The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 14):412
DOI 10.1186/s12859-016-1265-5

BMC Bioinformatics

RESEARCH Open Access

A continuous analog of run length distributions reflecting accumulated fractionation events

Zhe Yu and David Sankoff*

From 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop, Montreal, Canada.11-14 October 2016

Abstract

Background: We propose a new, continuous model of the fractionation process (duplicate gene deletion after polyploidization) on the real line. The aim is to infer how much DNA is deleted at a time, based on segment lengths for alternating deleted (invisible) and undeleted (visible) regions.

Results: After deriving a number of analytical results for "one-sided" fractionation, we undertake a series of simulations that help us identify the distribution of segment lengths as a gamma with shape and rate parameters evolving over time. This leads to an inference procedure based on observed length distributions for visible and invisible segments.

Conclusions: We suggest extensions of this mathematical and simulation work to biologically realistic discrete models, including two-sided fractionation.

Keywords: Genomics, Whole genome duplication, Analysis of runs, Probability modeling, Duplicate gene deletion

Background

In the course of evolution, new genomes occasionally arise by duplication or triplication of an existing genome, so that there are two or three identical copies of each maternal and each paternal chromosome. After a (usually) transient period of polyploidy marked by unusual patterns of meiosis where more than just one maternal and paternal chromosome are aligned and recombine, processes of sequence divergence and chromosome rearrangement lead to more familiar diploid patterns. At the same time a process of *fractionation* eliminates some or most of the duplicate genes, some from each chromosomal copy, but in the simplest model, never all members of a duplicate pair or triple - for reasons of viability. Fractionation processes have been surveyed across evolutionarily diverse types of eukaryote organisms [1].

Since one copy of a duplicate pair of genes must be retained, we can identify not only the chromosomal regions that have been retained – by simple observation of the genome – but also each region that is now invisible – by reference to the duplicate chromosome that has necessarily retained a copy of this region. Thus, the data on which inferences about the deletion process can be made consist of alternating segments of deleted and undeleted genome of varying lengths.

Among the important questions about the nature of the deletion process, we can ask whether deletion proceeds one gene at a time or by larger chromosomal fragments. In this paper, we model the process as the deletion of segments from the real line, with a biologically realistic treatment afforded to overlapping deletions. Previous work focused on the difficult question of how many overlapping deletion events are responsible for each contiguous deleted region [2–4], but was not able to account analytically for the dynamics of the process.

*Correspondence: sankoff@uottawa.ca
Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, K1N 6N5 Ottawa, Ontario, Canada

In the present paper we attack and solve the inference problem of the size, form and spacing of deletion events, allowing for a number of sweeps over the genome as a way of accounting for overlapping deletions. We carry this out in a continuous analog of the original discrete gene-order context, and address the "one-sided" version of the problem, where all deletions occur on one of the duplicate chromosomes.

There has been a certain amount of work on the quantification of the fractionation process, starting in 2006 with [5], which claimed deletions involved one gene at a time, and [6], which treated the number of genes deleted in a single event as a random variable with mean greater than 1. Other work of this kind includes [1] and [7]. However, the modelling of fractionation where the whole genome evolves as a stochastic process began with [2]. The previously unstudied phenomenon taken into account in that work was the overlap of deletion events, something that assumes much importance soon after the fractionation process commences. Overlap must be handled differently if all deletions occur from one copy of the genome or in either copy. To isolate the most important aspect of overlap, [2] gave analytical results for the case where deletions all occurred on one copy ("one-sided" model). Then [3] extended this to the more realistic case where deletion could occur at different rates, or the same rate, from either copy of the genome ("two-sided" model). This analysis was more difficult and could not be taken as far as with the one-sided model.

For the one-sided model, a closed form solution of how many deletion events contribute to a deleted region after a single event (i.e., at a single step in the fractionation process) was obtained in [4].

Methods

The proposed model

We model the fractionation process in terms of a number of successive sweeps of a point process with parameter ν on the positive reals, i.e., $\nu \in \mathbf{R}^+$, representing one copy of the genome. At the origin, we say that all points of this genome are "visible". A deletion event, rendering a segment of exponentially (mean μ) distributed length "invisible", occurs at each point determined by the point process. The second copy of the genome remains undisturbed throughout and retains a 1-to-1, length preserving, correspondence with the fractionating copy, without regard to any disruption caused by invisibility. In applications, the acceptance of the one-gene-at-a time theory of deletion depends on whether μ is below or above a certain absolute value, but the present work is part of the mathematical preliminaries to the practical questions. The eventual goal of this work is to determine the relative size of the "spacing" parameter ν and the deletion length parameter μ. The model innovation here is to introduce the parameter ν in the place of a rate parameter in previous work, which was awkward to work with.

