomes with gene families or if some of the genes in the cluster are missing. A cluster may be more significant if observed in multiple genomes, but the calculation of significance must take the number of genomes considered into account. Finally, cluster significance is influenced by the number of instances considered when searching for clusters. The interplay of these forces is subtle and cannot be easily intuited, underscoring the importance of formal statistical models such as those presented here. Our models take into account multiple genomes, gene families, incomplete clusters and the approach that was used to find the clusters. Despite a fairly simple and abstract

model, we have demonstrated that our results are practical by applying our models to a to range of examples from the biology literature.

In future work, we plan to develop more detailed, biologically motivated models. The current model treats the genome as an ordered set of genes. An extended analysis would model the chromosomal positions of genes, and would take tandem duplications and gene rich and gene poor regions into account. A parameterized model of gene family sizes that yields realistic, computationally tractable approximations is also needed. Finally, other types of biological information besides gene order can be brought to bear on the assessment of significance including gene orientation (e.g., Tamames 2001; Wolfe and Shields 1997) and divergence times (e.g., Chen et al. 2000; Friedman and Hughes 2001; Ruvinsky and Silver 1997.)



(a)



(b)

Figure 5: Expected number of orthologous clusters of m genes in a window of size r, where r ranges from m to $(g{+}1)(m{-}1){+}1$, $n = 3000$, $g = 3$ and gene family size ranges from $\phi = 1$ to $\phi = 8$. The threshold, $S^o_F() = 1$, is shown as a dashed line. **(a)** m = 3. **(b)** m = 9.

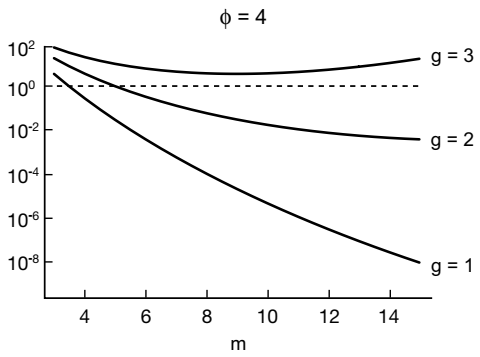


Figure 6: Expected number of orthologous clusters of m genes in a window of size $r = (g{+}1)(m{-}1){+}1$, where $\phi = 4$ and $n = 3000$. Gap sizes range from $g = 1$ to $g = 3$. The threshold, $S^o_F() = 1$, is shown as a dashed line.

8 Acknowledgments

Thanx and a tip o' the hat to Rose Hoberman and Louxin Zhang for carefully reading and correcting errors in this manuscript. D.D. was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship. D.S. was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. He holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

A Appendix: Test statistics

Test statistics derived in this paper are summarized in Tables 3 – 7. These include statistics for general models of gene families ($\phi_j \geq 1, \forall j$), genomes with uniform size gene families ($\phi_j = \phi, \forall j$) and genomes with no gene families ($\phi_j = 1, \forall j$). Table 3 gives statistics for individual clusters for the case where the clusters are found through comparison with a reference region and the reference and matching regions are located in different genomes. Statistics for the case where both regions are in the same genome are given in Table 4. Table 5 gives statistics based on the expected number of clusters seen during whole genome comparison, including comparison with a reference genome, comparison of two different genomes and paralogous genome self-comparison. Test statistics based on window sampling, for both individual clusters and genome-wide clustering, are given in Tables 6 and 7. Table 6 gives statistics for comparison of two different genomes while Table 7 addresses comparison of a genome with itself.

Test statistic		Eqn	Gene Families
$q()$	$\frac{\left[(n-r)\binom{r-1}{m-1}\right]+\binom{r}{m}}{\binom{n}{m}}$	2	$\phi_j = 1$
Probability of finding m prespecified genes in a window of size r.			
$S_\phi()$	$\Phi(M)q(n,m,r)$	5	$\phi_j \geq 1$
Expected number of clusters with gene families.			
$S_H()$	$\binom{m}{h} q(n,h,r)$	11	$\phi_j = 1$
$S_{\Phi H}()$	$\left(\sum_{H\in\mathcal{H}} \Phi(H)\right) q(n,h,r)$	17	$\phi_j \geq 1$
$S_{\Phi H}()$	$\binom{m}{h} \phi^h q(n,h,r)$	18	$\phi_j = \phi$
Expected number of incomplete clusters.			
$P_H()$	$m\, q_{HW}(n, h–1, m–1, r–1)$	14	$\phi_j = 1$
Upper bound on the probability of at least one incomplete cluster.			

