TWO PLUS TWO DOES NOT EQUAL THREE: STATISTICAL TESTS FOR MULTIPLE GENOME COMPARISON

NARAYANAN RAGHUPATHY[1,*] ROSE HOBERMAN[2*] AND DANNIE DURAND[1,2]

[1] *Dept. of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA.*
[2] *Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.*
E-mail: {rnarayan, roseh, durand}@cmu.edu

Gene clusters that span three or more chromosomal regions are of increasing importance, yet statistical tests to validate such clusters are in their infancy. Current approaches either conduct several pairwise comparisons, or consider only the number of genes that occur in all the regions. In this paper, we provide statistical tests for clusters spanning exactly three regions based on genome models of typical comparative genomics problems, including analysis of conserved linkage within multiple species and identification of large-scale duplications. Our tests are the first to combine evidence from genes shared among all three regions and genes shared between pairs of regions. We show that our tests of clusters spanning three regions are more sensitive than existing approaches and can thus be used to identify more diverged homologous regions.

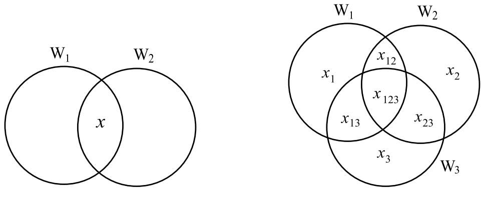
1. Introduction

An essential task in comparative genomics is to identify chromosomal regions that descended from a single ancestral region, either through speciation or duplication. Conserved homologous regions can be used to find evidence of functional selection or shared regulatory regions, and to analyze the history of large-scale duplications and rearrangements. In distantly related genomes, homologous genes are used as markers for identifying homologous regions. Gene content and order, although initially conserved, will diverge through local rearrangements, gene loss, and duplications[5]. Thus, distantly related homologous regions appear as *gene clusters*, distinct chromosomal regions that share a number of homologous gene pairs, where neither gene order nor gene content is perfectly preserved.

In order to distinguish regions that arose from the same ancestral region from unrelated regions that share homologous gene pairs, it is necessary to show that local similarities in gene content could not have occurred by chance. There is an emerging body of work on statistical tests for this purpose[2,3,4,9,10,14,15,18]. However, this work focuses almost exclusively on tests for comparisons of two regions. With the rapid rate of whole genome sequencing, analysis of gene clusters that span three or more chromosomal regions is of increasing interest.

When comparing two regions, the number of shared homologs (x, shown in Fig. 1(a)) is typically used as the measure of similarity. However, this approach cannot be directly

[*] these authors have contributed equally to this work



(a) (b)

Figure 1. Venn diagram representation of shared homologs in windows sampled from distinct chromosomal regions. (a) Pairwise comparison of windows, W_1 and W_2, which share x homologous genes. (b) Three-way comparison of W_1, W_2, and W_3, in which x_{123} homologs appear in all three windows. The variables x_{ij} represent the number of genes that appear in only W_i and W_j, and x_i represents the number of genes that appear in only a single window W_i.

extended for tests of clusters spanning more than two regions. When comparing three regions (W_1, W_2, and W_3), there are many more quantities to consider (Fig. 1(b)): the number of homologs observed in all three regions (x_{123}), the number of homologs observed in each pair of regions (x_{12}, x_{13} and x_{23}), and the number of genes observed only in a single window (x_1, x_2, and x_3). Evidence for homology comes not only from the set of homologs that appear in *all* the regions being compared (x_{123}), but also from the number of homologs that appear in only a subset of the regions ($x_{ij}'s$). How best to combine evidence from different subsets of regions remains an unsolved problem.

In this paper, we develop the first attempt to address this issue, for the problem of clusters spanning exactly three regions. Given a set of three windows sampled from three genomes, each containing r consecutive genes, we wish to determine whether the windows share more homologous genes than expected by chance. (If duplications are under consideration, two of the windows will be sampled from distinct regions of a single genome.) This problem, while restricted to three regions, exihits the basic challenges that arise in the more general problem.

