FRACTIONATION STATISTICS

Baoyong Wang

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy in Biology[1]

Mathematics and Statistics

Faculty of Science

University of Ottawa

[1]The Ph.D. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Biology

Abstract

Paralog reduction, the loss of duplicate genes after whole genome duplication (WGD) is a pervasive process. Whether this loss proceeds gene by gene or through deletion of multi-gene DNA segments is controversial, as is the question of fractionation bias, namely whether one homeologous chromosome is more vulnerable to gene deletion than the other. As a null hypothesis, we first assume deletion events, on one homeolog only, excise a geometrically distributed number of genes with unknown mean μ, and these events combine to produce deleted runs of length l, distributed approximately as a negative binomial with unknown parameter r, itself a random variable with distribution $\pi(\cdot)$. A biologically more realistic model requires deletion events on both homeologs distributed as a truncated geometric. We simulate the distribution of run lengths l in both models, as well as the underlying $\pi(r)$, as a function of μ, and show how sampling l allows us to estimate μ. We apply this to data on a total of 15 genomes descended from 6 distinct WGD events and show how to correct the bias towards shorter runs caused by genome rearrangements. Because of the difficulty in deriving $\pi(\cdot)$ analytically, we develop a deterministic recurrence to calculate each $\pi(r)$ as a function of μ and the proportion of unreduced paralog pairs. This is based on a computing formula containing nested sums. The parameter μ can be estimated based on run lengths of single-copy regions. We then reduce the computing formulae, at least in the one-sided case, to closed form. This virtually eliminates computing time due

to highly nested summations. We formulate a continuous version of the fractionation process, deleting line segments of exponentially distributed lengths in analogy to geometric distributed numbers of genes. We derive nested integrals and discover that the number of previously deleted regions to be skipped by a new deletion event is exactly geometrically distributed. We undertook a large simulation experiment to show how to discriminate between the gene-by-gene duplicate deletion model and the deletion of a geometrically distributed number of genes. This revealed the importance of the effects of genome size N, the mean of the geometric distribution, the progress towards completion of the fractionation process, and whether the data are based on runs of deleted genes or undeleted genes.

Acknowledgements

I would like to thank my advisor, Dr. David Sankoff, without whom this work would not have been possible. I am grateful to him for providing me many opportunities to explore and discover the new areas of genome rearrangement. I very much appreciate his guidance and encouragement over the years. My research was funded by an NSERC Discovery Grant awarded to Dr. Sankoff.

Thanks to all the members of the Sankoff lab, I have benefited much from their collaboration and helpful discussions.

Thanks to the members of my thesis jury for numerous helpful corrections and revisions that are incorporated into the final version; Professor Liqing Zhang from Viginia Tech, Professor Yiqiang Zhao from Carleton University, Professors Robert Smith? and Pierre-Jérôme Bergeron from my home department of Mathematics and Statistics, University of Ottawa.

Finally, I am grateful to my wife for her trust, support and patience throughout this work and in my daily life.

Dedication

To my daughters, Yiming and Emily, for giving me dreams and courage.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

From a statistical viewpoint, this thesis is a study of run lengths of zeros and ones in a certain type of binary sequence. The random process generating these sequences is inspired by the "fractionation" phenomenon in the evolution of genomes [1]. The higher animals, almost all fishes, some amphibians, yeasts, protists and, especially, the higher plants, trace their history to one or more of the rare events millions, tens of millions or even hundreds of millions years ago, known as "Whole Genome Doubling" (WGD). Such an event involves the creation of individuals, populations and eventually, new species having twice the number of chromosomes and twice the number of genes as its ancestral species. WGD triggered a massive wave of shedding of most of the tens of thousand of duplicate genes it created. For any duplicate pair, one member or the other may be lost, but not both. Along the length of any particular chromosome, then, we can abstract a sequence of zeros and ones, where the zeros represent genes that have been lost and the ones represent genes that are still present. We cannot observe the zeros directly on a chromosome, but we can always infer them in their correct position in the sequence through the presence of their duplicate copy on the "homeologous" (duplicate) chromosome. It is these binary

sequences, or pairs of homeologous sequences, that are the subject of this thesis. The biological motivation is how to formally characterize the fractionation process. Are pairs of duplicates chosen at random in the genome to lose one of their members? Or is there a process of loss affecting a number of contiguous genes simultaneously? Does fractionation affect two homeologous chromosomes in a balanced way, or is there a bias towards losing more genes from one of the chromosomes than from the other? How can we study fractionation in the presence of the many other evolutionary processes that affect the presence and position of genes on chromosomes, such as genome rearrangements, or frequent duplication of individual genes?

The biological implications of these questions are important because they may shed light on a wide range of genetic and evolutionary processes, such as the creation of novel gene functions (when a pair of genes resists fractionation) and new gene networks, radiation of species, the operation of "gene silencing", pseudogenization, the maintenance of stoichiometry among interacting genes, and many others.

The work I will report on represents one aspect of a team effort in studying WGD and its consequences dating back fifteen years. It includes methods to reconstruct the ancestor of a rearranged doubled genome [2] or tripled (or higher multiplicity) genome [3], incorporating doubled genomes into phylogenies [4], comparison of fractionation rates and biases across the tree of life [5], reconstructing fractionated genomes [6, 7], and others.

The first attempt at formal work on this problem was by Byrne et al [9] . They used simulations to show that deletion patterns in S. cerevisiae are closer to a gene-by-gene model than a Poisson model of mean either of 1 or 2. They gave no details of the simulations, such as the effect of a deletion of size zero, presumably no deletion, so that the means of the deleted segments are in reality much greater than 1 or 2. The next work was done by van Hoek and Hogeweg [11] who fitted geometric

distribution for deletion lengths and found, contrary to the previous work, that a geometric distribution with mean 1.1 was a better fit to the yeast data. These two analyses used very different biological assumptions in detecting deletion length data in the yeast data. Neither of them used analytical models to predict deletion length distributions or a test statistic.

I will present my research in the form of three self-contained papers. The first (reproduced here as Chapter 2) was presented to the RECOMB Satellite Workshop on Comparative Genomics in Galway, Ireland, 8–10 October, 2011. It was published in BMC Bioinformatics 12: S9, S5 (2011) [13]. After an exploration of the empirical evidence on fractionation, my paper contains the first formalization of the one-sided and two-sided models of overlapping geometric deletion events, and an explanation of why overlapping events differ from sums of independent geometric variables. The definitions of the "skipping" of overlapping events and, in the two-sided case, of "blocking" of part of an event on one chromosome by a previous event on the homeologous chromosomes, are original to this paper and are the basis of the rest of my research. The main result is the derivation of recurrence for the evolution of the fractionated sequence in the one-sided case. This recurrence has two parts. The first contains computing formula for the probability that a deletion event creates a new run of single-copy genes, touches or overlaps one previous single-copy run, or two single-copy runs; these suffice to allow comparison of the recurrence to simulated fractionation patterns. The second part of the recurrence is an update of the parameters necessitated by the new event. The computing formulae are exact, but contain large numbers of nested summations, which require excessive execution times and limit the possibility of extensions to larger overlaps.

The second paper (reproduced here as Chapter 3) is an extension of the first to the two-sided case, which requires that blocking (truncating the geometric variable)

be incorporated into the skipping-only model for overlap pair probabilities. This is considerably more difficult, but the recurrence, including the computing formulae, is successfully derived and compared to simulations. The two-sided model allows a consideration of biased fractionation, and this is also incorporated into the probability model. This work was presented to the Asia-Pacific Bioinformatics Conference in Melbourne, Australia, 17–19 January 2012, and was published in BMC Genomics 13:S1, S8 (2012) [14].

The third paper, Chapter 4, submitted for presentation at a forthcoming conference [15], contains three major advances on the previous work. The first is the discovery of how to reduce the computing formulae, at least in the one-sided case, to closed form. This virtually eliminates computing time, due to highly nested summations, as an impediment to further research. The second achievement is the formulation of a continuous version of the discrete fractionation process, again in the one-sided version. Instead of deleting one or more contiguous genes, the number being geometrically distributed, from a chromosome containing N terms, we delete line segments of exponentially distributed lengths from a long linear segment representing the chromosome, allowing skipping over previous deleted regions. Instead of nested sums for the overlap probabilities, we derive nested integrals, which are considerably easier to solve. This yields the surprising discovery that the number of previously deleted regions to be skipped by a new deletion event is exactly geometrically distributed. Indeed, it was this success with the continuous model that inspired the re-examination of the computing formulae for the discrete case, and thus led to the closed-form solutions.

Because the update part of the recurrence has not yet been put into exact and easily computable form, in the third paper I undertook a large simulation experiment to determine the kind of data that would be necessary to discriminate between the

gene-by-gene duplicate deletion model and the model of simultaneous deletion of a geometrically distributed number (mean greater than 1) of contiguous genes. Thus I was able to quantitatively determine, for the one-sided case, the importance of effects of genome size N, the mean of the geometric distribution, the progress towards completion of the fractionation process, and whether the data are based on runs of deleted genes or undeleted genes.

Finally, in Chapter 5, I list some motivations, results and directions for further research on these problems.

Chapter 2

Fractionation statistics

Baoyong Wang, Chunfang Zheng and David Sankoff. 2011. Fractionation statistics. *BMC Bioinformatics* 2011, 12(S9): S5.

Dr. David Sankoff helped plan the research and write the paper. I was responsible for designing the algorithm and getting the results in this work. Chunfang Zheng did simulations for our results.

2.1 Background

Whole genome doubling (WGD) triggers the wholesale shedding of duplicate genes through processes such as epigenetic silencing, pseudogenization, and deletion of chromosomal segments containing one or more genes [1, 5, 8, 10, 11]. The extent to which this *paralog reduction* is a gene-by-gene *inactivation* process [9] targeting redundant copies at random points throughout the genome, or a consequence of largely random *excision*, elimination of excess DNA [9], is controversial and likely varies from one phylogenetic domain to another. The distinction between these two processes is not sharp: the inactivation effect may be produced not only by pseudogenization and various suppression and silencing mechanisms but also by the actual excision of a small but critical region of a gene or promoter. Conversely, the apparent excision of two or more adjacent genes may rather be due to any of a variety of genetic, epigenetic or functional interactions, rather than the deletion of a DNA fragment. Nevertheless, the determination of whether paralog reduction is a gene-by-gene process or the deletion of longer stretches DNA is key to understanding the dynamics of genome evolution, not only following WGD, but as part of the continual innovative expansion and simplifying shrinkage of genomes over time.

The other face of paralog reduction is the process of fractionation. When a duplicate gene is lost, it may be lost from one copy (*homeolog*) of a chromosome or the other. When compared to the pre-WGD genome, or to a closely related but unduplicated genome, this creates an interleaving pattern, such that it is only by *consolidating* [9] the two homeologous single-copy regions that the full original gene complement becomes apparent. That the consolidated region is directly comparable to homologous regions in related genomes is due to the fact that single-copy genes are rarely deleted — of the two duplicates created by WGD, it is unlikely that both

are deleted, for obvious functional reasons.

Fractionation is an important evolutionary process whenever WGD occurs and is of particular interest for comparative genomics, since it results in a genome that is highly scrambled with respect to its pre-WGD ancestor. The study of fractionation also brings up the question of *bias*: are paralogs always or generally lost from the same "side", or are they lost randomly from one homeologous chromosome or the other [1, 5, 10, 12]?

In this paper, we analyze paralog reduction and fractionation in terms of two models, one easier to analyze but the other more realistic. First we model paralog reduction on only one of the two homeologous chromosomes as a series of excisions of geometrically distributed lengths and show how to use the observed run lengths of single-copy genes on the other chromosome to estimate the parameter of the deletion-length distribution.

In the second model, we allow excisions on both homeologous chromosomes and model deletion lengths in terms of truncated geometric distributions to account for the above-mentioned prohibition against deleting single-copy genes.

This work is essentially the creation of a simple, one-parameter "null" model of paralog reduction, where deletion is by random events involving geometrically distributed (with mean μ) numbers of genes on one homeologous chromosome or randomly on both of them. This sets up the possibility of statistical tests of real WGD descendants, to see if the geometric hypothesis is acceptable and to see if fractionation is unbiased or not. We will not explicitly investigate the alternative hypotheses of gene-by-gene excision or biased fractionation; our task here, aside from estimating the parameters of our model, is simply to set up the null statistical model with a view to eventually developing useful statistical tests of hypothesis for this problem.

In a previous study of post-WGD evolution [5], we took chromosomal rearrangement events into account. In the present paper, we do not incorporate rearrangement into our model, but we do reanalyze some of the data, to explore the effects of genome rearrangement processes in confounding the evidence of fractionation and to suggest a way of redressing the loss of information.

The lengths of runs of undeleted genes may be considered independent samples from a geometric distribution, and the lengths of runs of deleted genes are also independent, but we show that the deletion events making up a run of deleted genes are not independent. As a consequence, the distribution of deleted run-lengths seems beyond the scope of straightforward mathematical derivation. The major analytical and computational result of this paper is the construction, implementation and evaluation of a deterministic recurrence to calculate the distribution of the number of deletion events per run as a function of μ and the proportion θ of unreduced paralog pairs.

The models

The structure of the data

The data on paralog reduction are of the form (G, H), where G and H are binary sequences indexed by $\mathbb{Z}$, satisfying the condition that $g(i) + h(i) > 0$. This condition models the prohibition against deleting both copies of a duplicated gene. We may also assume that whatever process generated the 0's and 1's is homogeneous on $\mathbb{Z}$.

The sequence $G + H$ consists of alternating runs of 1's and 2's. We denote by $p(l), l \geq 1$ the probability distribution of length of runs of 1's. For any finite interval of $\mathbb{Z}$ we denote by $f(l), l \geq 1$ the empirical frequency distribution of length of runs of

1's.

The use of $\mathbb{Z}$ instead of a finite interval is consistent with our goal of getting to the mathematical essence of the process, without any complicating parameters such as interval length. In practice, we will use long intervals of 100,000 or 300,000 so that any edge effects will be negligible. See [10] and the section below on 15 WGD-descendant genomes for *ad hoc* ways of handling biological scale intervals.

2.1.1 One-sided deletion

In this case, $h(i) = 1$, for $-\infty < i < \infty$. We assume a continuous time process, parameter $\lambda(t) > 0$, only to ensure no two events occur at the same time. We start $(t = 0)$ with $g(i) = 1$ for all i. At any $t > 0$, consider any i where $g(i) = 1$. With probability $\lambda(t)dt$, the following *deletion event* occurs, *anchored* at position i: we choose a positive number a according to a geometric variable $\mathbf{y}$ with parameter $1/\mu$; i.e.,

$$\begin{aligned} P[\mathbf{y} = a] &= \gamma(a) \\ &= \frac{1}{\mu}\left(1 - \frac{1}{\mu}\right)^{a-1}, \quad a \geq 1, \end{aligned} \tag{2.1.1}$$

and we set $g(i) = 0, g(i+1) = 0, \ldots, g(i+a-1) = 0$, unless one or more of these is already 0, in which case we skip over it and continue to convert the next available 1's into 0's, until a total of a 1's have been converted. This is a natural way to model the excision process, since deletion of duplicates and the subsequent rejoining of the DNA directly before and directly after the excised fragment means that this fragment is no longer "visible" to the deletion process. Observationally, however, we know deletion has occurred because we have access to the sequence H, which retains copies of the deleted terms.

When the deletion event has to skip over previous 0's, this hides the anchor i and length a of previous deletion events. Denote by $\mathbf{r}$ the random variable indicating the total number of deletion events responsible for a run. Then, for a fixed $\mathbf{r} = r$, the run length $\mathbf{z}$ is distributed as the sum of r geometric variables, which would result in the negative binomial distribution

$$\begin{aligned} P[\mathbf{z} = l] &= p_{\text{negbin}}(l) \\ &= \binom{l-1}{l-r} \left(\frac{1}{\mu}\right)^r \left(1 - \frac{1}{\mu}\right)^{l-r}, \quad l \geq r, \end{aligned} \tag{2.1.2}$$

if these geometric variables were independent. As we shall see later, however, the hypothesis of independence does not hold. If we observe G at some point in time, as in the last row of Table 2.1, all we can observe are the run lengths of 0's and 1's. We cannot observe the a, i or r, while t and $\lambda(t)$ are unknown and, as we shall see, only mathematical conveniences that do not enter into our calculations. The parameter about which we wish to make statistical inferences is the deletion length distribution parameter μ, since it is this quantity that is at the heart of the biological controversy about paralog reduction. This inference therefore can only be based on the run lengths and the proportion of remaining 1's. If the probability distribution of $\mathbf{r}$ is $\pi(\cdot)$, the distribution of run length $\mathbf{x}$ is approximately

$$\begin{aligned} P[\mathbf{x} = l] &= p(l) \\ &= \sum_{r \geq 1} \pi(r) \binom{l-1}{l-r} \left(\frac{1}{\mu}\right)^r \left(1 - \frac{1}{\mu}\right)^{l-r}. \end{aligned} \tag{2.1.3}$$

The one-sided model is an extreme version of biased fractionation and is not meant to model any real situation. It is, however, relatively tractable and hence provides a mathematical framework for understanding more realistic cases.

event	i	a	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	r
start			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
1	-1	3	1	1	1	1	1	1	$\underline{0}$	$\underline{0}$	$\underline{0}$	1	1	1	1	1	1	1	1
2	-4	1	1	1	1	$\underline{0}$	1	1	0	0	0	1	1	1	1	1	1	1	1,1
3	4	4	1	1	1	0	1	1	0	0	0	1	1	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	1	1,1,1
4	-5	4	1	1	$\underline{0}$	0	$\underline{0}$	$\underline{0}$	0	0	0	$\underline{0}$	1	0	0	0	0	1	3,1

Table 2.1: One-sided model: Four deletion events leading to two runs of 0's. This illustrates the creation of a long run with $r = 3$ subsuming two previous shorter runs. Note that r is not observable.