Table 3: Test statistics for clusters orthologous to a reference region

Test statistic		Gene Families
$S_\phi()$	$\Phi(M)q(n-m,m,r)$	$\phi_j = 2$
Expected number of clusters.		
$S_{\Phi H}()$	$\left(\sum_{H\in\mathcal{H}} \Phi(H)\right) q(n-m,h,r)$	$\phi_j \geq 2$
$S_{\Phi H}()$	$\binom{m}{h} (\phi-1)^h q(n-m,m,r)$	$\phi_j = \phi$
Expected number of incomplete clusters.		

Table 4: Test statistics for clusters paralogous to a reference region. In the paralogous case, $\Phi(M)$ is defined to be $\Pi_{j\in M}(\phi_j - 1)$. These statistics are modified versions of the orthologous statistics and do not refer to equations in the text.

Test statistic		Eqn	Gene Families
$S_R()$	$(n-m+1)\, q(n,m,r)$	42	$\phi_j = 1$
$S_{RF}()$	$(n-m+1)\, \phi^m\, q(n,m,r)$	43	$\phi_j = \phi$
Reference genome.			
$S^o_C()$	$\binom{n}{m} q(n,m,r)^2$	53	$\phi_j = 1$
$S^o_F()$	$\binom{n_f}{m} [\phi^m q(n,m,r)]^2$	55	$\phi_j = \phi$
Orthologous whole genome comparison.			
$S^p_F()$	$\binom{n_f}{m} \binom{\phi}{2}^m q(n,m,r)^2$	59	$\phi_j = \phi$
Genome self-comparison.			

Table 5: Expected number of clusters in whole genome comparisons

Test statistic		Eqn	Gene Families
$q^o_W()$	$\sum_{i=m}^{r} \frac{\binom{r}{i}\binom{n-r}{r-i}}{\binom{n}{r}}$	22	$\phi_j = 1$
$q^o_{WF}()$	$\sum_{k=m}^{r} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2^o(l) \right]$	27	$\phi_j = \phi$
Probability that two windows share m families.			
$p_1()$	$\binom{n}{r}^{-1} \sum_{\mathcal{Z}} \left[\frac{k!}{z_{11}! \cdots z_{1\phi}!} \binom{\phi}{1}^{z_{11}} \cdots \binom{\phi}{\phi}^{z_{1\phi}} \right]$	32	$\phi_j = \phi$
Probability that k gene families appear in W_1			
$p_2^o()$	$\frac{\sum_{\zeta=0}^{r-l} \sum_{\mathcal{Z}(\zeta)} \left[\frac{l!}{z_{21}! \cdots z_{2\phi}!} \binom{\phi}{1}^{z_{21}} \cdots \binom{\phi}{\phi}^{z_{2\phi}} \binom{n_2 - k\phi}{\zeta} \right]}{\binom{n}{r}}$	38	$\phi_j = \phi$
Probability that exactly l of those k gene families appear in W_2.			
$S^o_W()$	$n_w q^o_W(n, m, r)$	56	$\phi_j = 1$
$S^o_{WF}()$	$n_w q^o_{WF}(n, m, r)$	56	$\phi_j = \phi$
Expected number of window pairs that share at least m genes.			
$P^o_W()$	$\sim 1 - [1 - q^o_W(n, m, r)]^{n_w}$	57	$\phi_j = 1$
$P^o_{WF}()$	$\sim 1 - [1 - q^o_{WF}(n, m, r)]^{n_w}$	57	$\phi_j = \phi$
Probability that at least one pair shares at least m genes.			