Statistical tests for gene clusters in multiple regions may be useful either because the researcher is studying more than two genomic regions or because comparison with additional genomes may increase confidence that a pair of regions arose from a single ancestral region. To identify regions duplicated in a whole genome duplication (WGD), in particular, comparisons with related genomes may be necessary. Although evidence of WGD can sometimes be found by comparing a genome with itself and looking for pairwise clusters, in many cases duplicated regions may not be identifiable by direct comparison due to *reciprocal gene loss*: Following a WGD, in many cases there is no immediate selective advantage for retaining a gene in duplicate, so one copy of most duplicates is lost. As a result, the gene content of duplicated regions is often disjoint.

A solution to this problem is comparison with the genome of a closely related species that diverged shortly before the whole genome duplication (a *pre-duplication* species). If two regions in the *post-duplication* species both have significant similarity to a single region in the pre-duplication species, they are likely to be homologous even if they share few or no homologous genes. This strategy provides more statistical power to detect duplicated regions and has been successfully employed to analyze duplications in fish[6], plants[8,16,17] and several yeast species[7,11].

The most common strategy for testing significance of multiple regions is to conduct multiple pairwise comparisons (reviewed by Simillion *et. al.*[12]). If region W_1 is significantly similar to W_2, and W_2 is significantly similar to region W_3, then homology between all three regions is inferred, even if W_1 and W_3 share few genes. This approach allows the use of existing statistical methods, which are designed for comparing two regions. However, this strategy is conservative as it will only identify a three-way cluster if at least two of the three pairwise comparisons are independently significant. Furthermore, it does not explicitly recognize the additional significance of genes that occur in all three regions.

In a second approach, once a significantly similar pair of regions is identified, the genes in these regions are merged to approximate their common ancestral region[12]. Then additional pairwise comparisons are conducted with this inferred ancestral segment as the search query. This approach still allows the use of pairwise statistical tests, but is more powerful than the above approach, since the second step considers the genes that occur in W_1 as well as those that occur in W_2 when searching for a third homologous region. However, it still requires that at least one pair of regions is independently significant. Moreover, when comparing with a third region, W_3, it does not consider the additional significance of genes that appear in W_1 *and* W_2, compared to genes that appear in only one of the regions.

The previous two approaches use sequential pairwise comparisons. Another model has been proposed that allows for *simultaneous* comparison of multiple regions[3]. However this model only considers x_{123}, the number of genes that are conserved in all regions. This approach is also conservative as it does not consider genes that occur in only a subset of the regions (the $x_{ij}'s$). Thus, current approaches account for either the genes that occur in all three regions, or those that occur in pairs of regions, but not both.

In this paper, we develop the first statistical tests that consider both the quantities x_{123} *and* x_{ij} simultaneously. We obtain expressions for the probability—under the null hypothesis of random gene order—that the number of shared genes is at least as large as the number observed. These expressions are derived for genome models that are appropriate for two common types of comparative genomics problems: (1) analyses of conserved linkage of genes in three regions from three genomes, and (2) identification of segments duplicated by a whole genome duplication, via comparison with the genome of a related, pre-duplication species. We show through simulations that our tests for comparing three regions are more sensitive than existing approaches, and have the potential to detect more diverged homologous regions.

2. Statistical tests for three regions

The significance of a cluster depends not only on properties of the windows (Fig. 1), but also on the properties of the genomes (Fig. 2). The relevant properties of the genomes are the total number of genes in each genome and the *gene content overlap*— the fraction of genes shared among the three genomes. Depending on which biological questions are being investigated, the processes of gene loss differ, and an appropriate model of gene con-