During the first sweep, illustrated at the top of Fig. 1 at time (or step) $t = 1$, the first *deletion point* x_1 is determined by sampling from the exponential distribution

$$\rho(x) = \frac{1}{\nu} e^{-\frac{x}{\nu}},\ x \geq 0, \tag{1}$$

with mean ν. Then a deletion length a_1 is chosen from another exponential distribution

$$\gamma(a) = \frac{1}{\mu} e^{-\frac{a}{\mu}},\ a \geq 0, \tag{2}$$

with mean μ. Normally, $\nu \gg \mu$, but this is not necessary to the analysis. The segment $[x_1, x_1 + a_1)$ is "deleted", or is designated as invisible. The next deletion point x_2 is chosen by sampling x'_2 from the first exponential distribution (mean ν), so that $x_2 = x'_2 + x_1 + a_1$. Then the length a_2 of the second deleted segment is determined by sampling from γ again. The process continues in this way to find $x_3, a_3, \ldots$ Concatenating only those segments that are still visible, we see that $x_1, x_2, \ldots$ are points determined by a point process with parameter ν. Associated with each of these points x is an "event counter" $C(x)$. Initially, each $C(x) = 1$. We define a function $\pi_t(i), i = 1, \ldots$ measuring the proportion of event counters registering i events at time $t \geq 1$. Thus $\pi_1(1) = 1$ and $\pi_1(j) = 0$, for all $j > 1$.

At times $t = 1, 2, \ldots$, the second, third, ... sweeps begin, all independent of the first sweep and each other, and each applied to the concatenated visible segments only. We sample $x_1^{(t)}$ and $a_1^{(t)}$ in the same way as x_1 and a_1 according to ρ and γ, respectively, to determine a deletion interval $[x_1^{(t)}, x_1^{(t)} + a_1^{(t)})$.

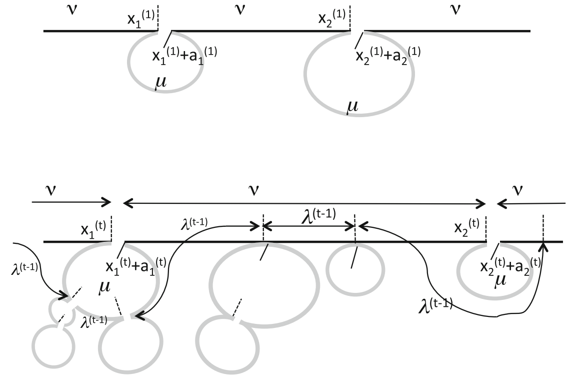

Fig. 1 Processes pertinent to first sweep and *t*-th sweep. *Solid horizontal bars* represent the visible regions of the genome. *Grey curves* represent invisible regions. *Dashed markers* represent deletion points, *solid markers* represent end of deletion segments. ν and μ are the means of the deletion point spacing and deletion segment length variables, while $\lambda^{(t-1)}$ is the mean space (=λ_{t-1} in the text) between visible deletion points after the $t-1$-st sweep

If the interval $[x_1^{(t)}, x_1^{(t)} + a_1^{(t)})$ contains no previously defined deletion point, a new event counter at $C(x_1^{(t)})$ is set at 1. If $[x_1^{(t)}, x_1^{(t)} + a_1^{(t)})$ already contains $j > 1$ deletion points $z_1, \ldots, z_j$, the event counter at $C(x_1^{(t)})$ is set at $1 + \sum_{i=1}^{j} C(z_i)$. The j deletion points $z_1, \ldots, z_j$ become invisible, along with the rest of the segment $[x_1^{(t)}, x_1^{(t)} + a_1^{(t)})$ that contains them.

We find the next deletion point by sampling $x_2^{(t)'}$ from ρ, and setting $x_2(t) = x_1^{(t)} + a_1^{(t)} + x_2^{(t)'}$. We continue the t sweep, adding visible deletion points and making others invisible. Some deletion points from the earlier sweep will remain unchanged, i.e. are still visible. The $x_i^{(t)}$ by themselves define a point process with parameter ν on the concatenated visible segments. But the $x_i^{(t)}$ and the additional deletion points remaining from the earlier sweep define a process with mean λ_t, a parameter that decreases with t, as the undeleted segments are interrupted by more and more deletions. This parameter is important as it is directly inferable from the observed genome at time t.