2.1.2 Two-sided deletion

In a more realistic model, deletions can occur both in sequence G and sequence H as in Table 2.2. Thus, before choosing a position i, we chose either G or H with probability ϕ and $1-\phi$, respectively. The default we shall study here, $\phi = \frac{1}{2}$, represents unbiased fractionation. Then we choose position i, where $g(i) + h(i) = 2$, and the geometric variable a as before. Suppose G is the chosen sequence. Then $g(i)$ is set to 0, $g(i+1)$ is set to 0, and so on until $g(a+i-1)$, unless we first reach a position j where $g(j)$ is already 0, in which case we skip as before, or until we reach a position k where $h(k) = 0$. In this case, we cannot continue to delete, because $g(k)$ is a single-copy gene, and we are prohibited from letting $g(k) + h(k) = 0$, for any k. In this case, we must truncate the geometric variable a, having already deleted only $k - i < a$ terms.

In this model, the deletion length is no longer geometric but a mixture of geometric and truncated geometric variables, and run length is no longer negative binomially distributed.

event	i	a	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	r
start			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
1	-1	3	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1
			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2			1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1,1
	-4	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	
3	4	4	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1,1,1
			1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	
4			1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	1	3,1
	-5	4	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	

Table 2.2: Two-sided model: Four deletion events affecting two homeologous chromosomes, leading to two runs of single-copy genes. The fourth step illustrates how further deletion (at $i = -1$) and the "skip" process (at $i = 2$) are blocked when a single-copy gene is encountered ($i = -1$) on the homeologous chromosome, truncating the geometric variable a.

2.2 Results

2.2.1 Simulations to determine π

We carried out a simulation of the one-sided model on an interval of $\mathbb{Z}$ of length 100,000. The top row of Fig. 2.1 compares $\pi(r)$ when $\theta = 0.5$ and $\theta = 0.1$, for $\mu = 2, 3, 6$ and 11. We can see that the number of deletion events contributing to a run is somewhat dependent on μ when half of the sequence has been deleted, but is strongly dependent when 90% has been deleted. In the bottom row, the graph on the left shows that run-length l is distributed very differently for $\mu = 2$ and $\mu = 11$, when the proportion of the sequence deleted is exactly the same. This strongly suggests that observing the run-length distribution and the overall proportion of deletions should allow us to infer μ.

Finally, the remaining graph in Fig. 2.1 shows that any edge effects in our simulation are negligible. Whether we work with G and H on an interval of length 100,000 or, as in another simulation, length 300,000, gives virtually the same results.

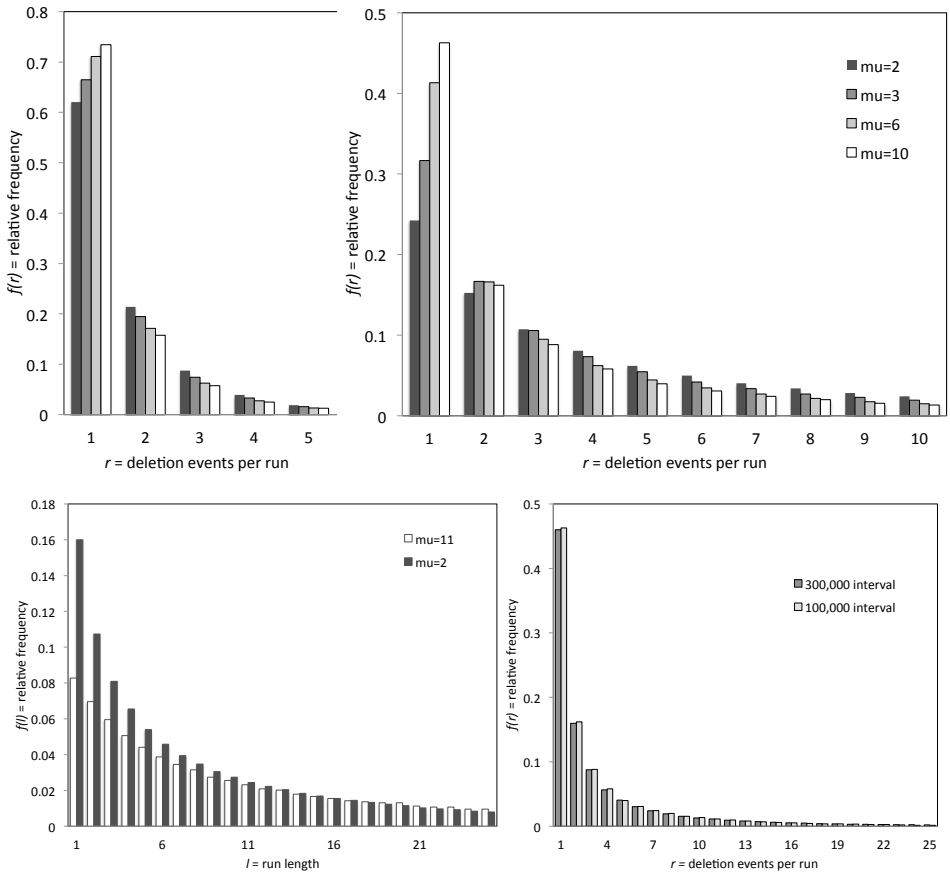

Figure 2.1: Distribution of number of deletion events r composing each run when the proportion of sequence deleted is 0.5 (top left) and 0.9 (top right). Distribution of run length, reflecting a mixture of negative binomial distributions, for two values of the parameter of the underlying geometric distribution (bottom left). Almost identical results for simulation interval of 100,000 or 300,000 genes (bottom right).

Fig. 2.2 shows the relationship, in the one-sided and two-sided models, between the proportion of genes deleted, on one chromosome or the other, and the average run length, for a range of values of μ. This confirms our impression that average run-length and overall proportion of deletion, both observable, can be used to infer

μ.

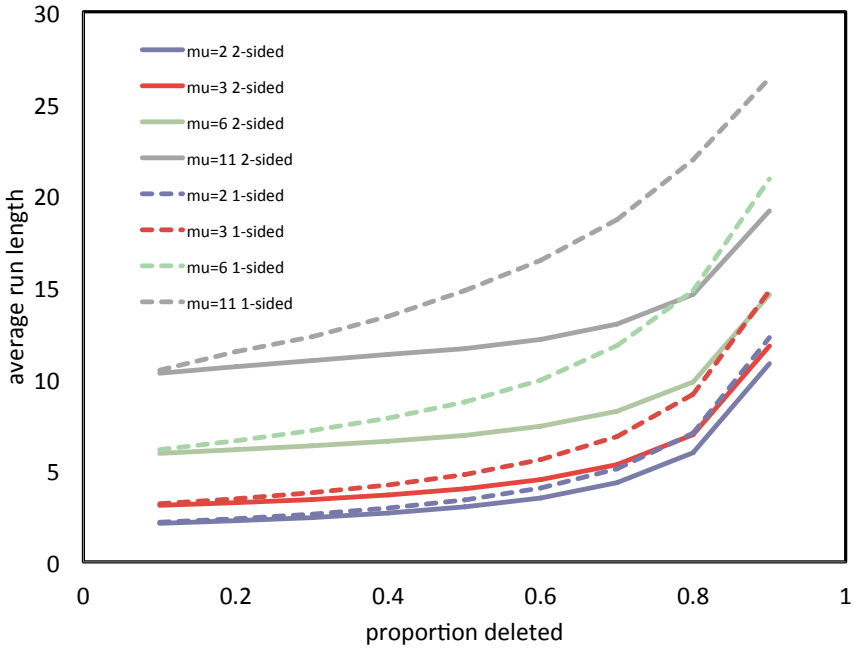

Figure 2.2: Average length of run of single-copy genes in one-sided and two-sided models for $\mu = 2, 3, 6$ and 11.

2.2.2 Non-independence of deletion events in a run

A long deletion event within a run of undeleted genes has a greater chance of including all the following genes in that run, and possibly successive runs as well, than a short event deleting, say, only one or two genes. This implies that longer deletion events will tend to be grouped together in an event while short events are more likely to be in short runs. Thus, the events making up a run are not chosen independently. This is reflected in the simulations in Fig. 2.3 for the case $\theta = 0.3$.

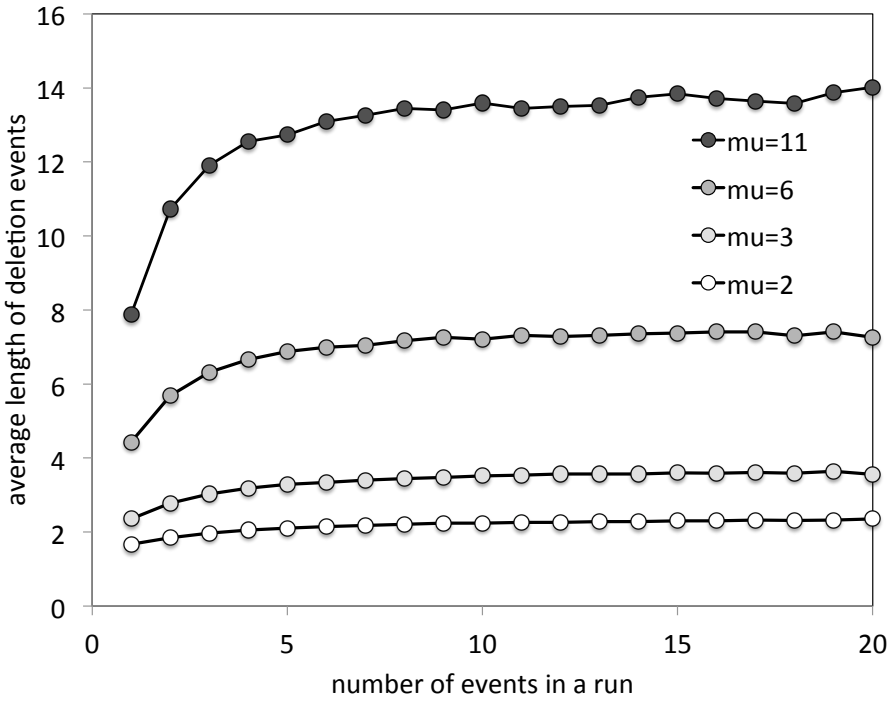

Figure 2.3: Non-independence of deletion events in a run. Association of shorter deletions with smaller values of r.

2.3 Application to 15 descendants of WGD events

To explore the relevance of our models for real genomes, we emphasize that we observe only the proportion θ of unreduced duplicates and the distribution of run lengths of single-copy genes on both homeologous chromosomes. (We can also observe the distribution of the run size of surviving paralog pairs, although models have not been developed for this.) We cannot observe t or λ. We cannot sample from the geometric distribution of deletion sizes, only their accumulation into runs, so that we cannot directly estimate its mean μ, the parameter of biological interest.

In [5], we studied 15 descendants of 6 ancient WGD events. In real genome sequences such as these, many or most runs of deleted paralogs will be impossible to identify; one or both of the homeologous regions will have been disrupted by

inversions, translocations or other rearrangement events that juxtapose the surviving genes in the run with genes originally remote on the chromosome or from elsewhere in the genome.

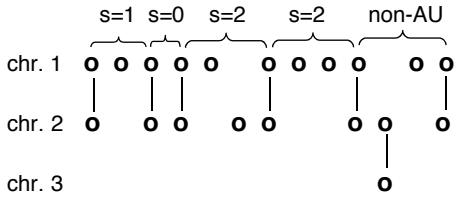

Figure 2.4: The number of single copy genes bounded by pairs of duplicates on chromosomes 1 and 2 is the sum of those on chromosome 1 and chromosome 2. The last two pairs of duplicates on chromosomes 1 and 2 do not border an AU because one of the genes between them has a paralog on chromosome 3. From [3].

We could, however, identify some two-sided undisrupted runs of single-copy genes, fractionated between two chromosomal regions. We searched for such *analytical units* (AU), two-sided runs flanked at either end by a pair of undeleted duplicate genes, with the two flanking genes on a chromosome having the same orientation, and including no intervening gene having a paralog somewhere outside the run, as in Fig. 2.4. It is statistically unlikely that such an AU configuration be produced by a series of compensating rearrangements, so that any rearrangements must have occurred entirely within the run, or have included the entire run intact plus the flanking duplicate pairs.

Among all the runs of single-copy genes in a WGD-descendant genome, it is only the AU that can be used as evidence for the paralog reduction process, because it is only from these that we can reconstruct common conserved homeologous regions on two chromosomes (or remote regions on one chromosome).

Key characteristics of the genomes, their global properties and the properties of the two-sided runs are given in Table 2.3. $D = d/(n - m)$ is the number of rearrangements per gene since WGD, where d is calculated only on the duplicates by the algorithm in [5], n is the total number of genes in the given genome and m is the number of single-copy genes. The way d is calculated, there are between one and two breakpoints per rearrangement. We do not know how many rearrangements have affected the whole genome, duplicates and single-copy, but, as a first approximation, we assume that the probability that any adjacency will be disrupted by a rearrangement since the WGD is proportional to D or αD. The proportionality constant $\alpha \leq 1$ is unknown, but experience suggests $\alpha = \frac{1}{2}$ is a reasonable value.

	t	n	m	d	$1-\theta$	$\bar{u}$
S. cerevisiae	150	5616	4498	135	0.89	6.0958
C. glabrata	150	5180	4382	252	0.92	5.3839
V. polyspora	150	5112	4164	202	0.9	4.922
S. bayanus	150	5857	4773	186	0.9	5.8297
N. castelli	150	5213	4053	221	0.88	5.0717
Paramecium	20	38626	14576	214	0.55	2.0299
populus	70	20082	7228	2600	0.53	1.6402
Arabidopsis	50	25655	13267	2701	0.68	3.6086
fugu	350	14251	12653	374	0.941	3.806
medaka	350	14564	13352	362	0.957	5.0629
stickleback	350	16726	14876	519	0.941	4.3792
tetraodon	350	17120	16088	310	0.969	6.876
chicken	450	10077	8495	686	0.915	3.6122
opossum	450	13339	11589	751	0.93	5.5507
human	450	13828	12144	673	0.935	3.818

Table 2.3: Evolutionary inference about 15 descendants of 6 WGD events. t: time in millions of years since event, n: total genes, m: single-copy genes, $1 - \theta = \frac{2m}{n+m}$, d: halving distance [15], $\bar{u}$ average run length.

In [5], the AU lengths in each of the 15 genomes were distributed as in Fig. 2.5. Since we model run length in terms of an unknown mixture of distributions involving

r geometric or truncated geometric variables, where $\pi(r)$ is unknown, we cannot infer μ directly. Nevertheless, we remark in the figure that the frequency distribution $f(u)$ of the run lengths $u \geq 1$ is closely approximated by a geometric distribution with mean $\bar{u}$ in all of the cases, except where there are few data. The mean $\bar{u}$ varies widely from genome to genome. In this section, we will continue to make use of this approximation to help understand the data.

Consider an AU of length u. There are $u+1$ possible breakpoints in an AU of length u, including the two at either end of the run of single-copy genes involving the flanking duplicate genes, that could destroy the AU, according to definition.

Each adjacency in an AU will survive (not be disrupted by a breakage) an evolutionary period equal to the time from WGD with probability approximately $(1-\alpha D)$. An AU of size u will survive with probability $(1-\alpha D)^{(u+1)}$. Then $f(u)/(1-\alpha D)^{(u+1)}$ is an estimate of the frequency of AUs of size u if there had been no rearrangements. The predicted relative frequency of run length becomes a geometric distribution with mean ν, where

$$1 - \frac{1}{\nu} = \left(1 - \frac{1}{\bar{u}}\right) / \left(1 - \frac{1}{z}\right), \tag{2.3.1}$$

where $\frac{1}{z} = \alpha D$, and

$$\nu = \frac{\bar{u}z - \bar{u}}{z - \bar{u}}. \tag{2.3.2}$$

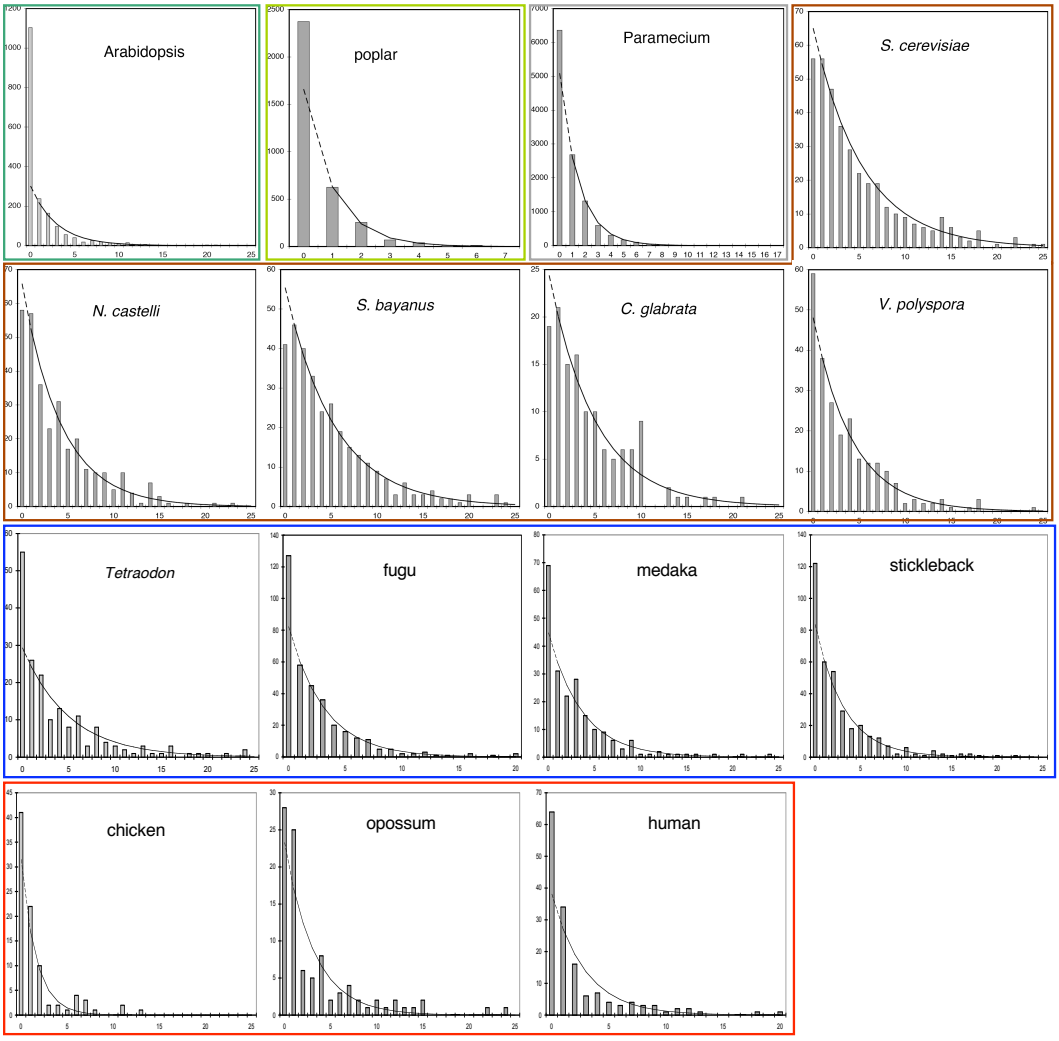

Figure 2.5: Distribution of length of run of single copy genes in 15 genomes descended from WGD events. Zero length indicates adjacent pairs of paralogs (i.e., not single-copy). Coloured boxes contain genomes descended from the same event. Frequencies of zero-length runs are not considered in the fitting by the geometric distributions shown. From [3].