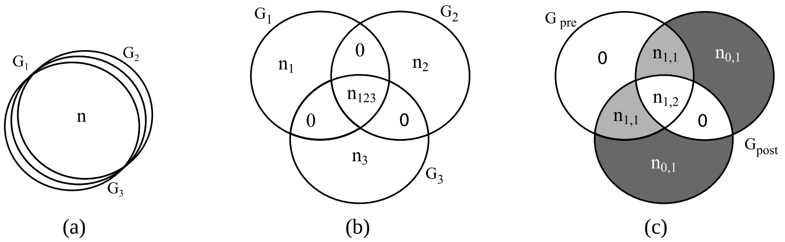


Figure 2. Gene content overlap models. The set of genes in each genome is represented as a circle. (a) Identical gene content model: all genes are shared between all three genomes. (b) Shared gene content model: n_{123} genes are shared between all three genomes. The remaining genes are singletons. (c) Pre/post duplication model: G_{pre} is the union of two ancestral, duplicated genomes embedded within it. $n_{1,2}$ genes appear twice in G_{post} (once in each embedded genome) and once in G_{pre}. These are the genes that are retained in duplicate. $n_{1,1}$ genes appear once in G_{pre} and once in G_{post}. These are the genes that were preferentially lost. $n_{0,1}$ genes appear once in G_{post} but do not appear in G_{pre}. These are the genes retained in singleton in G_{post} but lost in G_{pre}.

tent overlap will also differ. Here, we develop statistical tests for three different models of gene content overlap. The first two models are designed for comparisons of three genomes, while the third is for detection of duplicated regions by comparison with a pre-duplication genome. For each model we give analytical expressions for three statistical tests, and compute cluster probabilities for typical parameter values using Mathematica. We investigate the impact of different gene content overlap models and alternative test statistics on cluster significance, and compare the sensitivity of our tests with that of existing approaches.

2.1. *Identical gene content model*

We model a genome G_i as an ordered set of N_i genes, $G_i = 1, 2, \ldots N_i$. We ignore chromosome breaks and physical distance between genes, and assume genes do not overlap. In this, the simplest model, each genome contains n identical genes, *i.e.*, $n{=}N_1{=}N_2{=}N_3$ (Fig. 2(a)). Each gene in genome G_i has exactly one homolog each in G_j and G_k.

In order to determine the significance of gene clusters, we require test statistics that capture the essential properties of the clusters of interest. In the pairwise case, given a pair of chromosomal regions containing x observed homologs, significance is typically demonstrated by showing that $P(X \geq x)$ is small under the null hypothesis, where X is a random variable representing the number of homologs shared between the two regions. This probability can be computed using a combinatorial approach, counting the number of ways the two windows can be filled with genes, such that they share at least x genes, and normalizing by the number of ways of filling the windows without restrictions.

We illustrate this approach for the simpler case of a pairwise cluster, then present analytical expressions for the probabilities of three-region clusters under the null hypothesis. Given two windows, W_1 and W_2 of size r_1 and r_2, sampled from two genomes containing n identical genes, the number of ways the windows can share *exactly* x genes is $\binom{n}{x}\binom{n-x}{r_1-x}\binom{n-r_1}{r_2-x}$. The first binomial is the number of ways of choosing the x shared genes, and the remaining two binomials give the number of ways of choosing two sets of genes to fill the remainder of each window, such that the sets are disjoint. We normalize by the total number of ways of choosing genes to fill two windows of size r_1 and r_2 is $\binom{n}{r_1}\binom{n}{r_2}$. Thus, the probability that these windows share *exactly* x genes is[3]

$$P_2(X\!=\!x) = \frac{\binom{n}{x}\binom{n-x}{r_1-x}\binom{n-r_1}{r_2-x}}{\binom{n}{r_1}\binom{n}{r_2}} = \frac{\binom{n}{x,r_1-x,r_2-x}}{\binom{n}{r_1}\binom{n}{r_2}}, \tag{1}$$

where we define[a]

$$\binom{n}{i_1, i_2, ..., i_k} \equiv \binom{n}{i_1} \prod_{j=1}^{k-1} \binom{n - \sum_{l=1}^{j} i_l}{i_{j+1}} = \frac{n!}{i_1! i_2! \ldots (n - i_1 - i_2 \ldots - i_k)!}.$$

Thus, the probability that two windows share *at least* x genes is

$$P_2(X \geq x) = \sum_{h=x}^{r} P_2(X\!=\!h). \tag{2}$$

We use an analogous approach and notation for computing the probabilities for comparisons of three regions. In addition, we define $\vec{x} = (x_{123}, x_{12}, x_{13}, x_{23})$ and use $\vec{X} = \vec{x}$ as shorthand for $X_{123} = x_{123}$, $X_{12} = x_{12}$, $X_{13} = x_{13}$, and $X_{23} = x_{23}$. As above, we first derive an expression for the probability of observing exactly $\vec{x}$ genes, then sum over this expression to find the probability of observing at least as many shared genes.