More important, it is clear, that at each sweep, more and more of the genome becomes invisible. Since each concatenation of visible segments still extends to the positive reals, we cannot observe directly how much the genome has been reduced in absolute terms. But thanks to the length-preserving isomorphism between the second copy of the genome and the fractionating one, for any large finite interval we can observe the proportion of the genome that is left by time t and we can predict that it is approximately $(1 - \frac{\mu}{\nu+\mu})^t$.

We will calculate λ, the number of deletion points in $[x_i, x_{i+1})$, the distribution $p(j), j = 1, \ldots$ of the number j of pre-existing deletion points in intervals deleted during each sweep, and discuss how to calculate $\pi_1(j), j \geq 1$, the proportion of event counters with $C = j$.

Results

The length of undeleted segments λ

After the first sweep, x_i is the only deletion point in $[x_i^{(1)}, a_i^{(1)})$ and the only deletion point in the visible $[x_i^{(1)}, x_{i+1}^{(1)})$, so that $\lambda_1 = \nu$. During the second sweep, the number of these first-sweep deletion points that the visible $[x_i^{(2)}, x_{i+1}^{(2)})$ contains is Poisson distributed with mean $\frac{\nu}{\nu+\mu}$, while the remaining first-sweep deletion points that the invisible $[x_i^{(2)}, a_i^{(2)})$ contains are Poisson distributed with mean $\frac{\mu}{\nu+\mu}$. (These are approximations, since the true means are $\frac{x_{i+1}^{(2)} - x_i^{(2)}}{x_{i+1}^{(2)} + a_i^{(2)} - 2x_i^{(2)}}$ and $\frac{a_i^{(2)} - x_i^{(2)}}{x_{i+1}^{(2)} + a_i^{(2)} - 2x_i^{(2)}}$, respectively.) In addition the visible segment contains one new deletion

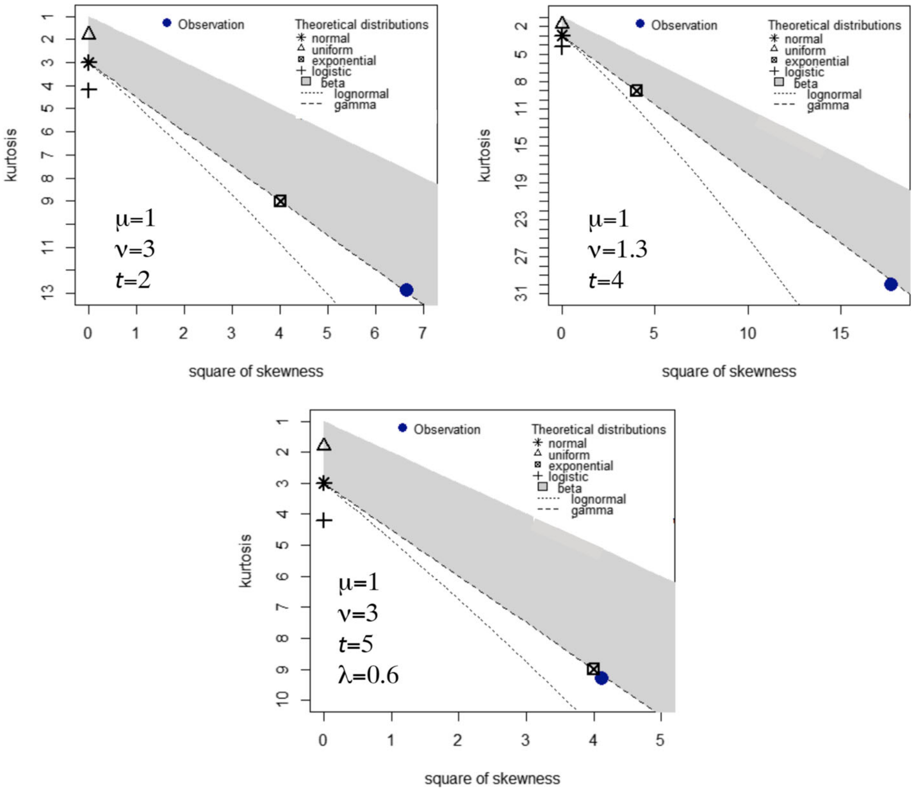

Fig. 2 Cullen-Frey diagrams for length distributions of invisible (*top*) and visible (*bottom*) segments