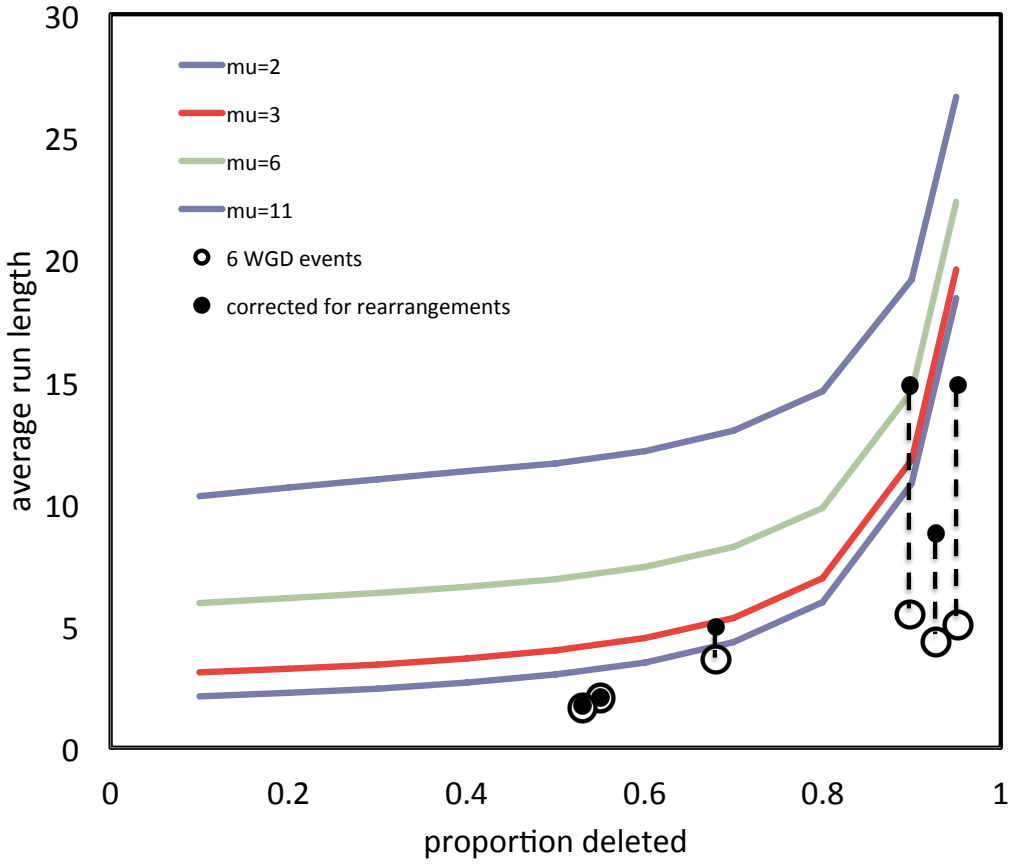

Figure 2.6: Mean deletion run-length in WGD descendants, uncorrected and corrected for rearrangements, compared to average length of run of single copy genes in the two-sided model for $\mu = 2, 3, 6$ and 11.

Fig. 2.6 shows the two-sided curve of run length versus proportion deleted as in Fig. 2.2, but with the mean run-length, $\bar{u}$, averaged over the descendants of each of the six distinct WGD events superimposed. Each point is connected in the figure to the corrected mean ν calculated from Eq. (2.3.1). We used $\alpha = 0.5$. This somewhat arbitrary choice is bounded above by the fact that z must be greater than $\bar{u}$ in Eq. (2.3.1).

This correction procedure is relatively unstable, since it is very sensitive to the arbitrary parameter α. All the more so with very low values of θ, as on the right of

Fig. 2.6, where the model begins to percolate; i.e., where the runs merge together at a rapidly increasing rate. Nevertheless, we see no evidence in the figure that μ is much greater than 1, leaving a gene-by-gene model very much a viable candidate alongside the geometric excision model.

2.4 A model for $\pi(r)$ in the one-sided model

We are interested in inferring μ from the observed distribution of run lengths and the proportion of undeleted terms θ. At the outset, $\theta = 1$. As $t \to \infty$, $\theta \to 0$. We are not, however, interested in t, since it is not observable and any time-based inference we can make about μ will depend only on run lengths and θ in any case. On the other hand, r, the number of deletion events per run, is an interesting variable since we can assume run length is $r\mu$ on average. And we can model the evolution of r directly in the one-sided model. We consider the distribution π as a function of θ.

As π changes, probability weight is redistributed among several types of run:

1. new runs ($r = 1$) falling completely within an existing run of undeleted terms, not touching the preceding or following run of deleted terms

2. runs that touch, overlap or entirely engulf exactly one previous run of deleted terms with $r \geq 1$, thus lengthening that run to $r + 1$ events,

3. runs that touch, overlap or engulf, by the skipping process, two previous runs of r_1 and r_2 events respectively, creating a new run of $r_1 + r_2 + 1$ events, and diminishing the total number of runs by 1, and

4. runs that touch, overlap or engulf, by the skipping process, $k > 2$ previous runs of of $r_1, \ldots, r_k$ events respectively, creating a new run of $r_1 + \cdots + r_k + 1$

events, and diminishing the total number of runs by $k-1$. Case 3 above may be considered a special case of this for $k=2$ and Case 2 for $k=1$.

The first process, involving a deletion event of length a, requires a run of undeleted terms of at least $a+2$. What can we say about runs of undeleted terms? We know that runs of deleted terms alternate with runs of undeleted terms, so that there is one run of the former for each of the latter. The mean length $\bar{u}$ of the deleted runs should be $(1-\theta)/\theta$ times the mean length $\bar{v}$ of the undeleted runs:

$$\bar{v} = \frac{\theta}{1-\theta}\mu \sum_{r=1}^{\infty} r\pi(r). \tag{2.4.1}$$

The distribution $\rho(l)$ of lengths of the undeleted runs is geometric, since each deletion event creates a randomly placed demarcation between two runs in the sequence consisting of all the remaining terms. The number of terms between two successive demarcations corresponds to the difference between successive order statistics. And is hence geometrically distributed.

The proportion of terms in runs of length l is $l\rho(l)/E_\rho$, where $E_\rho = \sum_{l>0} l\rho(l)$. As depicted in Fig. 2.7, the probability p_A that a deletion event falls within a run of length l without deleting the terms at either end is

$$\begin{aligned} p_A &= \sum_{l>2} \frac{l\rho(l)}{E_\rho} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{E_\rho} \sum_{l>2} \rho(l) \sum_{j=2}^{l-1} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{E_\rho} \sum_{l>2} \rho(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a) \end{aligned} \tag{2.4.2}$$

where j indexes the starting position of the deletion within the run, and a is the number of terms deleted in the event. The probability p_B that a deletion event

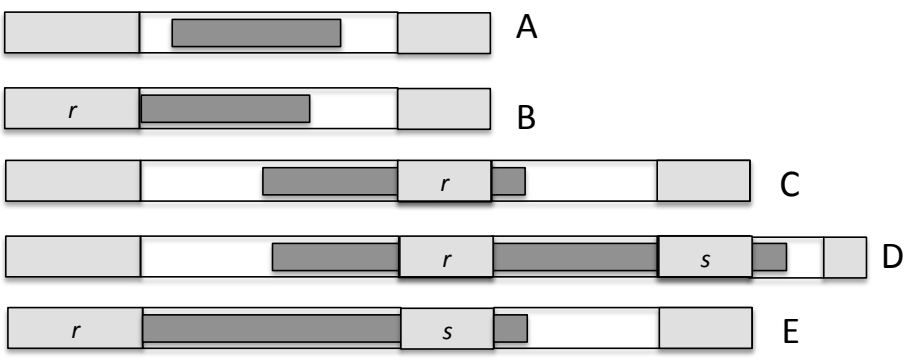

Figure 2.7: Types of deletion event affecting less than three pre-existing runs. White area indicates run of undeleted terms. Lightly shaded area indicates run of previously deleted terms. Darker area represents current deletion event. A: creates one new run with r = 1. B: lengthens left hand run to r + 1 events. C: lengthens right hand run to r + 1 events. D and E: merge two runs to create a single run with r + s + 1 deletion events.

touches only the run of deletions on the left of the run of undeleted terms is

$$p_B = \frac{1}{E_\rho} \sum_{l>1} \rho(l) \sum_{a=1}^{l-1} \gamma(a). \tag{2.4.3}$$

The probability p_C that a deletion event touches or overlaps the run of deletions on the right but does not extend over the entire run of undeleted terms beyond that is:

$$\begin{aligned} p_C &= \frac{1}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{j=2}^{l} \sum_{a=l-j+1}^{l-j+k} \gamma(a) \\ &= \frac{1}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \rho(l)\rho(k) \\ &\times \left(\sum_{a=1}^{\min[l-2,k-1]} a\gamma(a) + \sum_{a=\min[l-1,k]}^{\max[l-1,k]} \min[l-1,k]\gamma(a) + \sum_{a=\max[l,k+1]}^{l+k-2} (l+k-a-1)\gamma(a) \right). \end{aligned} \tag{2.4.4}$$

The probability p_D that a deletion event completely overlaps the run of deletions on

the right and touches or overlaps the run of deletions beyond that but does not extend over a further run of undeleted terms is:

$$\begin{aligned} p_D &= \frac{1}{E_\rho}\sum_{l>1}\sum_{k\geq 1}\sum_{h\geq 1}\rho(l)\rho(k)\rho(h)\sum_{j=2}^{l}\sum_{a=l-j+k+1}^{l-j+k+h}\gamma(a) \\ &= \frac{1}{E_\rho}\sum_{l>1}\sum_{k\geq 1}\sum_{h\geq 1}\rho(l)\rho(k)\rho(h) \\ &\times\left(\sum_{a=k+1}^{\substack{\min[l+k-2,\\ h+k-1]}}(a-k)\gamma(a)+\sum_{a=\min[l+k-1,\ k+h]}^{\substack{\max[l+k-1,\\ k+h]}}\min\left[l-1,h\right]\gamma(a)+\sum_{a=\max[l+k,\ k+h+1]}^{l+k+h-2}(l+k+h-a-1)\gamma(a)\right). \end{aligned} \tag{2.4.5}$$

The probability p_E that a deletion event touches the run of deletions on the left of the run of undeleted terms and touches or overlaps the run of deletions on the right but does not extend over the entire run of undeleted terms beyond that is:

$$p_E = \frac{1}{E_\rho}\sum_{l\geq 1}\sum_{k\geq 1}\rho(l)\rho(k)\sum_{a=l}^{l+k-1}\gamma(a). \tag{2.4.6}$$

The event A adds one new run with $r = 1$. The events B and C lengthen an existing run from r events to $r+1$. The events D and E join two existing runs of of r and s events to create a single run of length $r+s+1$. In our initial model, we neglect the merger of three or more runs of deletions. There is no conceptual difficulty in including three or more mergers, but the proliferation of embedded summations leads to computational problems. Thus we should expect the model to be adequate until θ gets very small, when mergers of several runs at a time become common.

The last lines of each of (2.4.2),(2.4.4) and (2.4.5) include the collection of terms, significantly cutting down on computing time when these formulae are implemented.

We define the change $\delta(r)$ in the number of runs of deleted terms with $r = 1, 2, \ldots$

as

$$\delta(1) = p_A - (p_B + p_C + 2p_D + 2p_E)\pi(1) \tag{2.4.7}$$

$$\delta(2) = (p_B + p_C)\pi(1) - (p_B + p_C + 2p_D + 2p_E)\pi(2). \tag{2.4.8}$$

For $r > 2$,

$$\delta(r) = (p_B+p_C)\pi(r-1)+(2p_D+2p_E)\sum_{s=1}^{r-2}\pi(s)\pi(r-s-1)-(p_B+p_C+2p_D+2p_E)\pi(r). \tag{2.4.9}$$

In an implementation on a finite interval of $\mathbb{Z}$, the number of runs of deleted terms will change from some value R to R', where

$$R' = R + \sum_{r=1}^{\infty}\delta(r), \tag{2.4.10}$$

and the distribution of run lengths will also change from π to π', with

$$\pi'(r) = \frac{R\pi(r) + \delta(r)}{R'} \quad , \tag{2.4.11}$$

where the mean increases accordingly from $\bar{u}$ to $\bar{u}'$, so that the mean $\bar{v}'$ of the new distribution ρ' of run lengths of undeleted terms satisfies

$$\bar{v}' = \frac{R}{R'}(\bar{u} + \bar{v}) - \bar{u}'. \tag{2.4.12}$$

The new proportion θ' of undeleted terms is $\bar{v}'/(\bar{u}' + \bar{v}')$.

We implement equations (2.4.1) to (2.4.12) as a recurrence with a step size parameter Λ to control the number of events using the same p_A, p_B, p_C, p_D and p_E and $\delta(\cdot)$ between successive normalizations, using $\Lambda\delta(\cdot)$ instead of $\delta(\cdot)$ in (2.4.10)—(2.4.12).

The choice of Λ determines the trade-off between computing speed and accuracy.

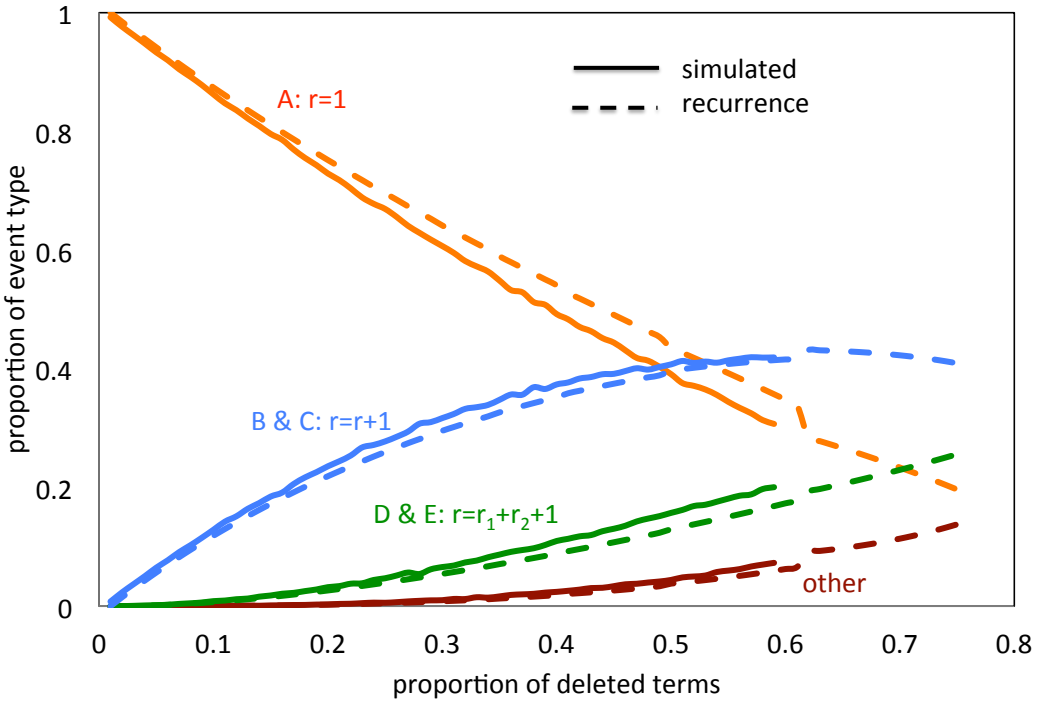

Figure 2.8: Changes in rates of different event types as calculated by recurrence, compared with simulation results.

Fig. 2.8 shows the results of our current implementation of our deterministic recurrence for the case $\mu = 2$. The results fit simulations of the stochastic model quite well. There are at least two reasons for the observed discrepancies. At the outset, since we used a large step size Λ for the computationally costly recurrence, its trajectory lags behind the simulation, especially with respect to the slower decrease in p_A and slower increase in $p_B + p_C$. Later discrepancies are partially due to not accounting for the merger of three or more runs. These can be estimated and are summarized as "other" in the diagram, but the quantities involved are not fed back to the recurrence through (2.4.11).