Table 6: Test statistics for window sampling, orthologous case

Test statistic		Eqn	Gene Families
$q^p_{WF}()$	$\sum_{k=m}^{r} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2(l) \right]$	41	$\phi_j = \phi$
Probability that two windows share m families			
$p_1()$	$\binom{n}{r}^{-1} \sum_{\mathcal{Z}} \left[\frac{k!}{z_{11}! \cdots z_{1\phi}!} \binom{\phi}{1}^{z_{11}} \cdots \binom{\phi}{\phi}^{z_{1\phi}} \right]$	32	$\phi_j = \phi$
Probability that k gene families appear in W_1			
$p^p_2()$	$\frac{\sum_{\mathcal{T}} \binom{\phi - x_{11}}{y_{21}} \cdots \binom{\phi - x_{1l}}{y_{2l}} \binom{n - k\phi}{r - \sum_{h=1}^{l} y_{2h}}}{\binom{n-r}{r}}$	40	$\phi_j = \phi$
Probability that exactly l of those k gene families appear in W_2*.*			
$S^p_{WF}()$	$n_w \, q^p_{FW}(n, m, r)$	60	$\phi_j = \phi$
Expected number of window pairs that share at least m genes			
$P^p_{FW}()$	$\sim 1 - [1 - q^p_{FW}(n_p, m, r)]^{n_w}$	61	$\phi_j = \phi$
Probability that at least one pair shares at least m genes.			

Table 7: Test statistics for window sampling, paralogous case

References

Adams MD, et al (2000). The genome sequence of *Drosophila melanogaster*. Science 287(5461):2185–9.

Amores A, Force A, l. Yan Y, Joly L, Amemiya C, Fritz A, Ho R, et al (1998). Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Bansal AK (1999). An automated comparative analysis of 17 complete microbial genomes. Bioinformatics 15:900–908.

Bergeron A, Corteel S, Raffinot M (2002). The Algorithmic of Gene Teams. In D. Gusfield and R. Guigo (Eds.), *Algorithms in Bioinformatics, Second International Workshop WABI2002*, Lecture Notes in Computer Science 2452, pp. 464–476.

Blanchette M, Kunisawa T, Sankoff D (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. Journal of Molecular Evolution 49:193–203.

Bork P, Snel B, Lehmann G, Suyama M, Dandekar T, Lathe III W, Huynen M (2000). Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 281–294. Dordrecht, NL: Kluwer Academic Press.

Chen K, Durand D, Farach-Colton M (2000). Notung: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. Journal of Computational Biology 7(3/4):429–447.

Chervitz S, Aravind L, Sherlock G, Ball C, Koonin E, Dwight S, Harris M, et al (1998). Comparison of the complete protein sets of worm and yeast: Orthology and divergence. Science 282:2022–2028.

Coghlan A, Wolfe KH (2002). Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in *Drosophila*. Genome Research 12(6):857–867.

Cosner ME, Jansen RK, Moret BME, Raubeson LA, Wang LS, Warnow T, Wyman S (2000). An empirical comparison of phylogenetic methods on chloroplast gene order data in *Campanulaceae*. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 99–121. Dordrecht, NL: Kluwer Academic Press.

Coulier F, Pontarotti P, Roubin R, Hartung H, Goldfarb M, Birnbaum D (1997). Of Worms and Men: An Evolutionary Perspective on the Fibroblast Growth Factor (FGF) and FGF Receptor Families. J. Mol Evol 44:43–56.

Durand D (2003). Vertebrate Evolution: Doubling and Shuffling with a Full Deck. Trends in Genetics 19(1):2–5.

Durand D, Sankoff D (2002). Tests for gene clustering. In *RECOMB2002, Sixth Annual International Conference on Computational Molecular Biology*, pp. 144–154.

Ehrlich J, Sankoff D, Nadeau J (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics 147(1):289–96.

El-Mabrouk N, Nadeau JH, Sankoff D (1998). Genome Halving. In Springer-Verlag (Ed.), *Combinatorial Pattern Matching*, pp. 235–250.