In the above pairwise comparison, we counted the number of ways to form three different sets: the x shared genes, the $r_1 - x$ genes unique to W_1, and the $r_2 - x$ genes unique to W_2. Computing the probability of three windows containing *exactly* the observed number of shared genes is a direct extension of the two-window problem, except there are seven sets to be selected (Fig. 1(b)) instead of three sets:

$$P_3(\vec{X}\!=\!\vec{x}) = \frac{1}{\binom{n}{r_1}\binom{n}{r_2}\binom{n}{r_3}} \cdot \binom{n}{x_{123},\ x_{12},\ x_{13},\ x_{23},\ x_1,\ x_2,\ x_3}. \tag{3}$$

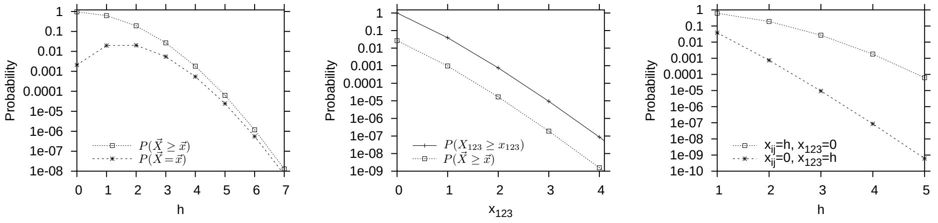
The probability of observing *at least* $\vec{x}$ shared genes is obtained by summing over all possible values of X_{123} and X_{ij},

$$P_3(\vec{X} \geq \vec{x}) = \sum_{v_{123}=x_{123}}^{u_{123}} \sum_{v_{12}=x_{12}}^{u_{12}} \sum_{v_{13}=x_{13}}^{u_{13}} \sum_{v_{23}=x_{23}}^{u_{23}} P_3(\vec{X} = \vec{v}), \tag{4}$$

where $u_{123} = \min(r_1, r_2, r_3)$, $u_{12} = \min(r_1, r_2) - v_{123}$, $u_{13} = \min(r_1 - v_{12}, r_3) - v_{123}$, $u_{23} = \min(r_2 - v_{12}, r_3 - v_{13}) - v_{123}$, and $\vec{v} = (v_{123}, v_{12}, v_{13}, v_{23})$. In the worst case, evaluating this expression takes $O(r^4)$ time. In practice, the computation time is substantially reduced, because the summand decreases exponentially as x_{123} and the $x_{ij}'s$ increase. Only the smallest values will contribute to the final probability, and most of the terms can be disregarded.

It might seem natural to use the probability of observing the *exact* number of shared homologs directly to test cluster significance. However, such an approach is risky. As shown in Fig. 3(a), for small values of x_{ij}, $P(\vec{X}\!=\!\vec{x})$ underestimates $P(\vec{X} \geq \vec{x})$ by several

[a] Note that this is a non-standard use of the multinomial notation since we do not require that $n = i_1 + i_2 + \ldots i_k$.



(a) (b) (c)

Figure 3. (a) A comparison of $P(\vec{X} \geq \vec{x})$ with $P(\vec{X} = \vec{x})$ for $n = 5000$, $r = 100$, $x_{123} = 0$, and $x_{12} = x_{13} = x_{23} = h$, as h ranges from zero to seven. (b) A comparison of $P(X_{123} \geq x_{123})$ with $P(\vec{X} \geq \vec{x})$ for $n = 5000$, $r = 100$, $x_{12} = x_{13} = x_{23} = 3$, as x_{123} ranges from zero to four.(c) A comparison of $P(\vec{X} \geq (h, 0, 0, 0))$ and $P(\vec{X} \geq (0, h, h, h))$, showing the impact of x_{123} and $x_{ij}'s$ on cluster significance, when $n = 5000$, $r = 100$.

orders of magnitude. For example, given the parameters in Fig. 3(a), when the three regions share *no* genes ($x_{123} = x_{ij} = 0$), the exact test reports a probability significantly less than one! This test will lead to false positives. However, as x_{ij} increases, the probabilities converge. This suggests that, for sufficiently large values of x_{ij}, the exact probability may be used as a fast approximation.