Other possible sources of error might be due to the cutoffs in x used for calculations involving $\gamma(x)$ and $\rho(x)$. However, extensive testing of various cutoff values has

indicated such errors to be negligible in our implementation.

2.5 Conclusions

We have developed a model for the fractionation process based on deletion events excising a geometrically distributed number of contiguous paralogs from one of a pair of homeologous chromosomes. This is extended to the mathematically less tractable case where both homeologs are susceptible to deletion events. The existence of data prompting this model is due to a functional biological constraint against deleting both copies of a duplicate pair of genes.

The mathematical framework we propose should eventually serve for testing the geometric excision hypothesis against alternatives such as gene-by-gene inactivations or imbalanced fractionation, although we have not developed these here.

A note on the sample size for the simulations: none of the simulations reported in this chapter were time-intensive. Our concern was simply to obtain results of sufficient precision so that no visible statistical fluctuation could by seen in the average values from one point to the next on the X-axis. In all cases, our sample size of 100 was more than adequate. In Fig. 2.8, for example, the departure of the curves for the recurrence from the desired curves were visualized by comparison with the simulations, which were considered to accurately represent the true behaviour of the model. A sample size of 10 would have been adequate, but we used 100 to place beyond question the accuracy of the simulations.

Simulations of these models indicate the feasibility of estimating the mean μ of the deletion event process from observations of the length of runs of single-copy genes and the overall proportion of single-copy genes. Application to real data from an earlier survey of 15 genomes descended from 6 WGD events, however, is hampered

by the accumulation of rearrangement events that have obscured most of the runs of single-copy genes. We have proposed a way of correcting for the missing runs, but this remains a rather approximate procedure.

The main outstanding question remains the exact derivation of π, the distribution of the number of deletion events contributing to a run of single-copy genes. The simulations are convenient in practice, since they depend on only one parameter μ as they evolve over time, but they give little mathematical insight. Our most important advance is a deterministic recurrence for $\pi(r)$ as the proportion θ of undeleted genes decreases, albeit for the one-sided model only. This takes into account the appearance of new runs over time, the lengthening of existing runs, as well as the merger of two existing runs with the new deletions to form a single, longer one. This calculation fits the process as simulated rather well and seems promising for further development.

Chapter 3

A model for biased fractionation after whole genome duplication

David Sankoff, Chunfang Zheng and Baoyong Wang. 2012. *BMC Genomics* 13:S1, S8 (2012). Dr. David Sankoff helped plan the research and write the paper. I was responsible for designing the algorithm and getting the results in this work. Also I collaborated with Dr. Sankoff in writing and preparing the manuscript, especially technical write-ups and graphics. Chunfang Zheng did simulations for our results.

3.1 Background

Whole genome doubling (WGD) creates two identical copies (*homeologs*) of each chromosome in a genome, with identical gene content and gene order. From this ensues the wholesale shedding of duplicate genes over evolutionary time through random *excision* — elimination of excess DNA — namely the deletion of chromosomal segments containing one or more genes, or through gene-by gene events such as epigenetic silencing and pseudogenization [1, 9, 10, 12, 15].

When a duplicate gene is lost, it may be lost from one copy (*homeolog*) of a chromosome or the other, but generally not both, because of the necessity of conserving function. This *fractionation* creates an interleaving pattern; the full original gene complement becomes apparent only by *consolidating* [13] the two homeologous single-copy regions. In most cases, there is a degree of bias, more genes being lost from one of the homeologous regions than the other [1, 10, 12, 13]. Fractionation is an important process in many evolutionary domains, in particular the flowering plants, since it results in a genome that is highly scrambled with respect to its pre-WGD ancestor. For this reason as well, fractionation raises a number of interesting and difficult problems for comparative genomics.

The study of fractionation is basically a study of runs; that is, runs of duplicate genes on two homeologous chromosomes alternating with runs of single-copy genes on one or both of these chromsomes. Because of the way these runs are generated biologically, and because they involve two chromosomes evolving in a non-independent way, standard statistical or combinatorial run analyses are not directly applicable.

In this paper, we present a detailed version of the excision model of fractionation with geometrically distributed deletion lengths, for which we previously analyzed a tractable, but biologically unrealistic, special case [14]. The key problem in this

field is to determine μ, the mean of the hypothesized geometric distribution $\rho(\frac{1}{\mu}, \cdot)$, since this bears directly on the main biological question of the relative importance of random excision versus gene-by-gene inactivation. The relevant data consist of runs of single-copy genes (whose duplicates have been lost from the homeologous region) as well as runs of remaining duplicate pairs in two homeologous regions. The inference of μ is complicated since each run of l single copies may have been produced by an unknown number r of deletion events, either $r = l$ events (the gene-by-gene model) or $1 \leq r < l-1$ (the random excision model), and these r samples of the distribution ρ turn out not to be independent. Thus a fundamental aspect of finding μ, and hence $\rho(\frac{1}{\mu}, \cdot)$, is to derive $\pi(r)$, the proportion of runs of single-copy genes with r terms, for $r = 1, 2, \ldots$

A further complication arises from the way deletion events accumulate into longer runs of single-copy genes. The deletion of a certain number of duplicate genes may overlap the site of a previous deletion event on the *same* chromosome, but it is blocked by the functional constraint (mentioned above) as soon as it starts to overlap the site of a previous deletion event on the *homeologous* chromosome.

Another biologically important question is to determine ϕ, the proportion of deletion events that operate on one of the homeologous chromosomes, while a proportion $1 - \phi$ operates on the other. We explored this question at some length in [10], but a detailed mathematical treatment of the effects of this "fractionation bias" remains to be done.

It is not difficult to simulate the fractionation process, but this gives little insight into its mathematical structure. Given that it is unlikely for any closed form of π to exist, nor for any simple computing formula, our goal here is to develop a recurrence for the distribution of $\pi(r)$ for $r = 1, 2, \ldots$ as a function of μ, ϕ and θ (the proportion of duplicate pairs remaining in the genome versus single-copy genes).

This work is an attempt at creating a rigorous "null" model of duplicate loss, based on parameters μ, ϕ and θ. This should provide a principal basis for developing statistical tests on real WGD descendants, to see if the geometric excision hypothesis is acceptable and to see if fractionation is unbiased or not. We will not explicitly investigate the alternative hypothesis of gene-by-gene deletion, nor do we take chromosomal rearrangement events into account; our task here is simply to set up the null statistical model with a view to enabling useful statistical tests of hypotheses for this problem.

3.2 The models

The structure of the data

The data on paralog reduction are of the form (G, H), where G and H are binary sequences indexed by $\mathbb{Z}$, satisfying the condition that $g(i) + h(i) > 0$. This condition models the prohibition against deleting both copies of a duplicated gene. We may also assume that whatever process generated the 0s and 1s is homogeneous on $\mathbb{Z}$.

The sequence $G + H$ consists of alternating runs of 1's and 2's. We denote by $p(l), l \geq 1$ the probability distribution of length of runs of 1's. For any finite interval of $\mathbb{Z}$ we denote by $f(l), l \geq 1$ the empirical frequency distribution of length of runs of 1's.

The use of $\mathbb{Z}$ instead of a finite interval is consistent with our goal of getting to the mathematical essence of the process, without any complicating parameters such as interval length. In practice, we use long intervals of at least 100,000 so that any edge effects will be negligible. See [11, 15] for *ad hoc* ways of handling biological scale intervals.

3.2.1 The deletion events

Let ϕ, where $0 \leq \phi \leq 1$, be the fractionation bias. We assume a continuous time process, parameter $\lambda(t) > 0$, only to ensure no two events occur at the same time.

- We start ($t = 0$) with $h(i) = g(i) = 1$ for all i.
- At any $t > 0$, consider any i where $h(i) = g(i) = 1$. With probability $\lambda(t)dt$, a *deletion event* occurs *anchored* at position i: we choose a positive number a according to a geometric variable $\mathbf{y}$ with parameter $1/\mu$; i.e., $P[\mathbf{y} = a] = \gamma(a) = \frac{1}{\mu}\left(1 - \frac{1}{\mu}\right)^{a-1}$, $a \geq 1$.
- Then, with probability ϕ, we choose to carry out the deletion on G with probability $1 - \phi$, on H.
- If the deletion is on G, we convert $g(i) = 0, g(i+1) = 0, \ldots, g(i+a-1) = 0$ unless a "collision" occurs.
- One type of collision, a *skippable* collision, arises when one or more of $g(i+1), \ldots, g(i+a-1)$ is already 0. In this case we skip over the existing 0 values and continue to convert the next available 1s into 0s, until a total of a 1s have been converted, or a collision of the second type is encountered.
- The second type of collision, *blocking* collision, arises when one or more of $h(i+1), \ldots, h(i+a-1)$ (or a further term if skipping has already occurred during this event) is already 0. In this case, further conversions of 1s to 0s are blocked, starting with the first $g(x)$ for which $h(x) = 0$.

Skippable collisions are a natural way to model the excision process, since deletion of duplicates and the subsequent rejoining of the DNA directly before and directly after the excised fragment means that this fragment is no longer "visible" to the

deletion process. Observationally, however, we know deletion has occurred because we have access to the sequence H, which retains copies of the deleted terms. Blocking collisions are a natural way of modeling the constraint against deleting single-copy genes.

When the deletion event has to skip over previous 0s, this hides the anchor i and length a of previous deletion events. Denote by $\mathbf{r}$ the random variable indicating the total number of deletion events responsible for a run. Then, given $\mathbf{r} = r$, the run length $\mathbf{z}$ is distributed as the sum of r geometric variables, which would result in a negative binomial distribution were these geometric variables independent. They are not, however, since events with large a tend to group together in runs with large r, while an event with small a is more likely to constitute by itself a run with $r = 1$ [13]. If we observe G at some point in time, as in the last pair of rows of Table 1, all we can observe are the run lengths of 0s and 1's. We cannot observe the a, i or r, while t and $\lambda(t)$ are unknown and, as we shall see, only mathematical conveniences that are supplanted by θ in our calculations. The parameters about which we wish to make statistical inferences are the deletion length distribution parameter μ, and the fractionation bias ϕ, since it is these quantities that are at the heart of the biological controversies about paralog reduction. This inference can only be based on the two observable quantities: the run lengths l and the proportion θ of remaining (undeleted) 1's.

3.3 Results

Simulations to determine π

We carried out simulations on an interval of $\mathbb{Z}$ of length 100,000. This enabled us to use a discrete-time process instead of the continuous-time process on $\mathbb{Z}$. The "anchors" for the deletion events were chosen at random among the currently undeleted genes. The remaining steps were carried out as described in the previous section and Table 3.1. Because each simulation run samples thousands of deletions, it sufficed to select 100 runs for each value of the parameters μ and ϕ studied.

event	i	a	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	r
start			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
1	-1	3	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1
			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2			1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1
	-4	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
3	5	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1,1
			1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
4	4	3	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1,2
			1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
5			1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	1	2
	-5	4	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	3

Table 3.1: Five deletion events affecting two homeologous chromosomes, leading to two runs of single-copy genes. The fourth step illustrates the "skip" process, at $i = 5$ where the pre-existing deletion is incorporated into a longer run with $r = 2$. The fifth step shows how further deletion (at $i = -1$) and the "skip" process (to $i = 2$) are blocked when a single-copy gene is encountered ($i = -1$) on the homeologous chromosome. This creates a single-copy run with length $l = 7$ and $r = 3$, part on one chromosome, part on the other. Note that r is not observable from the genome data.

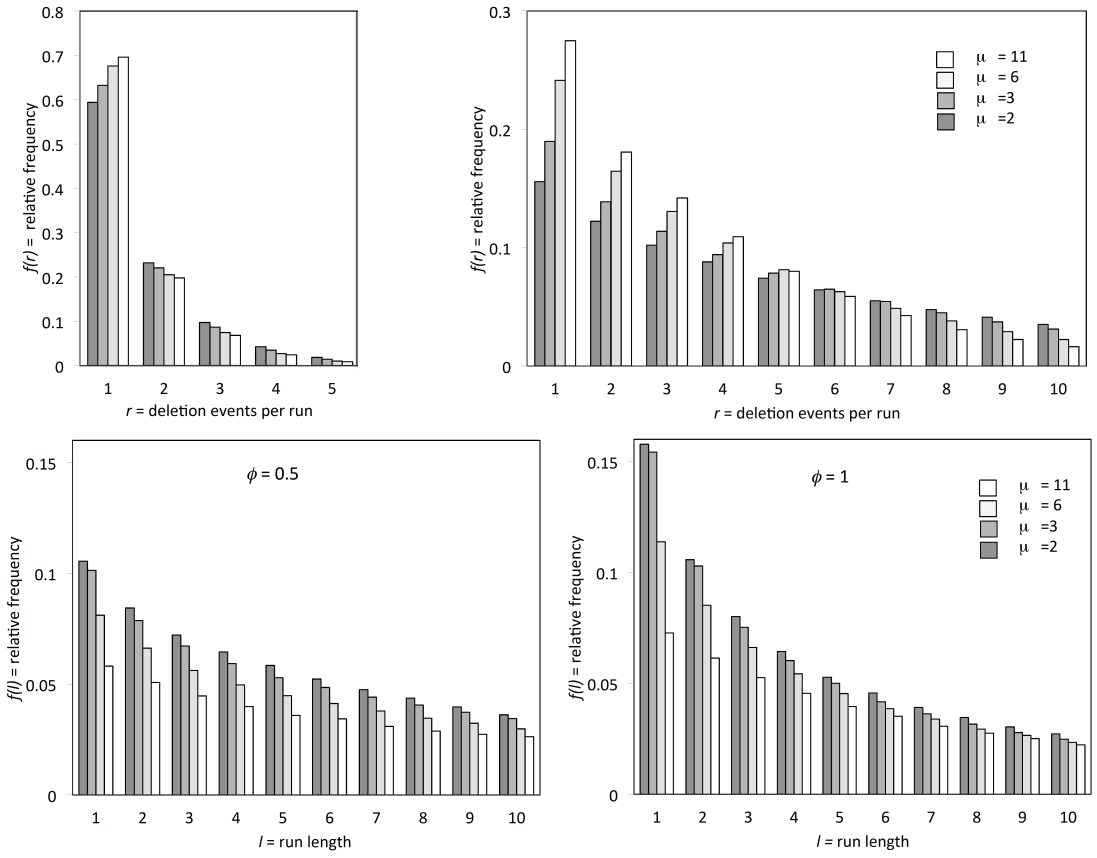

Figure 3.1: Distribution of number of deletion events r composing each run when $1 - \theta$, the proportion of sequence deleted, is 0.5 (top left) and 0.9 (top right); ϕ= 0.5 in both cases. Distribution of run lengths for for $\phi = 0.5$ (bottom left) and $\phi = 1$ (bottom right). For visibility, all diagrams show highest frequency parts of the distribution only.

Fig. 3.1 compares $\pi(r)$ when $\theta = 0.5$ and $\theta = 1$, for $\mu = 2, 3, 6$ and 11, when $\phi = 0.5$. We can see that the number of deletion events contributing to a run is somewhat dependent on μ when half of the the sequence has been deleted, but is strongly dependent when 90% has been deleted. In the bottom row, the graph on the left shows that run-length l is distributed very differently for $\mu = 2, 3, 6$ and $\mu = 11$ when the proportion of the sequence deleted is exactly the same. This strongly suggests that observing the run-length distribution and the overall proportion of deletions should allow us to infer μ. Moreover, the shape of these distributions is sensitive to ϕ.

We mention that any edge effects in our simulation are negligible. Whether we work with G and H on an interval of $\mathbb{Z}$ of length 100,000 or, as previously [13], length 300,000, gives virtually the same results.

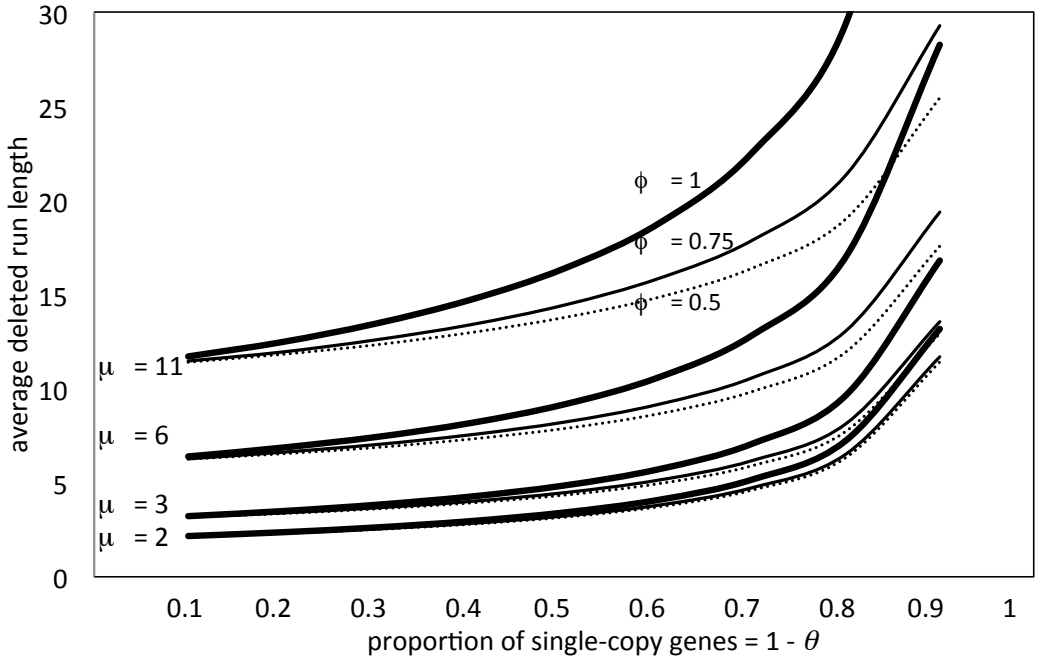

Figure 3.2: Average length of run of single copy genes in for $\theta = 0.5, 0.75, 1.0$, for $\mu = 2, 3, 6$ and 11

Fig. 3.2 shows the relationship, for three values of the fractionation bias ϕ and for a range of values of μ, between the proportion of genes deleted, on one chromosome or the other, and the average run length. This confirms that average run length and overall proportion of deletion θ, both observable, can be used to infer μ rather accurately and to infer ϕ, perhaps with somewhat less precision. The latter parameter can, however, be inferred from the shape of the run length distribution in Fig. 3.1 (bottom) or estimated directly from the proportion of single-copy genes on each homolog.