Table 1 Simulated values of shape and rate when $\frac{\mu}{\nu} = \frac{1}{3}$, for a range of values of μ, and $t = 2$

μ	ν	shape α	rate β	$1/\beta$
1	3	0.8994863	0.7801536	1.281798866
2	6	0.8713054	0.3645944	2.742773888
3	9	0.8943245	0.2557653	3.909834524
4	12	0.8551860	0.1863732	5.365578313
5	15	0.8479933	0.1504409	6.647128540
6	18	0.8673458	0.1250687	7.995605615
7	21	0.8793444	0.1044622	9.572840702
8	24	0.91907486	0.09607099	10.40896945
9	27	0.91503151	0.08817842	11.34064321
10	30	0.82931206	0.07483308	13.36307419

point, created during the second sweep itself. We can then predict λ_2 to be roughly

$$\hat{\lambda}_2 = \frac{\nu}{1 + \frac{\nu}{\nu+\mu}}. \tag{3}$$

Suppose λ_{t-1} is the parameter of the point process that generates the deletion points visible after sweep $t - 1$. Then, in the sweep at time t, the number of deletion points that the invisible $[x_i^{(2)}, a_i^{(2)})$ will contain is Poisson distributed with mean $\frac{\mu}{\lambda_{t-1}}$. The number of deletion points in the visible $[x_i, x_{i+1})$, not including x_i, is Poisson distributed with mean $\frac{\nu}{\lambda_{t-1}}$. In addition, the visible segment contains one new deletion point, created during the t-th sweep itself. λ_t can thus be predicted to be approximately

$$\hat{\lambda}_t = \frac{\nu}{1 + \frac{\nu}{\hat{\lambda}_{t-1}}}. \tag{4}$$

Since $\hat{\lambda}_1 = \nu$,

$$\hat{\lambda}_t = \frac{\nu}{t}. \tag{5}$$

The treatment of overlapping deletions

The discussions in this section and the next do not depend on t, so let Λ be the exponential distribution with mean λ. From [4], the probability p_0 that a deletion event contains no extant deletion points is

$$p_0 = \int_{l=0}^{\infty} \frac{l\Lambda(l)}{\lambda} \int_{x=0}^{l} \frac{1}{l} \int_{y=0}^{l-x} \gamma(y)dy\,dx\,dl. \tag{6}$$

Carrying out the integrations, we find

$$p_0 = \frac{\lambda}{\mu + \lambda}. \tag{7}$$

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

$$p_1 = \frac{1}{\lambda} \int_{l=0}^{\infty} \int_{z=0}^{\infty} \Lambda(l)\Lambda(z) \int_{x=0}^{l} \int_{y=l-x}^{l-x+z} \gamma(y)dy\,dx\,dz\,dl \tag{8}$$

$$= \frac{\lambda}{\mu + \lambda} \cdot \frac{\mu}{\mu + \lambda}. \tag{9}$$

It can be proved by induction that the probability a deletion event overlaps exactly q existing runs of deletions is:

$$p_q = \frac{\lambda}{\mu + \lambda} \left(\frac{\mu}{\mu + \lambda} \right)^q. \tag{10}$$

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 $\mu/(\mu + \lambda)$.

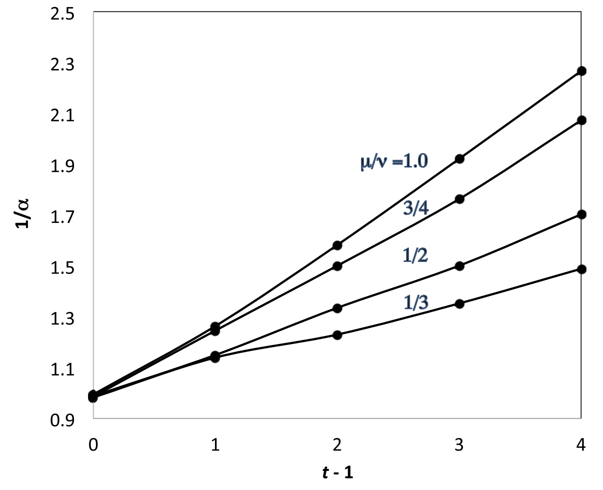

Fig. 3 Linear relation between $1/\alpha$ and $t - 1$ for fixed $\frac{\mu}{\nu}$