El-Mabrouk N, Sankoff D (2003). The reconstruction of doubled genomes. SIAM Journal of Computing 32:754–792.

Endo T, Imanishi T, Gojobori T, Inoko H (1997). Evolutionary significance of intra-genome duplications on human chromosomes. Gene 205(1–2):19–27.

Ermolaeva MD, White O, Salzberg S (2001). Prediction of operons in microbial genomes. Nucleic Acids Res 5(29):1216–1221.

Friedman R, Hughes A (2001). Gene Duplication and the Structure of Eukaryotic Genomes. Genome Research 11:373–381.

Gibson T, Spring J (2000). Evidence in favour of ancient octaploidy in the vertebrate genome. Biochem Soc Trans 2:259–264.

Goldberg D, McCouch S, Kleinberg J (2000). Algorithms for Constructing Comparative Maps. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 281–294. Dordrecht, NL: Kluwer Academic Press.

Goldberg LA, Goldberg PW, Paterson MS, Pevzner P, Sahinalp SC, Sweedyk E (2001). The Complexity of Gene Placement. Journal of Algorithms 41(2):225–2435.

Hannenhalli S, Chappey C, Koonin EV, Pevzner PA (1995). Genome Sequence Comparison and Scenarios for Gene Rearrangements: A Test Case. Genomics 30:299 – 311.

Heber S, Stoye J (2001a). Algorithms for finding gene clusters. In *Proceedings of WABI01*, Lecture Notes in Computer Science 2149, pp. 254–265.

Heber S, Stoye J (2001b). Finding all common intervals of k permutations. In *Proceedings of CPM01*, Lecture Notes in Computer Science 2089, pp. 207–218.

Hughes A, da Silva J, Friedman R (2001). Ancient Genome Duplications Did Not Structure the Human Hox-Bearing Chromosomes. Genome Research 11:771–780.

Hughes AL (1998). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. MBE 15(7):854–70.

Hughes AL (1999). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J. Mol Evol 48(5):565–76.

Huynen M, Bork P (1998). Measuring genome evolution. Proc Natl Acad Sci U S A 95:5849–56.

Huynen MA, van Nimwegen E (1998). The Frequency Distribution of Gene Family Sizes in Complete Genomes. Mol. Biol. Evol. 15(5):583–589.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409(682):860–921.

Kasahara M (1997). New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. Hereditas 127(1–2):59–65.

Katsanis N, Fitzgibbon J, Fisher E (1996). Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. Genomics 35(1):101–8.

Kolsto AB (1997). Dynamic bacterial genome organization. Molecular Microbiology 24:241–8.

Kumar S, Gadagkar SR, Filipski A, Gu X (2001). Determination of the Number of Conserved Chromosomal Segments Between Species. Genetics 157:1387–1395.

Lawrence J, Roth JR (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–60.

Li WH, Gu Z, Wang H, Nekrutenko A (2001). Evolutionary analyses of the human genome. Nature 409:847–9.

Lipovich L, Lynch ED, Lee MK, King MC (2001). A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: *SLC4A9* at 5q31. Genome Biology 2(4):0011.1–0011.13.

Lundin LG (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16(1):1–19.

McLysaght A, Hokamp K, Wolfe KH (2002). Extensive genomic duplication during early chordate evolution. Nat Genet. 31(2):200–204.

McLysaght A, Seoighe C, Wolfe KH (2000). High frequency of inversions during eukaryote gene order evolution. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 281–294. Dordrecht, NL: Kluwer Academic Press.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Nadeau J, Sankoff D (1998a). Counting on comparative maps. Trends Genet 14(12):495–501.

Nadeau J, Sankoff D (1998b). The lengths of undiscovered conserved segments in comparative maps. Mamm Genome 9(6):491–5.

Nadeau JH, Sankoff D (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147:1259–1266.

Nadeau JH, Taylor BA (1984). Lengths of Chromosomal Segments Conserved since the Divergence of Man and Mouse. Proc.Natl.Acad.Sci. USA 81:814–818.

O'Brien SJ, Wienberg J, Lyons LA (1997). Comparative genomics: lessons from cats. Trends Genet 10(13):393–399.