3.4 A recurrence for $\pi(r)$

We are interested in inferring μ from the observed distribution of run lengths and the proportion θ of undeleted terms; i.e., undeleted genes. At the outset, $\theta = 1$. As $t \to \infty$, $\theta \to 0$. We are not, however, interested in t, since it is not observable and any time-based inference we can make about μ will depend only on run lengths and θ in any case. On the other hand, r, the number of deletion events per run, is an interesting variable since we can assume run length is close to $r\mu$ on average, at least for small values of θ, and we can model the evolution of r directly We consider the distribution π as a function of μ, ϕ and θ.

As π changes, the probability weight is redistributed among several types of run:

1. new runs ($r = 1$) falling completely within an existing run of undeleted terms, not touching the preceding or following run of deleted terms, type A in Fig. 3.3,

2. runs that touch, overlap or entirely engulf exactly one previous run of deleted terms with $r \geq 1$, thus lengthening that run to $r + 1$ events, types B and C in Fig. 3.3,

3. runs that touch, overlap or engulf, by the skipping process, two previous runs of r_1 and r_2 events respectively, creating a new run of $r_1 + r_2 + 1$ events, and diminishing the total number of runs by 1, including types D and E in Fig. 3.3,

4. runs that touch, overlap or engulf, by the skipping process, $k > 2$ previous runs of $r_1, \ldots, r_k$ events respectively, creating a new run of $r_1 + \cdots + r_k + 1$ events, and diminishing the total number of runs by $k-1$, (not illustrated in Fig. 3.3). Case 3 above may be considered a special case of this for $k = 2$ and Case 2 for $k = 1$.

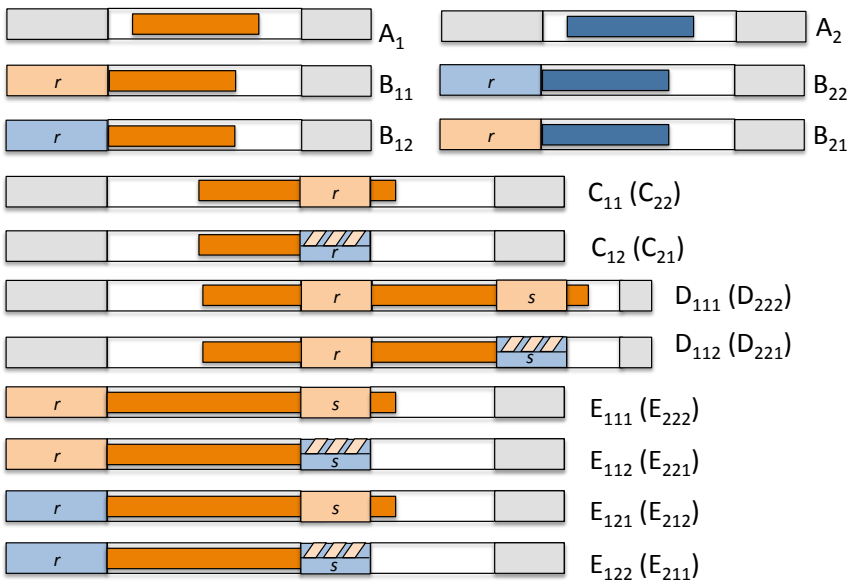

Figure 3.3: Types of deletion event affecting less than three pre-existing runs. Red and blue shading distinguishes between deletions from the two homeologous chromosomes. Grey areas represent previous deletions from either chromosome. White area indicates run of undeleted terms. Lightly shaded area indicates run of previously deleted terms. Darker area represents current deletion event. Hatched striped area above lightly shaded area indicates either previous deletions from both homeologous chromosomes, or only from the homeolog not affected by the current deletion. A: creates one new run with r = 1. B: lengthens left hand run to r + 1 events. C: lengthens the right hand-run to r + 1 events. D and E: merge two runs to create a single run with r + s + 1 deletion events.

The first process, involving a deletion event of length a requires a run of undeleted terms of at least $a + 2$. What can we say about runs of undeleted terms? We know that runs of deleted terms alternate with runs of undeleted terms, so that there is one run of the former for each of the latter. The mean lengths $\bar{u}$ and $\bar{v}$ of the deleted runs and the undeleted runs, respectively, should satisfy

$$\bar{v} = \frac{\theta}{1-\theta}\bar{u}. \tag{3.4.1}$$

The distribution $\rho(l)$ of lengths of the undeleted runs is assumed to be geometric. Similarly, the lengths of successive undeleted runs (indeed all undeleted runs) are assumed to be independent. While we do not have a rigorous proof of these assumptions, they have been confirmed by extensive simulations.

Let ϕ_1 and ϕ_2 be the proportion of deletion events affecting homeologous chromosomes 1 and 2, respectively, so that $\phi_1 + \phi_2 = 1$. Let $\tau(r)$ be the proportion of runs of single-copy genes with terms in both chromosomes. ($\tau(1) \equiv 0$ and, initially, $\tau(r) = 0$ for $r = 2, 3, \ldots$) Note that, in such a run, the term at the extreme left was deleted from chromosome i with probability ϕ_i and the same for the terms at the extreme right. The proportion of undeleted terms in runs of length l is $l\rho(l)/E_\rho$, where $E_\rho = \sum_{l>0} l\rho(l)$. As depicted in Fig. 3.3, the probabilities p_{A_1} and p_{A_2}, that a deletion event affects chromosomes 1 or 2, respectively, and falls within a run of

undeleted terms of length l without deleting the terms at either end is, for $i \in \{1, 2\}$

$$\begin{aligned} p_{A_i} &= \phi_i \sum_{l>2} \frac{l\rho(l)}{E_\rho} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{\phi_i}{E_\rho} \sum_{l>2} \rho(l) \sum_{j=2}^{l-1} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{\phi_i}{E_\rho} \sum_{l>2} \rho(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a), \end{aligned} \tag{3.4.2}$$

where j indexes the starting position of the deletion within the run and a is the number of terms deleted in the event. We define the *contribution to mean run length* of A events to be

$$\mu_A = \sum_{i=1}^{2} \frac{\phi_i}{E_\rho} \sum_{l>2} \rho(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a)a. \tag{3.4.3}$$

Events of type A_i create runs of deleted terms with $r = 1$ from one chromosome only. Note that the last line of equation (3.4.2), and equation (3.4.3), involve the collection of terms, reducing the number of nested summations in order to speed up calculation. While these are not lengthy calculations to start with, we display the speed-up as a simple illustration of the important efficiencies implemented for more difficult cases to be treated below. The probability $p_{B_{if}}$ that a deletion event on chromosome i touches only the run of deletions on chromosome f on the left of the run of undeleted terms is, for $i \in \{1, 2\}$ and $f \in \{1, 2\}$,

$$p_{B_{if}} = \frac{\phi_i \phi_f}{E_\rho} \sum_{l>1} \rho(l) \sum_{a=1}^{l-1} \gamma(a). \tag{3.4.4}$$

We define the contribution to mean run length of B events to be

$$\mu_B = \sum_{i=1}^{2}\sum_{f=1}^{2}\frac{\phi_i\phi_f}{E_\rho}\sum_{l>1}\rho(l)\sum_{a=1}^{l-1}\gamma(a)a. \tag{3.4.5}$$

Events of type B_{ii} turn a deleted run with r events from one chromosome, into a run with $r+1$ events. Events of type B_{if}, with $i \neq f$, turn a deleted run with r events, into a run with $r+1$ events. The probability $p_{C_{ii}}$ that a deletion event, on either chromosome, does not touch the run of deletions on the left, does touch or overlap the run of deletions on the right entirely on the same chromosome (homeolog), but does not extend over the entire run of undeleted terms beyond that is, for $i \in \{1,2\}$:

$$\begin{aligned} p_{C_{ii}} &= \frac{\phi_i^2(1-\tau)}{E_\rho}\sum_{l>1}\sum_{k\geq 1}\rho(l)\rho(k)\sum_{j=2}^{l}\sum_{a=l-j+1}^{l-j+k}\gamma(a) \\ &= \frac{\phi_i^2(1-\tau)}{E_\rho}\sum_{l>1}\sum_{k\geq 1}\rho(l)\rho(k) \\ &\times \left(\sum_{a=1}^{\min[l-2,k-1]} a\gamma(a) + \sum_{a=\min[l-1,k]}^{\max[l-1,k]}\min[l-1,k]\gamma(a) + \sum_{a=\max[l,k+1]}^{l+k-2}(l+k-a-1)\gamma(a)\right). \end{aligned} \tag{3.4.6}$$

We define the contribution to mean run length of C_{ii} events to be

$$\mu_{C_{ii}} = \sum_{i=1}^{2}\frac{\phi_i^2(1-\tau)}{E_\rho}\sum_{l>1}\sum_{k\geq 1}\rho(l)\rho(k)\sum_{j=2}^{l}\sum_{a=l-j+1}^{l-j+k}\gamma(a)a, \tag{3.4.7}$$

which can be calculated using an expansion such as that in (3.4.6). Events of type C_{ii} turn a deleted run with r events from one chromosome into a run with $r+1$ events.

The probability $p_{C_{if}}$ that a deletion event, on either chromosome, does not touch the run of deletions on the left but does touch the run of deletions on the right, partly

or entirely on the other chromosome, is, for $i \neq f \in \{1, 2\}$:

$$p_{C_{if}} = \frac{\phi_i\tau + \phi_i\phi_f(1-\tau)}{E_\rho} \sum_{l>1} \rho(l) \sum_{j=2}^{l} \sum_{a=l-j+1}^{\infty} \gamma(a). \tag{3.4.8}$$

We define the contribution to mean run length of C_{if} events to be

$$\mu_{C_{if}} = \sum_{i\neq f=1}^{2} \frac{\phi_i\tau + \phi_i\phi_f(1-\tau)}{E_\rho} \sum_{l>1} \rho(l) \sum_{j=2}^{l} (l-j+1) \sum_{a=l-j+1}^{\infty} \gamma(a). \tag{3.4.9}$$

Events of type C_{if}, with $i \neq f$, turn a deleted run with r events, into a run with $r+1$ events. Note that (3.4.9) does not contains terms of form $a\gamma(a)$ as do (3.4.3),(3.4.5) and (3.4.7), since deletion in this event is blocked beyond the existing run of deletions; the probability weight is thus concentrated on deletions of lesser length.

The probability $p_{D_{iii}}$ that a deletion event completely overlaps the run of deletions on the right and touches or overlaps the run of deletions beyond that, all on the same chromosome, but does not extend over a further run of undeleted terms is:

$$\begin{aligned} p_{D_{iii}} &= \frac{\phi_i^3(1-\tau)^2}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \sum_{h\geq 1} \rho(l)\rho(k)\rho(h) \sum_{j=2}^{l} \sum_{a=l-j+k+1}^{l-j+k+h} \gamma(a) \\ &= \frac{\phi_i^3(1-\tau)^2}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \sum_{h\geq 1} \rho(l)\rho(k)\rho(h) \\ &\times \left(\sum_{a=k+1}^{\min[l+k-2,\, h+k-1]} (a-k)\gamma(a) + \sum_{a=\min[l+k-1,\, k+h]}^{\max[l+k-1,\, k+h]} \min[l-1, h]\gamma(a) + \sum_{a=\max[l+k,\, k+h+1]}^{l+k+h-2} (l+k+h-a-1)\gamma(a) \right) \end{aligned} \tag{3.4.10}$$

in which the reduction of the number of nested summations is key to the computability of the entire calculation. We define the contribution to mean run length of D_{iii} events

to be

$$\mu_{D_{iii}} = \frac{\phi_i^3(1-\tau)^2}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \sum_{h\geq 1} \rho(l)\rho(k)\rho(h) \sum_{j=2}^{l} \sum_{a=l-j+k+1}^{l-j+k+h} \gamma(a)a, \tag{3.4.11}$$

which can be calculated using an expansion such as that in (3.4.10). Events of type D_{iii} turn two deleted runs with r and s events, respectively, both from the same chromosome, into a run with $r+s+1$ events. The probability $p_{D_{iif}}$ that a deletion event completely overlaps the run of deletions on the right, on the same chromosome, and touches the run of deletions beyond that, partly or entirely on the other chromosome, is:

$$p_{D_{iif}} = \frac{\phi_i^2(1-\tau)\tau + \phi_i^2\phi_f(1-\tau)^2}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{j=2}^{l} \sum_{a=l-j+k+1}^{\infty} \gamma(a) \tag{3.4.12}$$

and the contribution to mean run length is

$$\mu_{D_{iif}} = \frac{\phi_i^2(1-\tau)\tau + \phi_i^2\phi_f(1-\tau)^2}{E_\rho} \sum_{l>1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{j=2}^{l} (l-j+k+1) \sum_{a=l-j+k+1}^{\infty} \gamma(a). \tag{3.4.13}$$

Events of type D_{iif}, with $i \neq f$, turn two deleted runs with r and s events, respectively, with the latter containing terms from both chromosomes, into a single run with $r+s+1$ events.

The probability $p_{E_{iii}}$ that a deletion event touches the run of deletions on the left of the run of undeleted terms and touches or overlaps the run of deletions on the right, all on the same chromosome, but does not extend over the entire run of

undeleted terms beyond that is:

$$p_{E_{iii}} = \frac{\phi_i^3(1-\tau)}{E_\rho} \sum_{l\geq 1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{a=l}^{l+k-1} \gamma(a), \tag{3.4.14}$$

The contribution to mean run length is

$$\mu_{E_{iii}} = \frac{\phi_i^3(1-\tau)}{E_\rho} \sum_{l\geq 1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{a=l}^{l+k-1} \gamma(a)a. \tag{3.4.15}$$

The probability $p_{E_{iif}}$ that a deletion event touches the run of deletions on the left of the run of undeleted terms, both from the same chromosome, and touches the run of deletions on the right, partly or entirely on the other chromosome, is:

$$p_{E_{iif}} = \frac{\phi_i^2\tau + \phi_i^2\phi_f(1-\tau)}{E_\rho} \sum_{l\geq 1} \rho(l) \sum_{a=l}^{\infty} \gamma(a). \tag{3.4.16}$$

The contribution to mean run length is

$$\mu_{E_{iif}} = \frac{\phi_i^2\tau + \phi_i^2\phi_f(1-\tau)}{E_\rho} \sum_{l\geq 1} \rho(l)l \sum_{a=l}^{\infty} \gamma(a). \tag{3.4.17}$$

The probability $p_{E_{iii}}$ that a deletion event touches the run of deletions on the left of the run of undeleted terms and touches or overlaps the run of deletions on the right, all on the same chromosome, but does not extend over the entire run of undeleted terms beyond that is:

$$p_{E_{ifi}} = \frac{\phi_i^2\phi_f(1-\tau)}{E_\rho} \sum_{l\geq 1} \sum_{k\geq 1} \rho(l)\rho(k) \sum_{a=l}^{l+k-1} \gamma(a). \tag{3.4.18}$$

The contribution to mean run length is

$$\mu_{E_{ifi}} = \frac{\phi_i^2 \phi_f (1-\tau)}{E_\rho} \sum_{l\geq 1} \sum_{k \geq 1} \rho(l)\rho(k) \sum_{a=l}^{l+k-1} \gamma(a) a \tag{3.4.19}$$

The probability $p_{E_{iff}}$ that a deletion event touches the run of deletions on the left of the run of undeleted terms and touches or overlaps the run of deletions on the right, all on the same chromosome, but does not extend over the entire run of undeleted terms beyond that is:

$$p_{E_{iff}} = \frac{\phi_i \phi_f \tau + \phi_i \phi_f^2 (1-\tau)}{E_\rho} \sum_{l\geq 1} \rho(l) \sum_{a=l}^{\infty} \gamma(a). \tag{3.4.20}$$

The contribution to mean run length is

$$\mu_{E_{iff}} = \frac{\phi_i \phi_f \tau + \phi_i \phi_f^2 (1-\tau)}{E_\rho} \sum_{l\geq 1} \rho(l) l \sum_{a=l}^{\infty} \gamma(a) \tag{3.4.21}$$

Events of type E_{iii} turn two deleted runs with r and s events, respectively, all from one chromosome, into a single run with $r+s+1$ events. Events of type E_{iif}, E_{ifi} and $E_{iff},$, with $i \neq f$, turn two deleted runs with r and s events, respectively, into a single run with $r+s+1$ events.

We reiterate here that the last lines of each of (3.4.2), (3.4.6) and (3.4.10) include the collection of terms, significantly cutting down on computing time when these formulae are implemented, especially in the case of (3.4.10).

In this initial model, we neglect the merger of three or more runs of deletions. There is no conceptual difficulty in including three or more mergers, but the proliferation of embedded summations would require excessive computation. Thus we should expect the model to be adequate until θ gets very small, when mergers of several runs

at a time become common.