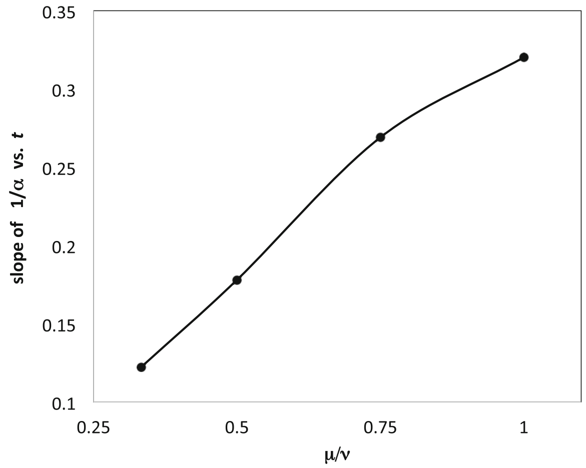

Fig. 4 Relation between slope of $1/\alpha$ as a function of t, and $\frac{\mu}{\nu}$

The distribution of event counts π

The event count $C(x)$ at a visible deletion point x tells us how many deletion events have occurred to make up the invisible segment adjacent to x. In contrast to the undeleted segments, where we know that no events occurred, observing that a segment has been deleted does not tell us $C(x)$. Some work has focused on the distribution $\pi(i)$ of the probabilities that a deletion point x has $C(x) = i$, and we are able to calculate how π changes with each sweep. Then we can update π_t by a linear