In order to assess the additional sensitivity gained by incorporating genes that are shared between only two of three regions into the statistical test, we compare $P(\vec{X} \geq \vec{x})$ with $P(X_{123} \geq x_{123})$, the probability of observing at least x_{123} homologs shared between all three windows. To ensure that all three windows share *exactly* x_{123} genes with no restrictions on the $x_{ij}'s$, it is necessary to select x_{12}, x_{13} and x_{23} so that they have no homologs in common. Otherwise, X_{123} would be greater than rather than equal to x_{123}. This can be achieved using the following expression for the number of windows that share *exactly* x_{123} genes:

$$q(X_{123} = x_{123}) = \sum_{x_{12}=0}^{r_1 - x_{123}} \binom{r_1}{x_{123}, x_{12}} \binom{n - r_1}{r_2 - x_{123} - x_{12}} \binom{n - x_{123} - x_{12}}{r_3 - x_{123}}, \qquad (5)$$

where the second term ensures that W_1 and W_2 share exactly x_{12} genes, and the third term ensures that exactly x_{123} genes are shared in all three windows. We then obtain the probability of observing *at least* x_{123} genes in common by summing over q as follows:

$$P(X_{123} \geq x_{123}) = \binom{n}{r_2}^{-1} \binom{n}{r_3}^{-1} \sum_{k=x_{123}}^{u_{123}} q(X_{123} = k). \qquad (6)$$

We analyzed the impact of considering the $x_{ij}'s$, by comparing Eq. 6 with Eq. 4 (Fig. 3(b)). $P(X_{123} \geq x_{123})$ is consistently two orders of magnitude greater than $P(\vec{X} \geq \vec{x})$. This is because a test based only on x_{123} fails to capture evidence of homology from genes that occur in only a subset of the windows (*i.e.*, the $x_{ij}'s$), and will severely underestimate cluster significance. For example, given a significance threshold of $\alpha = .01$ and the parameters used in Fig. 3(b), a cluster with $x_{12} = x_{13} = x_{23} = 3$ and $x_{123} = 1$ would not be considered significant using a test based on x_{123} alone, even though such a cluster is unlikely to arise by chance.

To further understand the relative importance of x_{123} and x_{ij}, we analyzed how much more a gene shared by all three windows contributes to significance than a gene shared by only two windows. Consider a cluster in which h genes are shared by all three windows (*i.e.,* $x_{123} = h, x_{ij} = 0$), compared to a cluster where there are h distinct genes shared between each *pair* of windows (*i.e.,* $x_{123} = 0, x_{ij} = h$). Notice that in both cases, each pair of windows shares h genes. However, in the first case each region only contains h shared genes, whereas in the second case each region shares $2h$ genes with the other regions. Although the total number of shared genes is larger in the second scenario, Fig. 3(c) shows that the first scenario is always much more significant. Even a small increase in x_{123} results in a large increase in significance—much more so than an increase of an equivalent number of homologous matches between pairs of regions.