Let $p_A = p_{A_1} + p_{A_2}$, and similarly let each of $p_B, \ldots, p_E$ be the sums of their respective subscripted terms (with all combinations of i and f). We define the change $\delta_\pi(r)$ in the number of runs of deleted terms with $r = 1, 2, \ldots$

$$\delta_\pi(1) = p_A - (p_B + p_C + 2p_D + 2p_E)\pi(1) \tag{3.4.22}$$

$$\delta_\pi(2) = (p_B + p_C)\pi(1) - (p_B + p_C + 2p_D + 2p_E)\pi(2). \tag{3.4.23}$$

For $r > 2$,

$$\delta_\pi(r) = (p_B+p_C)\pi(r-1)+(2p_D+2p_E)\sum_{s=1}^{r-2}\pi(s)\pi(r-s-1)-(p_B+p_C+2p_D+2p_E)\pi(r). \tag{3.4.24}$$

In an implementation on a finite interval of $\mathbb{Z}$, the number of runs of deleted terms will change from some value R to R', where

$$R' = R + \sum_{r=1}^{\infty} \delta_\pi(r). \tag{3.4.25}$$

The distribution of number of events per run will also change from π to π', where

$$\pi'(r) = \frac{R\pi(r) + \delta_\pi(r)}{R'}, \tag{3.4.26}$$

and where the mean of the number of deleted genes per run increases from $\bar{u}$ to $\bar{u}'$, so that

$$\bar{u}' = \frac{R\bar{u} + \sum_{X=A,B,C.,D.,E.} \mu_X}{R'}. \tag{3.4.27}$$

The mean $\bar{v}'$ of the new distribution ρ' of run lengths of undeleted terms satisfies

$$\bar{v}' = \frac{R}{R'}(\bar{u} + \bar{v}) - \bar{u}'. \tag{3.4.28}$$

The new proportion θ' of undeleted terms is $\bar{v}'/(\bar{u}' + \bar{v}')$.

In the same interval of $\mathbb{Z}$, we define the change $\delta_\tau(r)$ in the number of runs containing single copy genes in both chromosomes with $r = 1, 2, \ldots$

$$\delta_\tau(1) = 0. \tag{3.4.29}$$

$$\delta_\tau(2) = (p_{B_{12}} + p_{B_{12}} + p_{C_{12}} + p_{C_{21}})\pi(1) - (p_B + p_C + 2p_D + 2p_E)\pi(2)\tau(2) \tag{3.4.30}$$

For $r > 2$,

$$\begin{aligned}\delta_\tau(r) &= (p_B + p_C)\pi(r-1)\tau(r-1) + (p_{B_{12}} + p_{B_{12}} + p_{C_{12}} + p_{C_{21}})\pi(r-1)(1 - \tau(r-1)) \\ &+ (2p_D + 2p_E)\sum_{s=1}^{r-2}\pi(s)\pi(r-s-1)(1 - (\phi_1^3 + \phi_2^3)[1 - \tau(r-s-1)][1 - \tau(s)]) \\ &- (p_B + p_C + 2p_D + 2p_E)\tau(r)\pi(r). \end{aligned} \tag{3.4.31}$$

In the implementation, the number of runs of deleted terms with genes on both chromosomes will change from $T(r)$ to $T'(r)$, where

$$T'(r) = T(r) + \delta_\tau(r). \tag{3.4.32}$$

The proportions of runs with deletion events from both chromosomes will also change from τ to τ', where

$$\tau'(r) = \frac{T'(r)}{R'\pi'(r)}. \tag{3.4.33}$$

We implement equations (3.4.1) to (3.4.33) as a recurrence with a step size pa-

rameter Λ to control the number of events using the same $p_A, p_B, p_C, p_D, p_E, \delta_\pi(\cdot)$ and $\delta_\tau(\cdot)$ between successive normalizations, and using $\Lambda\delta_\pi(\cdot)$ and $\Lambda\delta_\tau(\cdot)$ instead of $\delta_\pi(\cdot)$ and $\delta_\tau(\cdot)$ in (3.4.25)—(3.4.33). The choice of Λ determines the trade-off between computing speed and accuracy.

Figure 3.4 shows the results of our current implementation of our deterministic recurrence for the cases $\mu = 2$ and $\mu = 11$, for unbiased fractionation ($\phi = 0.5$) and for extremely biased fractionation ($\phi = 1$). The results fit simulations of the stochastic model quite well and reveal a number of tendencies. One is that unbiased fractionation with small deletions leads to the fastest drop in events of type A as θ decreases. Biased fractionation with large deletion sizes leads to slow initial growth in the proportions of events of types D and E and "other".

There are at least two reasons for the discrepancies between the simulations and the recurrences observed in Fig. 3.4. At the outset, since we used a large step size Λ for the computationally costly recurrence, its trajectory lags behind the simulation, especially with respect to the slower decrease in p_A and slower increase in $p_B + p_C$. Later discrepancies are partially due to not accounting for the merger of three or more runs. These can be estimated and are summarized as "other" in the diagram, but the quantities involved are not fed back to the recurrence through Eq. (3.4.26).

Other possible sources of error might be due to the cutoffs in x used for calculations involving $\gamma(x)$ and $\rho(x)$. However, extensive testing of various cutoff values has indicated such errors to be negligible in our implementation.

3.5 Conclusions

We have developed a model for the fractionation process based on deletion events excising a geometrically distributed number of contiguous paralogs from either one

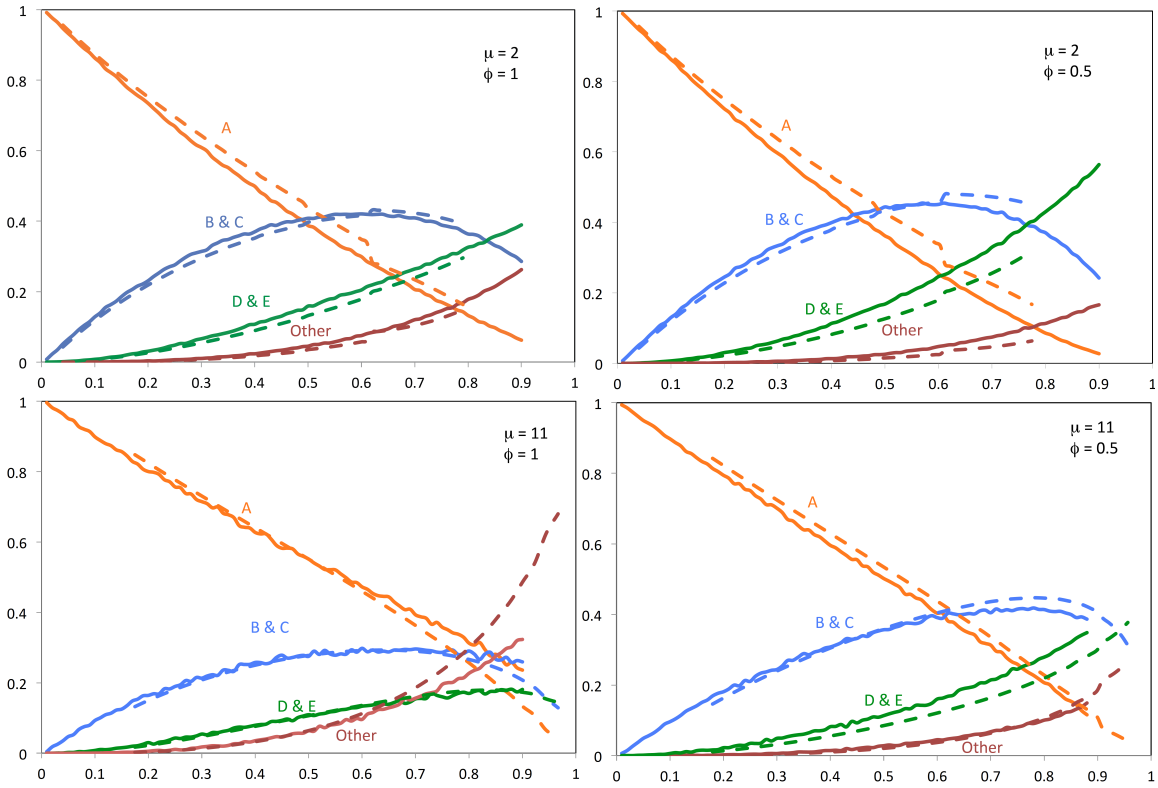

Figure 3.4: Changes in rates of different event types as calculated by recurrence (dashed lines), compared with simulation results (solid lines). Horizontal axis: Proportion of duplicates deleted is $1 - \theta$. Vertical axis: proportion of event type.

of a pair of homeologous chromosomes. The existence of data prompting this model is due to a functional biological constraint against deleting both copies of a duplicate pair of genes.

The mathematical framework we propose should eventually serve for testing the geometric excision hypothesis against alternatives such as single gene-by-gene inactivations, although we have not developed this in this paper. In addition, further developments could treat the gene-by-gene inactivation model as the null hypothesis, and the geometric excision model, with mean greater than 1, as the alternative hypothesis.

Simulations of these models indicate the feasibility of estimating the mean μ of

the deletion event process and the fractionation bias ϕ from observations of the length of runs of single-copy genes and the overall proportion of single-copy genes.

The main question we have explored is the exact derivation of π, the distribution of the number of deletion events contributing to a run of single-copy genes. The simulations are convenient in practice, since they depend only on the parameters μ and ϕ as they evolve over time, but they give little mathematical insight. Our most important advance is a deterministic recurrence for $\pi(r)$ as the proportion θ of undeleted genes decreases. This takes into account the appearance of new runs over time, the lengthening of existing runs, and the merger of two existing runs with the new deletions to form a single, longer one. This calculation fits the process as simulated rather well and seems promising for further development.

In order to validate our fractionation model empirically, we will have to expand it to incorporate the rearrangement events that are pervasive in genome evolution. Our previous work on this problem shows that the effect of rearrangement is to seriously bias the observable, credible instances of fractionation towards smaller runs of deleted genes [12, 13]. Future work on this difficult problem will either have to rely on careful modeling of this ascertainment bias or else find a way to incorporate into the model deleted runs that have been interrupted by rearrangements.

Chapter 4

Discriminating between structural and functional mechanisms for duplicate gene loss following whole genome doubling

Baoyong Wang, Chunfang Zheng, Carlos Fernando Buen Abad Najar, David Sankoff submitted to *ICCABS 2014*.

Dr. David Sankoff helped plan the research and write the paper. I was responsible for designing the algorithm and getting the results in this work. Also I collaborated with Dr. Sankoff in writing and preparing manuscript, especially technical write-ups and graphics. Chunfang Zheng and Carlos Fernando Buen Abad Najar carried out the simulations for our results.

4.1 Abstract

The loss of duplicate genes after WGD is the subject to a debate as to whether it proceeds gene by gene or through deletion of multi-gene chromosomal segments. We assume deletion events excise a geometrically distributed number of genes with mean $\mu \geq 1$, and these events can combine to produce deleted runs of length l. If $\mu = 1$, the process is gene-by-gene. If $\mu > 1$, the process at least occasionally excises more than one gene at a time. In the latter case if deletions overlap, the later one simply extends the existing run of single-copy genes. We explore aspects of the predicted distribution of the lengths of single-copy regions analytically, but resort to simulations to show how observing run lengths l allows us to discriminate between the two hypotheses.

4.2 Introduction

The process of WGD gives rise to two copies of each chromosome in a genome, containing the same genes in the same order. One of each pair of duplicate genes is lost over evolutionary time, with the exception of a small number of pairs where the two copies are retained, usually by virtue of diverging in both sequence and function.

An important biological controversy arises from the question of whether duplicated genes are deleted through random excision — elimination of excess DNA — namely, the deletion of chromosomal segments containing one or more genes [11], which we term the "structural" mechanism, or through gene-by gene events such as epigenetic silencing and pseudogenization [9], which are "functional" mechanisms. This question is important to evolutionary theory because it speaks directly to the role of WGD, and gene duplication in general, in disrupting gene order, in creating functional innovation, and in the radiation of new species. It is a question of whether

selection operates on the level of simply permitting non-lethal deletions or whether more subtle effects are in play, such as dosage balance of interacting genes.

This debate may be formulated in terms of deletion events removing a number X of contiguous genes, where X is drawn from a geometric distribution γ with mean μ. Here the one-at-a-time deletion model is represented by $\mu = 1$, while the random number of deletions at a time holds if $\mu > 1$. In the latter case, the possibility of two overlapping events is handled by a biologically realistic additive run-length assumption.

In this paper we investigate the discrimination problem of choosing between the two models based on deletion run-length statistics (resulting from overlapping deletion events). This involves comparing an observed genome containing single-copy genes, originally members of duplicate pairs, to the predictions of the models for $\mu = 1$ and for $\mu > 1$. This requires knowledge of the run-length distribution, given a total number of deleted genes and remaining duplicate pairs. While this is easily calculated for the case $\mu = 1$, the the distribution for the opposing scenario $\mu > 1$ is not known.

In the first part of this paper, Sections 4.3 and 4.4, we analyze aspects of the deletion run-length distribution ψ when $\mu > 1$ for the deletion-length distribution γ, including some new and surprising analytical results, the clearest of which pertain to a continuous analog of the problem. We then show why it is difficult to describe in closed form or other easily computable format. In the second part, Section 4.6, we simulate the distribution and carry out a study of the discrimination problem for various values of μ, genome size N and θ the proportion of undeleted genes at time t.

4.3 The models

For modeling purposes, we consider a doubled genome made up, at the outset, of a pair of identical linear chromosomes each containing genes $g_1, \ldots, g_N$, where N is large enough so that we can neglect end effects. At each time $t = 1, 2, \ldots$ one such doubled gene g_i is chosen at random, and a value X is chosen from a geometric distribution γ with mean μ. If $X = a$, then $g_i, g_{i+1}, \ldots, g_{i+a-1}$ are deleted from one of the genomes — they become single-copy genes — unless some of these are already single-copy. In the latter case, we skip existing single-copy genes and proceed to convert the next double-copy genes we encounter until a total of a double-copy genes have been converted to single-copy. Note that this overlapping of deletion events never occurs if $\mu = 1$ since, in this case, by definition, exactly one double-copy gene is selected and deleted in each step. For simplicity, we assume all deletions take place from one and the same genome, though the more realistic and complicated case of deletion events occurring on one or the other chromosome, with probabilities ϕ and $1 - \phi$, has also been modeled [14].

The "skipping" procedure, introduced in [13], is a natural way to model the deletion process, since deletion of part of a chromosome and the subsequent rejoining of the chromosome directly before and directly after the deleted fragment means that this fragment is no longer "visible" to the deletion process. We have a record of the deleted genes, however, as a single copy of each gene is retained on the unaffected genome.

Overlapping deletion events and skipping result in the creation of runs of single-copy genes whose length is the sum of a number of geometric variables. The sum of r identical geometric variables produces a negative binomial distribution with parameter r, but the skipping process does not involve the sum of identical random

variables, since a deletion with a large value of a is more likely to overlap an existing single copy region than a deletion with small a. Thus, at any point of time $t > 0$, the distribution ψ_t of single-copy run lengths will tend to contain a higher frequency of runs of length 1, and of very long runs, than the negative binomial. On the other hand, the distribution of run lengths of the remaining double-copy genes is geometrically distributed with a probability distribution ρ_t , with a mean ν_t that decreases with t [13, 14].

4.4 Analysis of overlap probabilities

An attempt to determine ψ_t analytically starts with the calculation of how many deletion events have overlapped to form a run of single-copy genes at time t. In [13], we derived a formula to predict whether a deletion event would create a new run of single-copy genes, probability p_0; overlap exactly one existing run, thus extending it without changing the total number of runs, probability p_1; overlap two runs, producing one larger combined run in place of the two pre-existing ones, probability p_2; and so on. Other probabilities deal with the events that a run "touches" a pre-existing run without overlapping it. These probabilities all depend solely on γ and ρ_t. For example, we examine the case of p_0. The other probabilities are all formulated in analogous ways.

The proportion of terms in runs of length l is $l\rho_t(l)/\nu_t$, where $\nu_t = \sum_{l>0} l\rho_t(l)$. The probability p_0 that a deletion event falls within a run of double-copy genes without deleting the terms at either end is

$$\begin{aligned} p_0 &= \sum_{l>2} \frac{l\rho_t(l)}{\nu_t} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{\nu_t} \sum_{l>2} \rho_t(l) \sum_{j=2}^{l-1} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{\nu_t} \sum_{l>2} \rho_t(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a) \end{aligned} \tag{4.4.1}$$

where j indexes the starting position of the deletion within a run of length l, and a is the number of terms deleted in the event.

This formula requires quadratic computing time, but the p_i for higher i requires polynomial time of degree $i+2$. Here we exemplify with p_0 to show that these probabilities can in fact be reduced to closed form, so that computing time is a negligible constant. The formula in (4.4.1), when expanded, consists of a number of partial sums of the geometric distributions γ and ρ_t and means of these distributions, all of which are readily reduced to closed form, plus sums of terms of the form $[(1-\frac{1}{\mu})(1-\frac{1}{\nu_t})]^l$ and $l[(1-\frac{1}{\mu})(1-\frac{1}{\nu_t})]^l$, which themselves can be considered in terms of a geometric distribution with mean ζ, where

$$1 - \frac{1}{\zeta} = (1 - \frac{1}{\mu})(1 - \frac{1}{\nu_t}). \tag{4.4.2}$$

Then (4.4.1) reduces to

$$p_0 = \frac{(\nu_t - 1)^2}{(\mu + \nu_t - 1)\nu_t}. \tag{4.4.3}$$

For large ν_t, i.e. during the early stages of the process,

$$p_0 \approx \frac{\nu_t}{\mu + \nu_t}(1 - \frac{1}{\nu_t}). \tag{4.4.4}$$

Typically, $\mu > 1$ is a small number, often between 1 and 2, [9, 11], and ν_t of the order of 10^3 or 10^4. Thus p_0 is initially only slightly less than 1 but rapidly declines (see Fig. 3.4), as ν_t initially decreases by a factor of around $\frac{1}{2}$ at every step.