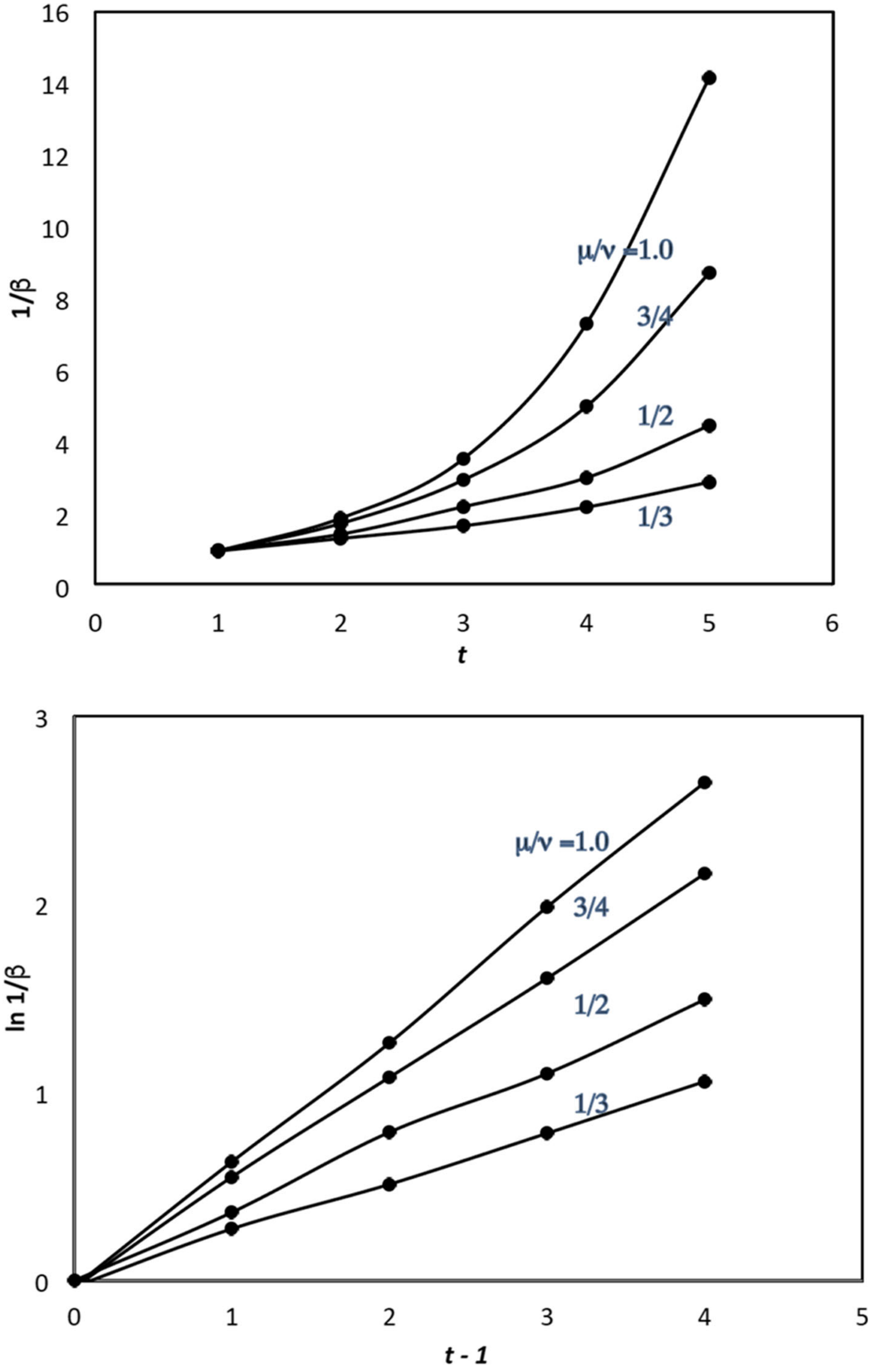

Fig. 5 Relation between 1/rate ($1/\beta$) as a function of t for fixed $\frac{\mu}{\nu}$

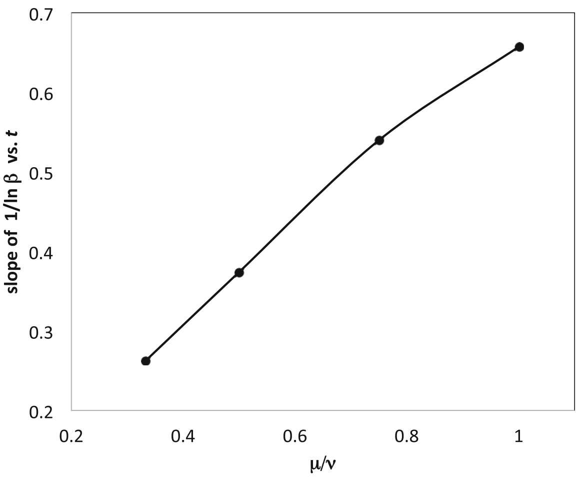

Fig. 6 Relation between slope of ln $1/\beta$ as a function of t, and $\frac{\mu}{\nu}$

combination of the distribution of changes due to the deletion and the existing π_{t-1}. Let $\Delta(i)$ represent the change in π_i at any sweep t. This can be calculated from Eq. (10) and the net effect that a deletion overlapping q existing runs has on the various π. Without giving details here,

$$\begin{aligned}\Delta(1) &= p_0 - p_1\,[\pi(1)] - 2p_2\,[\pi(1)] - 3p_3\,[\pi(1)]\\ &\quad - 4p_4\,[\pi(1)] - \ldots \qquad (11)\\ \Delta(2) &= p_1\,[\pi(1)] - p_1\,[\pi(2)] - 2p_2\,[\pi(2)] - 3p_3\,[\pi(2)]\\ &\quad - 4p_4\,[\pi(2)] - \ldots \qquad (12)\\ \Delta(3) &= p_1\,[\pi(2)] + p_2\left[(\pi(1))^2\right] - p_1\,[\pi(3)]\\ &\quad - 2p_2\,[\pi(3)] - 3p_3\,[\pi(3)] - 4p_4\,[\pi(3)] - \ldots \qquad (13)\\ \Delta(4) &= p_1\,[\pi(3)] + 2p_2\,[\pi(1)\pi(2)] + p_3\left[(\pi(1))^3\right]\\ &\quad - p_1\,[\pi(4)] - 2p_2\,[\pi(4)] - 3p_3\,[\pi(4)] \qquad (14)\\ &\quad - 4p_4\,[\pi(4)] - \ldots\\ \Delta(5) &= p_1\,[\pi(4)] + p_2\left[2\pi(1)\pi(3) + (\pi(2))^2\right]\\ &\quad + 3p_3\left[(\pi(1))^2\,\pi(2)\right] + p_4\left[(\pi(1))^4\right] \qquad (15)\\ &\quad - p_1\,[\pi(5)] - 2p_2\,[\pi(5)] - 3p_3\,[\pi(5)]\\ &\quad - 4p_4\,[\pi(5)] - 5p_5\,[\pi(5)] - \ldots \qquad (16)\\ &\ldots\end{aligned}$$

Unfortunately, even knowing the dynamics of C does not help us with the inference problem, since the number of events associated with an invisible segment, is not directly associated with the total length of the segment. It is known that the overlapping gamma variables making up each segment are related in a complex way, and cannot simply be treated as the sum of gammas drawn a single population.

This leads us to the approach in the next two sections, where simulations strongly suggest the functional form of the distribution of invisible segment lengths, including shape and rate parameters that can be observed, leading to inference of the simulation parameters based on the observations.