2.2. *Shared gene content model*

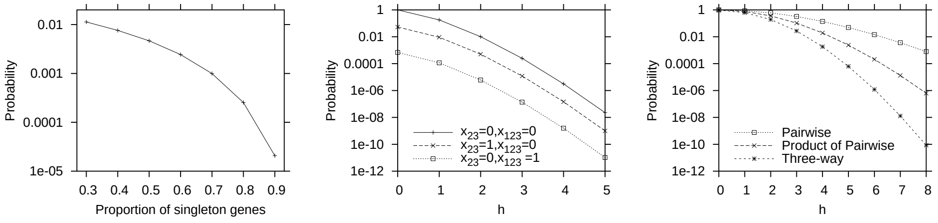
In contrast to the assumptions of the identical gene content model, in most cases, a genome will have *singleton* genes that do not have a detectable homolog in related genomes. How does this difference affect cluster significance? In the shared gene content model, we assume the genomes share a common set of $n_{123} \leq N_i$ homologs (Fig. 2(b)). In addition, each genome G_i contains $n_i = N_i - n_{123}$ singleton genes. Homology between gene pairs that have no homolog in the third genome is disregarded, with such genes being treated as singletons. This models the situation that would result if homologs were identified according to the triangle method used in COGs [13].

To compute the probability of observing exactly $\vec{x}$ shared genes, we must count the number of ways of choosing the $\vec{x}$ shared genes, as well as the genes that are unique to each window (x_1, x_2, and x_3). As in the case of identical gene content, the shared genes must be selected from the n_{123} genes common to the three genomes. However, the x_i genes that are unique to each window W_i can be selected either from the remaining common genes, or from the singletons of that genome (n_i). In the former case, care must be taken to ensure that a gene is only assigned to one window. As a result, two additional summations are required, since the number of ways to choose the x_3 genes unique to W_3 depends on how many genes from the n_{123} common genes were used to fill W_1 and W_2. The probability is:

$$P_S(\vec{X} = \vec{x}) = \binom{N_1}{r_1}^{-1} \binom{N_2}{r_2}^{-1} \binom{N_3}{r_3}^{-1} \binom{n_{123}}{x_{123},\ x_{12},\ x_{13},\ x_{23}} \sum_{i=0}^{x_1} \sum_{j=0}^{x_2} \binom{n_{123} - s}{i, j} \binom{n_1}{x_1 - i} \binom{n_2}{x_2 - j} \binom{N_3 - s - i - j}{x_3}, \tag{7}$$

where $s = x_{123} + x_{12} + x_{13} + x_{23}$ is the total number of shared genes. $P_S(\vec{X} \geq \vec{x})$, the probability of observing *at least* as many shared genes under this model, can be computed from Eq. 7 by summing over $P_S(\vec{X} = \vec{x})$, similar to Eq. 4.

We use this expression to study how cluster significance depends on the extent of gene content overlap among the genomes. As the proportion of singleton genes in the genomes increases from 0.3 to 0.9, the probability of observing a cluster drops from 0.01 to 10^{-5}



(a) (b) (c)

Figure 4. (a) The effect of singleton genes on cluster significance. The x-axis shows the proportion of singletons in each genome $(1 - n_{123}/N)$. The y-axis shows the probability $P_S(\vec{X} \geq (1, 1, 1, 1))$, when $N = N_1 = N_2 = N_3 = 5000$, and $r = 100$. (b) The effect of reciprocal loss on cluster significance in comparing pre- and post-duplication genomes, when $n_{1,2} = 450, n_{1,1} = 3600, n_{0,1} = 500, r = 50$, and $x_{12} = x_{13} = h$, as h ranges from 0 to 5. (c) Comparing pairwise probabilities, the product of two pairwise probabilities, and three-way probabilities, when $N = 5000$, $r = 100$, $x_{123} = 0$, and $x_{12} = x_{13} = x_{23} = h$.

(Fig. 4(a)). This is because as fewer homologs are shared between the genomes, it is more surprising to find them clustered together. This shows the importance of considering the extent of gene content overlap among the genomes when evaluating cluster significance.