We proceed in an analogous way to derive closed forms for $p_1, p_2, \ldots$, but it is perhaps more instructive here to present the continuous version of the deletion process. Here the two identical chromosomes at time $t = 0$ are linear segments, long enough in comparison with the other parameters of the model so that end effects can be ignored. At each time $t = 1, 2, \ldots$, a random point g is chosen on the chromosome, and a value X is chosen from an exponential distribution

$$f(a) = \frac{1}{\mu} e^{-\frac{a}{\mu}}, \quad a \geq 0, \tag{4.4.5}$$

with mean μ. If $X = a$, then the segment $[g, g+a]$ is deleted from one of the genomes — $[g, g+a]$ becomes a single-copy region — unless part of it is already single-copy. In the latter case, we skip existing single-copy regions and proceed to convert the next double-copy region we encounter until a total of a double-copy regions have been converted to single-copy.

In analogy with the discrete model, the distribution of the lengths of remaining double-copy segment is exponentially distributed with a probability distribution $\rho(\ell)_t$, with a mean ν_t that decreases with t.

The proportion of undeleted lines accounted for by segments of length ldl is $\frac{l\rho(l)}{\nu_t} dl$, where $\nu_t = \int_0^\infty l\rho(\ell) dl$. Then the probability p_0 that a deletion event falls

completely within an undeleted segment is

$$p_0 = \int_{l=0}^{\infty} \frac{l\rho_t(l)}{\nu_t} \int_{x=0}^{l} \frac{1}{l} \int_{y=0}^{l-x} f(y)dy\,dx\,dl. \tag{4.4.6}$$

Carrying out the integrations, we find

$$p_0 = \frac{\nu_t}{\mu + \nu_t}, \tag{4.4.7}$$

reminiscent of the relation (4.4.4) in the discrete case with large ν_t.

The probability p_1 that a deletion event overlaps exactly one existing run of deletions is:

$$p_1 = \frac{1}{\nu_t} \int_{l=0}^{\infty} \int_{z=0}^{\infty} \rho_t(l)\rho_t(z) \int_{x=0}^{l} \int_{y=l-x}^{l-x+z} f(y)dy\,dx\,dz\,dl \tag{4.4.8}$$

$$= \frac{\nu_t}{\mu + \nu_t} \cdot \frac{\mu}{\mu + \nu_t}. \tag{4.4.9}$$

Using the same techniques, we can prove by induction that the probability a deletion event overlaps exactly q existing runs of deletions is:

$$p_q = \frac{\nu_t}{\mu + \nu_t} \left(\frac{\mu}{\mu + \nu_t} \right)^q. \tag{4.4.10}$$

Thus we have the surprisingly uncomplicated result that the number q of pre-existing runs of single-copy regions overlapped by a new deletion event is geometrically distributed on $q = 0, 1, \ldots$ with parameter $\mu/(\mu + \nu_t)$.

4.5 On the run-length distribution

Although having a closed form for p_q constitutes progress towards to the computation of the run-length distribution ψ_t, or eventually towards to some analytical results on it, how to find this distribution remains a difficult question. As mentioned in Section 4.3, long deletion events will be involved in more skipping than small ones. This is illustrated in Figure 4.1, where runs built from a small number of events tend to be composed of shorter deletions especially when μ is large. Had we just added independent samples from a geometric distribution, the curves in the figure would have been horizontal lines.

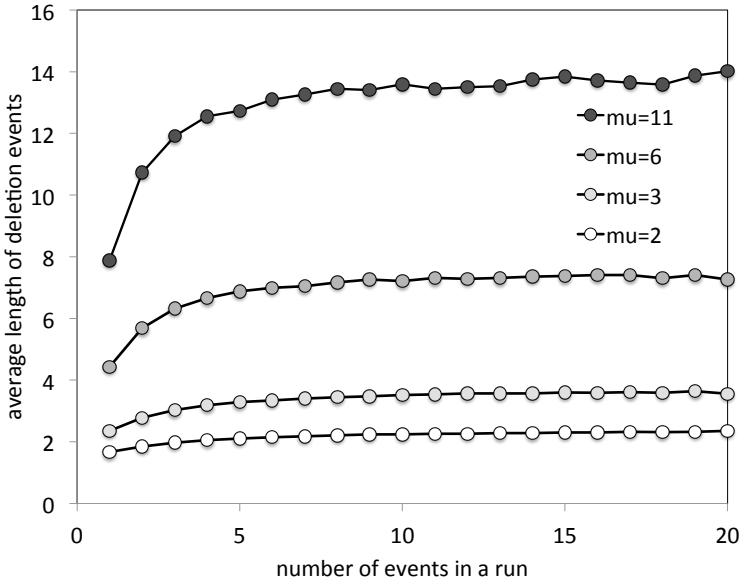

Figure 4.1: Simulation of the number of overlapping deletion events making up a single-copy region, when 70% of the genes are single copy. With a large number of events in a run, the individual events tend to have greater lengths. From [13].

How to account for the distorting effect of skipping on the run-length distribution will require additional insight and research. In the interim, we may use simulations to study the discrimination problem.

4.6 Simulations

We first simulated the fractionation process for all combinations of the following parameter values:

- $N = 100, 200, 300, 400, 500, 600, 700, 800, 900.$
- $\mu = 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4.$
- Proportion of the genes deleted, $1 - \theta = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

For each combination of the parameters μ, N, θ, we calculated the distribution of run lengths for both single-copy regions and double-copy regions. The simulation was repeated 1000 times and the frequencies of length $(1, 2, 3, \ldots)$ of runs of deleted genes were averaged over the 1000 trials to get a reasonably accurate estimate of the cumulative $F_{\mu,N,1-\theta}$. Similarly for the cumulative $G_{\mu,N,1-\theta}$ for runs of double-copy genes.

Once the cumulative distributions were established, we then carried out the actual discrimination study. For each value of μ and N, we sampled 1000 new individual trajectories of the deletion process until $1 - \theta = 90\%$ of the genes were deleted. For each value of $1 - \theta = 0.1, 0.2, \ldots, 0.9$, we created "bins" corresponding to the fifteen values of μ for which we had constructed cumulatives. Then for each sample S_i, at each $1 - \theta = 0.1, \ldots, 0.9$ we counted the frequency of runs of deleted genes of length $1, 2, \ldots$ and constructed a cumulative distribution. We calculated the Kolmogorov-Smirnov statistic $D^{(S_i)}_{\mu,N,1-\theta}$ between the sample cumulative and the distribution $F_{\mu,N,1-\theta}$ for each fifteen values of μ and assigned the sample to the bin corresponding to the minimal value of this statistic, which we called $\hat{\mu}$ for that sample.

Note that we use the Kolmogorov Smirnov statistic $D^{(S_i)}_{\mu,N,1-\theta}$ not for any testing purpose but as criterion for finding which μ, that is which $F_{\mu,N,1-\theta}$ is closest to the

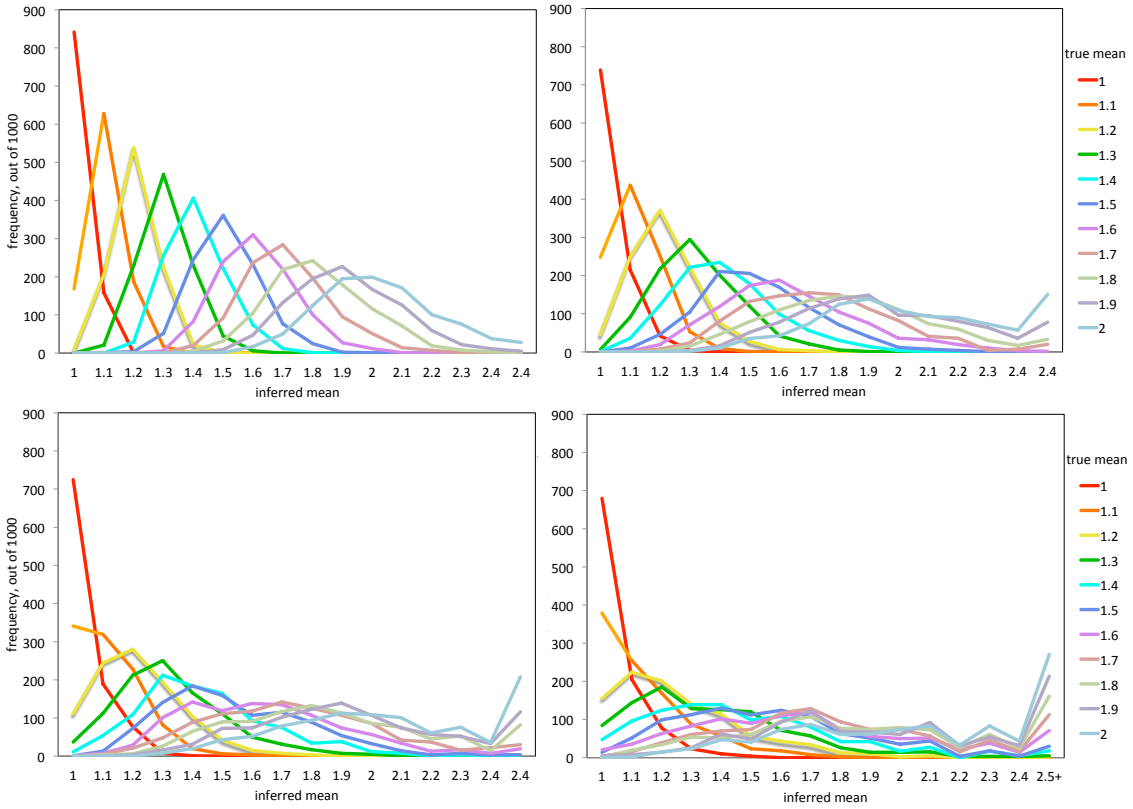

Figure 4.2: Frequency of $\hat{\mu}$, the value for which $D^{(S_i)}_{\mu,N,1-\theta}$ between the sample cumulative and the distribution $F_{\mu,N,1-\theta}$ is minimal. All data involve a proportion of $1-\theta = 0.20$ deleted genes. Top left: $N = 900$. Top right: $N = 300$. Bottom left: $N = 200$. Bottom right: $N = 100$.

sample cumulative. Other measures of goodness of fit could have been used, but Kolmogorov Smirnov statistic is convenient, because there are no intervals to set up as with χ^2 and no parameters to estimate.

Figure 4.2 shows the distributions of $\hat{\mu}$ for the 1000 samples $S_1, \ldots, S_{1000}$. The four panels are the results of $N = 900, 300, 200$ and 100. A separate distribution is drawn for each of the trial values of μ used to generate the samples. For $N = 900$ (top left), there is a clear pattern of the mode of the distribution to occur at the same value of μ that generated the data, though the distributions become more spread out for higher value of μ. The same pattern may be seen for $N = 300$ (top right),

though considerably degraded. This loss of accuracy of $\hat{\mu}$ continues through $N = 200$ (bottom left) and $N = 100$ (bottom right), where the modes for $\hat{\mu}$ when $\mu = 1.1$ are in the $\mu = 1.0$ bin.

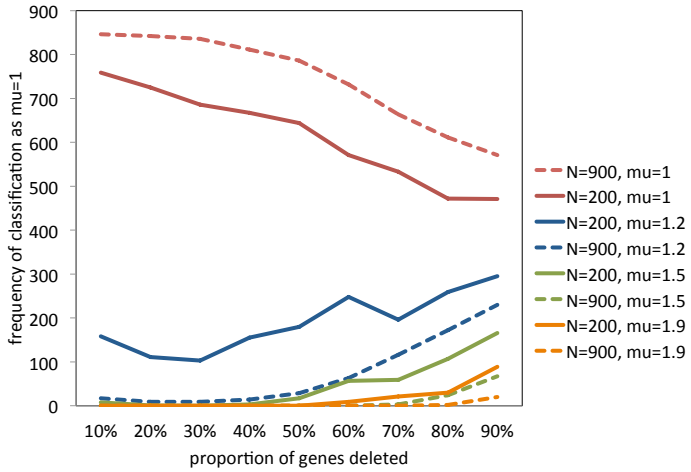

Figure 4.3: Frequency of $\hat{\mu} = 1$ as a function of $1 - \theta$, for various values of μ and N. The curves for μ =1, represent estimates of $1 - \alpha$, where α is the type 1 error when μ=1. The curves for μ >1 represent estimates of α, when μ is not 1.

With all four values of N in Figure 4.2, the most accurate inference is made for $\mu = 1$, the gene-by-gene model. This brings us back to the original problem of discriminating between the gene-by-gene "functional model" ($\mu = 1$ and the random excision "structural" model ($\mu > 1$). Figure 4.3 shows the frequency with which we estimate $\hat{\mu} = 1$, for various values of μ and $N = 200$ or 900, as a function of $1-\theta$, the proportion of genes deleted. The upper curve in the figure show that we can correctly identify the $\mu = 1$ model around 80% of the time; more for $N = 900$ and less for $N = 200$, as long as $1 - \theta < 50\%$. The lower curves show that incorrectly inferring $\hat{\mu} = 1$ occurs around 20% of the time when $\mu = 1.2$, but very rarely for $\mu = 1.9$ or even $\mu = 1.5$, until $1 - \theta$ begins to exceed 50%.

Up to now, we have examined only runs of single-copy genes. What of the runs of remaining double-copy genes? Figure 4.4 compares some of the results from Figure

4.3, but using the cumulative $G_{\mu,N,1-\theta}$ for runs of double-copy genes as well as $F_{\mu,N,1-\theta}$ for runs of single-copy genes. The main observation is that the double-copy approach systematically infers $\mu = 1$ with higher frequency for small values of $1-\theta$, whether or not this inference corresponds to the generating μ. It systematically infers $\mu = 1$ with lower frequency for large values of $1-\theta$, again whether or not this is correct.

These simulations establish ranges of values of N, μ and $1-\theta$ for which we can and cannot discriminate between the two models.

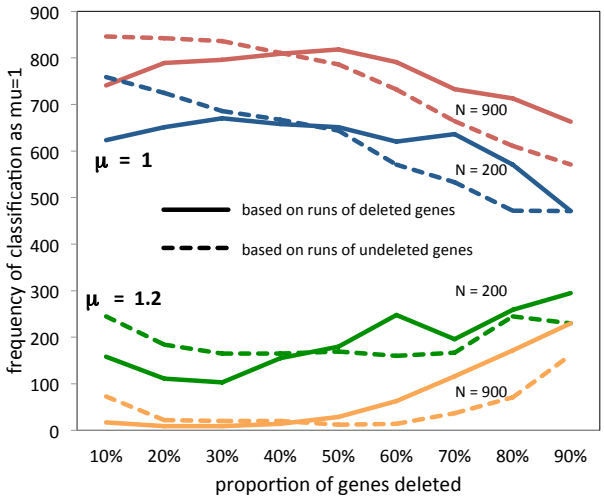

Figure 4.4: Frequency of $\hat{\mu} = 1$ as a function of $1-\theta$, for $\mu = 1$ and 1.2 and $N = 900$ and 200. Results based on runs of single-copy (deleted) genes contrasted with results from double-copy (undeleted) genes. For $\mu = 1$, the curves represent $1-\alpha$ and for $\mu = 1.2$ the curves represent α, where α is the size of the type 1 error.

4.7 Discussion

In this work, we have made some progress in deriving the run-length distribution ψ_t for single-copy regions, although this problem is still not completely resolved. From an analytical point of view, it is unexpected and interesting that, in the continuous version of the problem, the number of pre-existing runs overlapped by a deletion event

follows a geometric distribution.

The simulation study showed the much greater difficulty in distinguishing between the structural and functional models when the mean μ of the deletion size distribution is 1.1 rather than 1.9, when N is 100 rather than 900, and when the proportion of genes deleted is bigger than 50% rather than less than 40%. It also showed the differential powers of discrimination of double-copy runs versus single-copy runs at different levels of deletion.

Our simulation results are based on a "binning" strategy for determining $\hat{\mu}$ for the purposes of discrimination, rather than an asymmetrical testing approach comparing the hypotheses $\mu = 1$ and $\mu > 1$. This is justified by the lack of any biological significance, and high rates of error, in comparing $\mu = 1 + \epsilon$ and $\mu = 1$ for very small ϵ, as well as the global picture it offers of the degradation of discriminatory power as a function of μ, N and θ.

The calculation of $F_{\mu,N,1-\theta}$ and $G_{\mu,N,1-\theta}$ was based on a thousand samples for each combination of μ and N. Different stages of each sample provided the values of F and G for the different 1- θ. A sample size of one hundred would have been adequate. The average over the all samples can be considered a very accurate simulation of the true cumulative. Likewise, the Kolmogorov Smirnov-based binning of one thousand samples for each μ and N results in the great regularity of the curves in Fig. 4.2. Some statistical fluctuation is apparent in the secondarily derived graphs in Fig. 4.3, but the trends are unmistakable.

This work has for the first time enabled the systematic discrimination between the two models of duplicate deletion following WGD. Future research will continue on the analytical determination of ψ_t as well as extension to the "two-sided" deletion models proposed in [15]. It is the latter kind of model that will eventually be useful for analyzing data from real genomes.

Chapter 5

Conclusion

This thesis is the first systematic statistical accounts of the results of fractionation following whole genome duplication. We have drawn a sharp distinction between two biological account of fractionation data. Indeed, the extent to which duplicate, or *paralog*, deletion is a gene-by-gene *inactivation* process [8] targeting redundant copies at random points throughout the genome, or a consequence of largely random *excision*, elimination of excess DNA [11], is controversial and likely varies from one phylogenetic domain to another. However, the distinction between these two processes is not sharp: the inactivation effect may be produced not only by pseudogenization and various suppression and silencing mechanisms but also by the actual excision of a small but critical region of a gene or promoter. This could be captured by our continuous model in Chapter 4, but it is not clear whether there could be any real data we could apply this to, since they would have to be at a sub-genic level and involve various different kinds of structure, of which many are currently unknown or hard to locate. Conversely, the apparent excision of two or more adjacent genes may rather be due to any of a variety of genetic, epigenetic or functional interactions, rather than the deletion of a DNA fragment. Nevertheless, the determination of whether

deletion is a gene-by-gene process or the deletion of longer stretches of DNA is key to understanding the dynamics of genome evolution. this holds not only following WGD, but as part of the continual innovative expansion and simplifying shrinkage of genomes over time.