Simulation

Our simulation experiments were based on initial visible segments of length 10,000, which is very long in comparison to the deletion lengths with $\mu \leq 10$. In other words, we do not risk artificial effects, like a disappearing genome, after a few sweeps, $t \leq 10$. Moreover, after each sweep, if the total undeleted length $= L$, we add, to the end of the remaining visible portion, segments where the lengths of the visible portions total $10,000 - L$, copied from a replicate trial. The program, written in Java, was repeated 5 times for each configuration of the parameters μ, ν and t. Each set of 5 trials averaged a total of less than 3 min on a Lenovo Y50 laptop.

After each sweep, we calculated the distribution of segment lengths for both the invisible and visible parts of the model genome.

Parameter estimation

The results of the simulations strongly suggest that the lengths of the invisible segments are gamma distributed, as illustrated in the Cullen-Frey graphs at the top of Fig. 2. As the parameters ν, μ and t change, the moments of the simulated distributions also change, but remain those of a gamma distribution. Similarly, the distribution of the lengths of the visible segments is always exponential, as at the bottom of Fig. 2, with rate

$$\lambda^{-1} = \frac{t}{\nu}. \qquad (17)$$

Table 2 $\mu = 1, \nu = 3, t = 5, \lambda^{-1} = 0.16656, \alpha = 0.6711, \beta = 0.3504$

par\time t	3	4	5	6	7
μ	1.231661635	1.063543459	1.021578166	1.018615845	1.033112259
ν	1.801158145	2.401544193	3.001930241	3.602316289	4.202702338
β_t	0.30056014	0.338494894	0.338654644	0.325314125	0.306788889
100δ	14.23403354	3.409208693	**3.363623543**	7.170390806	12.4566366

Bold entry indicates the t most consistent with the observed data on α, β and λ

As a first step towards the ability to infer μ and ν from the length distributions of invisible and visible segments, we would like to predict α and β, the shape and rate parameters of the gamma distribution, from t, μ and ν. Table 1 suggests, for a fixed value of t and a fixed value $\frac{\mu}{\nu}$, that shape is constant as μ changes, and that the rate is inversely proportion to μ.

Similar results hold for each combination of t and $\frac{\mu}{\nu}$, with different shape constants and rate proportions. Figure 3 shows how the shape constant varies with t for four values of $\frac{\mu}{\nu}$.

The four coefficients of the linear relationships inferred from Fig. 3 are plotted in Fig. 4. Fitting this curve with a quadratic yields

$$\alpha^{-1} - 1 =[-0.1725(\frac{\mu}{\nu})^2 + 0.5333\frac{\mu}{\nu} - 0.039]\,(t-1). \tag{18}$$

As for the rate parameter of the gamma, Fig. 5 shows that it is the logarithm of the rate that behaves linearly over time for a fixed value of $\frac{\mu}{\nu}$.

The four coefficients of the linear relationships inferred from Fig. 5 are plotted in Fig. 6. Fitting this curve with a quadratic yields

$$\beta^{-1} = \mu \exp\left[\left(-0.2458\left(\frac{\mu}{\nu}\right)^2 + 0.9257\frac{\mu}{\nu} - 0.0212\right)(t-1)\right] \tag{19}$$

The observable quantities in our model are the distribution of visible segment lengths, predicted to be exponential with mean λ, and the shape and rate parameters α and β of the predicted gamma distribution of invisible segment lengths. These three observable quantities are related to the unknown model parameters μ, ν, and t through Eqs. (17), (18) and (19). With the given value of these parameters, we can estimate the values of μ, ν, and t.

Lacking a closed form solution for μ, ν, and t in terms of λ, α and β, we use the following procedure. Since t must be an integer, we can find values of ν_t and μ_t for each $t = 1, 2, \ldots$ with Eqs. (17) and (18). Then we can solve Eq. (19) to find β_t.

Table 3 $\mu=6, \nu=12, t=2, \lambda^{-1} = 0.17, \alpha = 0.8488, \beta = 0.12063$

par\time t	2	3	4	5	6	7
μ	5.7892	4.7235	4.7286	4.9648	5.2990	5.6560
ν	12	18	24	30	36	42
β_t	0.1195	0.1406	0.1342	0.1220	0.1094	0.0976
100δ	**0.9107**	16.5215	11.2325	1.1655	9.3410	19.0791