2.3. *Pre/Post Duplication Model*

We propose a third genome overlap model specifically for analyzing duplications. Let G_{post} be a genome that has undergone a WGD and G_{pre} be a genome that diverged prior to the WGD (Fig. 2(c)). Let $n_{i,j}$ be the number of genes that appear i times in G_{pre} and j times in G_{post}, where $i \leq 1, j \leq 2$. This model only recognizes paralogs that arose through WGD, ignoring lineage specific duplications. Thus, it assumes that each gene in G_{post} has at most one paralog and that genes in G_{pre} have no paralogs; *i.e.*, $n_{2,0} = n_{2,1} = n_{2,2} = 0$. Furthermore, this model assumes that every gene that appears twice in the post-duplication genome also has a homolog in the pre-duplication genome; *i.e.*, $n_{0,2} = 0$. This assumption is based on the rationale that genes retained in duplicate are functionally important and, hence, are retained in G_{pre} as well. This assumption is supported by empirical observation. For example, in post-WGD yeast species over 95% of genes retained in duplicate are also present in each pre-WGD yeast genome[1]. Similarly, in this model every gene in G_{pre} has at least one homolog in G_{post} ($n_{1,0} = 0$). We use the convention that W_1 is the window sampled from G_{pre}, and W_2 and W_3 are sampled from G_{post}.

To compute the probability of observing *exactly* $\vec{x}$ shared homologs under the null hypothesis, we make the additional assumption that at most one copy of a duplicated gene appears in a given window. Given this condition,

$$P_D(\vec{X} = \vec{x}) = \frac{\binom{n_{1,2}}{x_{123}, x_{23}} \binom{N_{pre} - x_{123} - x_{23}}{x_{12}, x_{13}} \binom{N_{pre} - s}{x_1} \binom{N_{post} - n_{1,2} - s - x_1}{x_2, x_3}}{\binom{N_{pre}}{r_1} \sum_{i=0}^{\min(r_2, r_3)} \binom{n_{1,2}}{i} \binom{N_{pre} + n_{0,1} - i}{r_2 - i} \binom{N_{pre} + n_{0,1} - r_2}{r_3 - i}},$$

where $N_{pre} = n_{1,2} + n_{1,1}$ and $N_{post} = 2n_{1,2} + n_{1,1} + n_{0,1}$. $P_D(\vec{X} \geq \vec{x})$, the probability of observing *at least* $\vec{x}$ shared homologs under the null hypothesis, is then obtained as before by summing over $P_D(\vec{X} = \vec{x})$.

We calculated $P_D(\vec{X} \geq \vec{x})$ with parameter values based on a recent study of pre- and post-duplication yeast species[1,11]. In our simulations, $N_{post} = 5000$ and $n_{1,2} = 450$, consistent with the observation that only 16% of genes in *S. cerevisiae* are duplicate genes that arose during the WGD. Since the number of genes that occur twice in G_{post} is small, even small values of x_{123} and x_{23} will have a large impact on cluster significance. Fig 4(b) compares the significance of clusters for three reciprocal gene loss scenarios: when no genes are shared between the two regions selected from the post-duplication genome ($x_{123} = 0, x_{23} = 0$), when a single gene is shared ($x_{123} = 0, x_{23} = 1$), and when a single gene is shared among all three regions ($x_{123} = 1, x_{23} = 0$). The shape of the three curves is similar, but the probabilities drop by an order of magnitude from one to the next. Even the addition of a single gene retained in duplicate has a large impact on cluster significance! This is particularly noteworthy because current methods compare the pre-duplication region independently with each of the post-duplication regions, and thus ignore the values of x_{23} and x_{123}[6,7,8,11,16,17]. Our results show that these current methods could fail to detect clearly significant clusters, thus resulting in a substantial decrease in sensitivity.

3. Discussion

We have presented three different models of gene content overlap and proposed novel statistical tests for evaluating the significance of gene clusters spanning three regions. Our tests are the first to combine evidence from genes shared among all three regions and genes shared only between pairs of regions.

How do our three-way tests compare to current approaches reviewed in Sec. 1? Unlike tests that consider only x_{123}, our tests also consider $x_{ij}{}'s$, and thus can detect significant clusters even when x_{123} is small (Fig. 3(b)). Our tests also have advantages over current approaches based on pairwise statistical tests alone. These approaches construct multi-region clusters by merging pairwise clusters. However, this method does not explicitly consider the number of genes shared among all three regions. Our results (Fig. 3(c)) show that even a few genes conserved in all three regions dramatically increases the statistical significance of gene clusters. This effect is particularly strong when the shared gene content of the genomes is small (Fig. 4(a)). Thus, unlike pairwise tests, our approach can detect related regions where each pair of regions share only a few genes (*i.e.*, $x_{ij}{}'s$ are small), but where a few genes are also shared among all the regions (*i.e.*, x_{123} is non-zero but small).