The skipping model we have adopted, where the lengths of two overlapping deletions are added in forming a single-copy region, is very simple, but the run-length probabilities ψ_t of the resulting regions are not easy to derive. I was finally able to find the probabilities for the number of pre-existing regions skipped by each new sample from the geometric distribution, but the resulting lengths are not distributed according to the negative binomial; Progress on this problem will require further insight. The exact derivation of π_t, the distribution of the number of deletion events contributing to a run of single-copy genes at time t is not known. Even if it were, this would not suffice to determine ψ_t.

Before extending these results to the two-sided model, we will have to reconsider the assignment to the non-truncated portions of the blocked samples from the geometric. This should be motivated by biological considerations. If the fractionation process proceeded linearly, one gene at a time until it was blocked after gene i, then the remaining tail probabilities should all be added to the segment ending in i. On the other hand, if deletions occurred according to the geometric, but all those greater than i turned out to be lethal, then all the probabilities of segments smaller or equal to i should increase, by normalizing by (1 – the tail probabilities). The simulation study showed the importance of simultaneously considering μ, N and θ for whether discrimination is possible between the two models, as well as the differential powers of discrimination of double-copy runs versus single-copy runs. Additional simulations for the two-sided model, at various bias levels (ϕ) will be needed before application to real data.

Appendix A

Appendix to Chapter 4 Detailed Calculations

A.1 Overlap Probabilities Formula—Discrete Case

$$
\begin{aligned}
p_0 &= \sum_{l>2} \frac{l\rho(l)}{\nu_t} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\
&= \frac{1}{\nu_t} \sum_{l>2} \rho(l) \sum_{j=2}^{l-1} \sum_{a=1}^{l-j} \gamma(a) \\
&= \frac{1}{\nu_t} \sum_{l>2} \rho(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a) \\
&= \frac{1}{\nu_t} \sum_{l>2} \rho(l) \left(\sum_{a=1}^{l-2} (l-1)\gamma(a) - \sum_{a=1}^{l-2} a\gamma(a) \right) \\
&= \frac{1}{\nu_t} \sum_{l>2}^{\infty} \rho(l) \left((l-1)[1 - \sum_{a=l-1}^{\infty} \gamma(a)] - [\mu - \sum_{a=l-1}^{\infty} a\gamma(a)] \right) \\
&= \frac{1}{\nu_t} \sum_{l>2}^{\infty} \rho(l) \left((l-1) - (l-1) \sum_{a=l-1}^{\infty} \gamma(a) - \mu + \sum_{a=l-1}^{\infty} a\gamma(a) \right) \\
&= \frac{1}{\nu_t} \left(\sum_{l>2}^{\infty} l\rho(l) - \sum_{l>2}^{\infty} \rho(l) - \sum_{l>2}^{\infty} \rho(l)(l-1) \sum_{a=l-1}^{\infty} \gamma(a) - \mu \sum_{l>2}^{\infty} \rho(l) + \sum_{l>2}^{\infty} \rho(l) \sum_{a=l-1}^{\infty} a\gamma(a) \right) \\
&= \frac{1}{\nu_t} \left(\sum_{l>2}^{\infty} l\rho(l) - \sum_{l>2}^{\infty} \rho(l) - A - \mu \sum_{l>2}^{\infty} \rho(l) + B \right) \qquad \text{(A.1.1)}
\end{aligned}
$$

where

$$\begin{aligned}
A &= \sum_{l>2}^{\infty} \rho(l)(l-1) \sum_{a=l-1}^{\infty} \gamma(a) \\
&= \sum_{l>2}^{\infty} \rho(l)(l-1) \sum_{b=1}^{\infty} \frac{1}{\mu}(1-\frac{1}{\mu})^{b+l-3} \\
&= \sum_{l>2}^{\infty} \rho(l)(l-1) \sum_{b=1}^{\infty} \frac{1}{\mu}(1-\frac{1}{\mu})^{b-1}(1-\frac{1}{\mu})^{l-2} \\
&= \sum_{l>2}^{\infty} \rho(l)(l-1)(1-\frac{1}{\mu})^{l-2} \\
&= \sum_{l>2}^{\infty} \frac{1}{\nu_t}(1-\frac{1}{\nu_t})^{l-1}(l-1)(1-\frac{1}{\mu})^{l-1}(1-\frac{1}{\mu})^{-1} \\
&= \frac{1}{\nu_t}(1-\frac{1}{\mu})^{-1} \sum_{l>2}^{\infty}(l-1)\left((1-\frac{1}{\nu_t})(1-\frac{1}{\mu})\right)^{l-1}
\end{aligned} \tag{A.1.2}$$

Letting $x = (1-\frac{1}{\nu_t})(1-\frac{1}{\mu})$ then

$$\begin{aligned}
A &= \frac{1}{\nu_t}(1-\frac{1}{\mu})^{-1} \sum_{l>2}^{\infty}(l-1)\left((1-\frac{1}{\nu_t})(1-\frac{1}{\mu})\right)^{l-1} \\
&= \frac{1}{\nu_t}(1-\frac{1}{\mu})^{-1} \sum_{l>2}^{\infty}(l-1)(x)^{l-1} \\
&= \frac{1}{\nu_t}(1-\frac{1}{\mu})^{-1}\left(\frac{(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})}{[1-(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})]^2} - (1-\frac{1}{\nu_t})(1-\frac{1}{\mu})\right) \\
&= \frac{\mu^2(\nu_t-1)}{(\nu_t+\mu-1)^2} - \frac{\nu_t-1}{\nu_t^2}
\end{aligned} \tag{A.1.3}$$

where

$$
\begin{aligned}
B &= \sum_{l>2}^{\infty}\rho(l)\sum_{a=l-1}^{\infty}a\gamma(a)\\
&= \sum_{l>2}^{\infty}\rho(l)\sum_{a=l-1}^{\infty}a\frac{1}{\mu}(1-\frac{1}{\mu})^{a-1}\\
&= \sum_{l>2}^{\infty}\rho(l)\sum_{b=1}^{\infty}(b+l-2)\frac{1}{\mu}(1-\frac{1}{\mu})^{b+l-3}\\
&= \sum_{l>2}^{\infty}\rho(l)\sum_{b=1}^{\infty}(b+l-2)\frac{1}{\mu}(1-\frac{1}{\mu})^{b-1}(1-\frac{1}{\mu})^{l-2}\\
&= \sum_{l>2}^{\infty}\rho(l)(1-\frac{1}{\mu})^{l-2}\left(\sum_{b=1}^{\infty}b\frac{1}{\mu}(1-\frac{1}{\mu})^{b-1}+\sum_{b=1}^{\infty}(l-2)\frac{1}{\mu}(1-\frac{1}{\mu})^{b-1}\right)\\
&= \sum_{l>2}^{\infty}\rho(l)(1-\frac{1}{\mu})^{l-2}[\mu+(l-2)]\\
&= \sum_{l>2}^{\infty}\mu\rho(l)(1-\frac{1}{\mu})^{l-2}+\sum_{l>2}^{\infty}(l-2)\rho(l)(1-\frac{1}{\mu})^{l-2}\\
&= \sum_{l>2}^{\infty}\mu\frac{1}{\nu_t}(1-\frac{1}{\nu_t})^{l-1}(1-\frac{1}{\mu})^{l-1}(1-\frac{1}{\mu})^{-1}+\sum_{l>2}^{\infty}(l-2)\frac{1}{\nu_t}(1-\frac{1}{\nu_t})^{l-1}(1-\frac{1}{\mu})^{l-1}(1-\frac{1}{\mu})^{-1}\\
&= \frac{\mu}{\nu_t}(1-\frac{1}{\mu})^{-1}\sum_{l>2}^{\infty}\left((1-\frac{1}{\nu_t})(1-\frac{1}{\mu})\right)^{l-1}+(1-\frac{1}{\mu})^{-1}\frac{1}{\nu_t}\frac{(1-\frac{1}{\nu_t})^2(1-\frac{1}{\mu})^2}{[1-(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})]^2}\\
&= \frac{\mu}{\nu_t}(1-\frac{1}{\mu})^{-1}\left(\frac{1}{1-(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})}-1-(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})\right)\\
&+ (1-\frac{1}{\mu})^{-1}\frac{1}{\nu_t}\frac{(1-\frac{1}{\nu_t})^2(1-\frac{1}{\mu})^2}{[1-(1-\frac{1}{\nu_t})(1-\frac{1}{\mu})]^2}\\
&= \frac{(\nu_t-1)(2\mu^2\nu_t+\mu^3-2\mu^2-\mu\nu_t+\mu)}{(\nu_t+\mu-1)^2\nu_t}-\frac{\mu(\nu_t-1)}{\nu_t^2} \qquad \text{(A.1.4)}
\end{aligned}
$$

Finally,

$$
\begin{aligned}
p_0 &= \frac{1}{\nu_t}\left(\sum_{l>2}^{\infty} l\rho(l) - \sum_{l>2}^{\infty}\rho(l) - A - \mu\sum_{l>2}^{\infty}\rho(l) + B\right) \\
&= \frac{1}{\nu_t}\left([\nu_t - \frac{1}{\nu_t} - 2\frac{1}{\nu_t}(1-\frac{1}{\nu_t})] - [1 - \frac{1}{\nu_t} - \frac{1}{\nu_t}(1-\frac{1}{\nu_t})] - \frac{\mu^2(\nu_t-1)}{(\nu_t+\mu-1)^2} + \frac{\nu_t-1}{\nu_t^2}\right) \\
&- \frac{1}{\nu_t}\left(\mu[1 - \frac{1}{\nu_t} - \frac{1}{\nu_t}(1-\frac{1}{\nu_t})] - \frac{(\nu_t-1)(2\mu^2\nu_t+\mu^3-2\mu^2-\mu\nu_t+\mu)}{(\nu_t+\mu-1)^2\nu_t} - \frac{\mu(\nu_t-1)}{\nu_t^2}\right) \\
&= \frac{(\nu_t-1)^2}{(\mu+\nu_t-1)\nu_t} \qquad \text{(A.1.5)}
\end{aligned}
$$

A.2 Overlap Probabilities Formula—Continuous Case

$$
\begin{aligned}
p_0 &= \int_0^{\infty}\frac{l\rho(l)}{\nu_t}\int_{j=0}^{l}\frac{1}{l}\int_{a=0}^{l-j}\gamma(a)dadjdl \\
&= \frac{1}{E_\rho}\int_0^{\infty}\rho(l)\int_{j=0}^{l}\int_{a=0}^{l-j}\frac{1}{\mu}e^{-\frac{a}{\mu}}dadjdl \\
&= \frac{1}{E_\rho}\int_0^{\infty}\rho(l)\int_{j=0}^{l}(-e^{-\frac{a}{\mu}}|_{a=0}^{l-j})djdl \\
&= \frac{1}{E_\rho}\int_0^{\infty}\rho(l)\int_{j=0}^{l}(1-e^{\frac{j-l}{\mu}})djdl \\
&= \frac{1}{E_\rho}\left(\int_0^{\infty}\rho(l)ldl - \int_0^{\infty}\rho(l)(\mu e^{\frac{j-l}{\mu}}|_{j=0}^{l})dl\right) \\
&= \frac{1}{E_\rho}\left(\int_0^{\infty}\rho(l)ldl - \int_0^{\infty}\rho(l)(\mu-\mu e^{\frac{-l}{\mu}})dl\right) \\
&= \frac{1}{E_\rho}\left(\int_0^{\infty}\frac{1}{\nu_t}e^{-\frac{l}{\nu_t}}ldl - \mu\int_0^{\infty}\frac{1}{\nu_t}e^{-\frac{l}{\mu'}}dl + \mu\int_0^{\infty}\frac{1}{\mu'}e^{-\frac{l}{\nu_t}}e^{\frac{-l}{\mu}}dl\right) \\
&= \frac{1}{E_\rho}(\nu_t - \mu + \frac{\mu^2}{\mu+\nu_t}) \\
&= \frac{1}{E_\rho}\frac{\nu_t^2}{\mu+\nu_t} \\
&= \frac{\nu_t}{\mu+\nu_t} \qquad \text{(A.2.1)}
\end{aligned}
$$

$$\begin{aligned}
p_1 &= \frac{1}{\nu_t}\int_0^\infty\int_0^\infty \rho(l)\rho(k)\int_{j=0}^{l}\int_{a=l-j}^{l-j+k}\gamma(a)dadjdkdl \\
&= \frac{1}{\nu_t}\int_0^\infty\int_0^\infty \rho(l)\rho(k)\int_{j=0}^{l}\int_{a=l-j}^{l-j+k}\frac{1}{\mu}e^{-\frac{a}{\mu}}dadjdkdl \\
&= \frac{1}{\nu_t}\int_0^\infty \left(\rho(l)(\mu-\mu e^{-\frac{l}{\mu}}-\frac{\mu^2}{\mu+\mu'}+\frac{\mu^2}{\mu+\mu'}e^{-\frac{l}{\mu}}\right)dl \\
&= \frac{1}{\nu_t}(\mu-\frac{2\mu^2}{\mu+\nu_t}+\frac{\mu^3}{(\mu+\nu_t)^2}) \\
&= \frac{\mu\nu_t}{(\mu+\nu_t)^2} \\
&= \frac{\nu_t}{\mu+\nu_t}\cdot\frac{\mu}{\mu+\nu_t} \qquad \text{(A.2.2)}
\end{aligned}$$

$$p_q = \frac{\nu_t}{\mu+\nu_t}\cdot(\frac{\mu}{\mu+\nu_t})^q.$$

We are going to use induction to prove this formula; that is, $p_q = p_A \cdot p^q$.

We already checked $q = 0$ and $q = 1$, and it is true. We suppose that the result $q = k$. That is,

$$\begin{aligned}
p_k &= p_A\cdot p^k \\
&= \frac{\mu'}{\mu+\mu'}\cdot(\frac{\mu}{\mu+\mu'})^k \\
&= \frac{\mu}{\mu'}\left((\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k-1}-2(\frac{\mu}{\mu+\mu'})^k+\frac{\mu}{\mu+\mu'})^{k+1}\right) \qquad \text{(A.2.3)}
\end{aligned}$$

Now if $q = k + 1$, we have that

$$\begin{aligned}
p_{k+1} &= \frac{\mu}{E_\rho}\int_0^\infty \rho(l)\left[\left((\frac{\mu}{\mu+\mu'})^k - (\frac{\mu}{\mu+\mu'})^k e^{-\frac{l}{\mu}}\right) - \left((\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k+1} - (\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k+1}e^{-\frac{l}{\mu}}\right)\right] dl \\
&= \frac{\mu}{E_\rho}\left((\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^k - 2(\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k+1} + (\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k+2}\right) \\
&= \frac{\mu}{E_\rho}\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'}\left((\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k-1} - 2(\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k} + (\frac{1}{\mu'}\frac{\mu\mu'}{\mu+\mu'})^{k+1}\right) \\
&= \frac{\mu}{\mu+\mu'}p_k \\
&= \frac{\mu}{\mu+\mu'}p_A \cdot p^k \\
&= p_A \cdot p^{k+1} \qquad \text{(A.2.4)}
\end{aligned}$$

finishing the proof.

Bibliography

[1] Langham Richard, Walsh Justine, Dunn Molly, Ko Cynthia, Goff Stephen, Freeling Michael: Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 2004, 166:935-945.

[2] El-Mabrouk N, David Sankoff: The reconstruction of doubled genomes. SIAM Journal on Computing 2003,32:754-792.

[3] Robert Warren, David Sankoff: Genome aliquoting with double cut and join. BMC Bioinformatics 10, S2 (2009).

[4] Chunfang Zheng, Qian Zhu, David Sankoff: Descendants of whole genome duplication within gene order phylogeny, Journal of Computational Biology 15, 947-964 (2008).

[5] David Sankoff, Chunfang Zheng, Qian Zhu: The collapse of gene complement following whole genome duplication, BMC Genomics 11, 313 (2010)

[6] Chunfang Zheng, David Sankoff: Fractionation, rearrangement and subgenome dominance Bioinformatics 28, i402-i408 (2012)

[7] Katharina Jahn, Chunfang Zheng, Jakub Kovac, David Sankoff: A consolidation algorithm for genomes fractionated after higher order polyploidization. BMC Bioinformatics 13, S19:S8 (2012)

[8] Byrne Kevin, Wolfe Kenneth: The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Research 2005, 15:1456-1461.

[9] Byrnes JK, Morris GP, Li WH: Reorganization of Adjacent Gene Relationships in Yeast Genomes by Whole-Genome Duplication and Gene Deletion. Molecular Biology and Evolution June 2006, 23(6):1136-1143.

[10] Thomas B, Pedersen B, Freeling M: Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Research 2006, 16:934-946.

[11] Van Hoek Milan, Hogeweg Paulien: The role of mutational dynamics in genome shrinkage. Molecular Biology and Evolution 2007, 24:2485-2494.

[12] Edger Patrick, Pires J: Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 2009, 17:699-717.

[13] Baoyong Wang, Chunfang Zheng, David Sankoff: Fractionation statistics. BMC Bioinformatics 2011, 12(Suppl 9):S5.

[14] David Sankoff, Chunfang Zheng, Baoyong Wang: 2012. A model for biased fractionation after whole genome duplication. BMC Genomics 13:S1, S8 (2012)

[15] Baoyong Wang, Chunfang Zheng, Carlos Fernando Buen Abad Najar, David Sankoff: Discriminating between structural and functional mechanisms for duplicate gene loss following whole genome doubling. Submitted.