Bold entry indicates the t most consistent with the observed data on α, β and λ

Table 4 $\mu = 1, \nu = 3, t = 3, \lambda^{-1} = 1.017737, \alpha = 0.7977859, \beta = 0.5649623$

par\time t	2	3	4	5	6
μ	1.4006	1.0332	0.9909	1.0102	1.0523
ν	1.9651	2.9477	3.9303	4.9129	5.8954
β_t	0.4271	0.5606	0.5596	0.5246	0.4810
100δ	24.39	**0.7780**	0.9497	7.15	14.87

Bold entry indicates the t most consistent with the observed data on α, β and λ

We then compare all the β_t, for $t = 1, 2, \ldots$ with the β observed in the simulation, and set

$$\hat{t} = \arg\min\left\{\frac{\beta - \beta_t}{\beta}\right\} \tag{20}$$

As an example, in one set of simulations where $\mu = 1, \nu = 3$ and $t = 5$, the experimental value of parameters are $\lambda^{-1} = 1.665595, shape = 0.6711252$ and $\beta = 0.3504422$. When $t \leq 2$, there is no solution for μ. For $t > 2$, Table 2 shows the results of this procedure, where 100δ is $100 \times$ the normalized difference between β and β_t in Eq. (20).

The minimum value of 100δ occurs when $t = 5$, expressing the fact that the inferred values of μ and ν, together with $t = 5$, are the parameter values most consistent with the observed values of α, β and λ. Other typical examples spanning a range of parameter values are given in Tables 3, 4 and 5.

It can be seen, at least in these diverse examples, that the inference procedure generally identifies the correct value of t, and good estimates of μ and ν.

Discussion

The introduction of sweeps consisting of alternating jumps and deletions, with time-invariant parameters ν and μ, provide us with an improved possibility of solving the fractionation model completely. We do announce such a solution, though it has much room for improvement. Though the exponential distribution of visible segment lengths should be easy to establish analytically, it is also possible that the gamma distribution of invisible segment lengths could be proved, including the α and β parameters as a function of the number of sweeps. Depending on

Table 5 $\mu = 5, \nu = 15, t = 8, \lambda^{-1} = 0.532632, \alpha = 0.53869147, \beta = 0.03107084$

par\time t	4	5	6	7	8	9	10
μ	6.2529	5.4956	5.2256	5.1247	5.1051	5.1313	5.1861
ν	7.5099	9.3873	11.2648	13.1423	15.0198	16.8972	18.7747
β_t	0.0281	0.0317	0.0324	0.0318	0.0306	0.0292	0.0277
100δ	9.43	2.18	4.22	2.33	**1.39**	6.00	10.97

Bold entry indicates the t most consistent with the observed data on α, β and λ

the functional form of such a solution, the inference of t, μ and ν might be amenable through closed form formulae rather than the quadratic modeling. Nevertheless, we have succeeded for the first time in inferring the parameters of a fractionation model, albeit a "one-sided" model and a continuous analog of more realistic discrete fractionation models.

Conclusions

Aside from theoretical improvements, the first priority for this work should be the return to a discrete gene-order model of fractionation with the insights gained in the current report. This should be extended to, or at least tested on simulations of, two-sided fractionation models with subgenome dominance (higher deletion rates on one copy of the genome than the other).

Acknowledgements
Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

Declarations
The publication charges for this article were funded by the Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2016-05585.
This article has been published as part of *BMC Bioinformatics* Vol 17 Suppl 14, 2016: Proceedings of the 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-14.

Availability of data and material
Not applicable.

Authors' contributions
The study was planned by DS and ZY, who also wrote the paper. The research was carried out by ZY. Both authors read and approved the paper.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 11 November 2016

References

1. Sankoff D, Zheng C, Zhu Q. The collapse of gene complement following whole genome duplication. BMC Genomics. 2010;11:313.
2. Wang B, Zheng C, Sankoff D. Fractionation statistics. BMC Bioinforma. 2011;12:S9—S5.
3. Sankoff D, Zheng C, Wang B. A model for biased fractionation after whole genome duplication. BMC Genomics. 2012;13:S1—S8.
4. Sankoff D, Zheng C, Wang B, Buen Abad Najar CF. Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling. BMC Bioinforma. 2015;16:S17—S9.
5. Byrnes JK, Morris GP, Li WH. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol Biol Evol. 2006;23:1136–43.
6. van Hoek MJ, Hogeweg P. The role of mutational dynamics in genome shrinkage. Mol Biol Evol. 2007;24:2485–94.
7. Zheng C, Wall PK, Leebens-Mack J, dePamphilis C, Albert VA, Sankoff D. Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral Populus diploid. J Bioinforma Comput Biol. 2009;7:499–520.