Ohno S (1970). Evolution by Gene Duplication. Springer-Verlag.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999). The use of gene clusters to infer functional coupling. PNAS 96:2896–2901.

Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. MBE 15(9):1145–59.

Pevzner PA (2000). Computational Molecular Biology: An Algorithmic Approach. Cambridge, MA: MIT Press.

Rubin GM, et al (2000). Comparative genomics of the eukaryotes. Science 287(5461):2204–2215.

Ruvinsky I, Silver LM (1997). Newly Indentified Paralogous Groups on Mouse Chromosomes 5 and 11 Reveal the Age of a T-Box Cluster Duplication. Genomics 40:262–266.

Sankoff D (2001). Gene and genome duplication. Current Opinion in Genetics and Development 11:681–4.

Sankoff D (2002). Short inversions and conserved gene clusters. Bioinformatics 18:1305–1308.

Sankoff D, Bryant D, Deneault M, Lang BF, Burger G (2000a). Early eukaryote evolution based on mitochondrial gene order breakpoints. J Comput Biol 3–4:521–535.

Sankoff D, Deneault M, Bryant D, Lemieux C, Turmel M (2000b). Chloroplast gene order and the divergence of plants and algae from the normalized number of induced breakpoints. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 89–98. Dordrecht, NL: Kluwer Academic Press.

Sankoff D, El-Mabrouk N (2002). Genome rearrangement. In T. Jiang, T. Smith, Y. Xu, and M. Zhang (Eds.), *Current Topics in Computational Biology*, pp. 135–155. MIT Press.

Sankoff D, Ferretti V, Nadeau JH (1997). Conserved Segment Identification. Journal of Computational Biology 4:559–565.

Semple C, Wolfe KH (1999). Gene duplication and gene conversion in the *Caenorhabditis elegans* Genome. JME 48(5):555–64.

Seoighe C, et al (2000). Prevalence of short inversions in yeast genome evolution. PNAS 97:14433–14437.

Seoighe C, Wolfe K (1998). Extent of genomic rearrangement after genome duplication in yeast. Proc Natl Acad Sci U S A 95(8):4447–52.

Seoighe C, Wolfe KH (1999). Updated map of duplicated regions in the yeast genome. Gene 238:253–261.

Skrabanek L, Wolfe K (1998). Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev 8(6):559–565.

Smith NGC, Knight R, Hurst LD (1999). Vertebrate genome evolution: a slow shuffle or a big bang. BioEssays 21:697–703.

Snel B, Lehmann G, Bork P, Huynen MA (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28(18):3442–4.

Spring J (2002). Genome duplication strikes back. Nature Genetics 31:128–129.

Suyama M, Bork P (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 1(17):10–13.

Tamames J (2001). Evolution of gene order conservation in prokaryotes. Genome Biol 6(2):0020.1–11.

Tamames J, Casari G, Ouzounis C, Valencia A (1997). Conserved clusters of functionally related genes in two bacterial genomes. JME 44::66–73.

Tamames J, Gonzalez-Moreno M, Valencia A, Vicente M (2001). Bringing gene order into bacterial shape. Trends Genet 3(17):124–126.

Tatusov RL, Koonin EV, Lipman D (1997). A Genomic Perspective on Protein Families. Science 278:631–637.

Tiuryn J, Radomski JP, Slonimski PP (2000). A formal model of genomic DNA multiplication and amplification. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 503–5013. Dordrecht, NL: Kluwer Academic Press.

Trachtulec Z, Forejt J (2001). Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. Mamm Genome 3(12):227–231.

Venter JC, et al (2001). The Sequence of the Human Genome. Science 291(5507):1304–51.

Vision TJ, Brown DG, Tanksley SD (2000). The origins of genomic duplications in Arabidopsis. Science 290:2114–2117.

Wolfe KH (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genetics 2(5):333–41.

Wolfe KH, Shields DC (1997). Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome. Nature 387:708–713.

Yanai I, Camacho CJ, DeLisi C (2000). Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 85(12):2641–2644.