Even when $x_{123} = 0$, we gain sensitivity over pairwise approaches. This is because the pairwise approach requires two of the three pairwise tests to be independently significant, whereas our approach considers the three regions jointly. Figure 4(c) illustrates this difference, for a scenario in which $n = 5000$ and $r = 100$. In this case, given a significance threshold of $\alpha = 0.01$, for a *pair* of regions to be significantly similar ($P_2(X \geq x)$), they must share at least seven genes. Thus, to find a three-way cluster with the pairwise approach, W_1 must share seven genes each with W_2 and W_3. In contrast, using our test $P_3(\bar{X} \geq \bar{x})$, a cluster is significant when each pair of regions shares only four genes, even when none of these genes appear in all three regions.

Since the comparison of two windows W_1 and W_2 is independent of the comparison of W_1 and W_3, one could try using the *product of two pairwise probabilities* as an approximation of the joint probability of all three windows. This approximation, though closer to the three-way probabilities, still underestimates the multi-region significance (Fig. 4(c)). This is because the product of pairwise probabilities fails to consider the genes shared between the third pair of windows (W_2 and W_3), and also does not give more weight to the genes that are shared among all the windows. Thus, we argue that pairwise tests are not always sufficient and multi-region tests will be able to identify more distantly related homologous regions. Here, we have presented initial results in this direction, yet many important problems remain. A more general test would take all paralogs into account. In addition, to investigate hypotheses of multiple WGDs within the same lineage, tests for more than three regions sampled from the same genome are required.

Acknowledgments

D.D. was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship. R.H. was supported in part by a Barbara Lazarus Women@IT Fellowship, funded in part by the Alfred P. Sloan Foundation. We thank A. McLeod, A. Goldman, J. Joseph, M. Stolzer, and N. Song for help with manuscript preparation.

References

1. K. P. Byrne and K. Wolfe. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, 15(10):1456–1461, 2005.
2. P. Calabrese, S. Chakravarty, and T. Vision. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, 19(suppl 1):i74–80, 2003.
3. D. Durand and D. Sankoff. Tests for gene clustering. *J Comput Biol*, 10(3–4):453–482, 2003.
4. R. Hoberman, D. Sankoff, and D. Durand. The statistical analysis of spatially clustered genes under the maximum gap criterion. *J Comput Biol*, 12(8):1081–1100, 2005.
5. L. Hurst, C. Pál, and M. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, 5(4):299–310, 2004.
6. O. Jaillon et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.
7. M. Kellis, B. Birren, and E. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624, 2004.
8. H. Ku, T. Vision, J. Liu, and S. Tanksley. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *PNAS*, 97(16):9121–9126, 2000.
9. A. McLysaght, K. Hokamp, and K. Wolfe. Extensive genomic duplication during early chordate evolution. *Nat Genet*, 31(2):200–204, 2002.
10. N. Raghupathy and D. Durand. Individual gene cluster statistics in noisy maps. In *RECOMB 2005 Workshop on Comparative Genomics*, volume 3678 of *LNBI*, pages 106–120. Springer-Verlag, 2005.
11. D. Scannell, K. Byrne, J. Gordon, S. Wong, and K. Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–345, 2006.
12. C. Simillion, K. Vandepoele, and Y. Van de Peer. Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, 26(11):1225–35, 2004.

13. R. Tatusov, E. Koonin, and D. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.
14. Z. Trachtulec and J. Forejt. Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm Genome*, 3(12):227–231, 2001.
15. K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res*, 12(11):1792–801, 2002.
16. K. Vandepoele, C. Simillion, and Y. Van de Peer. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet*, 18(12):604–6, 2002.
17. K. Vandepoele, C. Simillion, and Y. Van de Peer. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell*, 15(9):2192–2202, 2003.
18. J. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.