




# Gaps and Runs in Syntenic Alignments

Zhe Yu, Chunfang Zheng, and David Sankoff<sup>(✉)</sup> 

University of Ottawa, Ottawa, Canada  
{zyu096,czhen033,sankoff}@uottawa.ca

**Abstract.** Gene loss is the obverse of novel gene acquisition by a genome through a variety of evolutionary processes. It serves a number of functional and structural roles, compensating for the energy and material costs of gene complement expansion.

A type of gene loss widespread in the lineages of plant genomes is “fractionation” after whole genome doubling or tripling, where one of a pair or triplet of paralogous genes in parallel syntenic contexts is discarded.

The detailed syntenic mechanisms of gene loss, especially in fractionation, remain controversial.

We focus on the the frequency distribution of gap lengths (number of deleted genes – not nucleotides) within syntenic blocks calculated during the comparison of chromosomes from two genomes. We mathematically characterize a simple model in some detail and show how it is an adequate description neither of the *Coffea arabica* subgenomes nor its two progenitor genomes.

We find that a mixture of two models, a random, one-gene-at-a-time, model and a geometric-length distributed excision for removing a variable number of genes, fits well.

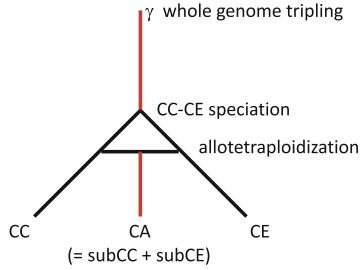
**Keywords:** Gene loss · Tetraploidy · Fractionation · Plant genomes · Coffee · Run length

## 1 Introduction

The evolutionary process of gene loss, through DNA excision, pseudogenization or other mechanism, is the obverse of novel gene acquisition by a genome through processes such as tandem duplication, gene family expansion, whole genome doubling, neo- and subfunctionalization and horizontal transfer. Loss serves a number of functional and structural roles, mainly compensating for the energetic, material and structural costs of gene complement expansion.

A type of gene loss widespread in the lineages of plant genomes, and also occurring in a few yeast, fish and amphibian genomes, is “fractionation” after whole genome doubling or tripling, where one of a pair or triplet of paralogous genes in parallel syntenic contexts is discarded.

Quantitative studies have focused on many aspects of gene loss. In this paper, we study the evolutionary history of the allotetraploid *Coffea arabica* (CA) and



**Fig. 1.** *Coffea* phylogeny. Fractionation operates in lineages coloured red. (Color figure online)

its two diploid progenitors, *Coffea canephora* (CC) and *Coffea eugenoides* (CE), annotated genome assemblies being provided by the Arabica Coffee Genome Consortium [1]. This history is summarized in Fig. 1. We survey gene loss in three periods. These are

- loss from the ancestral lineage leading from the  $\gamma$  whole genome tripling event [2] 120 million years ago, due at least partly to fractionation,
- independent losses from the CC and CE genomes after speciation (but before allotetraploidization) around 10 million years ago [3], and
- loss from the CC and CE, and the subCC and subCE subgenomes of CA, following the allotetraploidization event. Loss from the two subgenomes (namely those chromosomes in CA deriving from CC and those deriving from CE) can be attributed to fractionation.

We first study the distribution of gene pair similarities derived from the comparison of the four genomes and subgenomes. This will serve to confirm the validating parallels between CC and CE evolution, and between subCC and subCE evolution.

We then introduce our main analytical construct, the frequency distribution of gap lengths within syntenic blocks calculated during the comparison of chromosomes from two genomes or subgenomes. In the simplest model, proposed over ten years ago [4–6], at each step a random gene pair is selected to lose one member. In a new version of this model that takes into account chromosome length, we develop an exact recurrence to calculate the expected number of gaps of each length after a given number of steps. We then provide evidence from the *Coffea* data that demonstrates a systematic departure from this model.

In a competing class of models [7,8], gene loss is effected by excision of a variable length fragment of a chromosome, often formulated in terms of a gamma distribution. In the *Coffea* data, there are far too many single-gene deletions for this solution, but a mixture of the two models, where the gamma is actually a single-parameter geometric distribution, fits well.

## 2 Methods

Our research is based on the homologous gene pairs in syntenic context as produced from the data on pairs of genomes by the SYNMAP procedure on the CoGE platform [9,10]. At a general level, we used the “peaks” method [11] for the three events that generate duplicate genomes in the evolution of CA: gamma hexaploidization, CC/CE speciation and CA tetraploidization (which is effectively a speciation of CC/CA-subCC and of CE/CA-subCE). In this method, the local modal values (peaks) of the distribution of the entire set of homologous gene pairs, as calculated by the R function `geom_density`, are estimates of the time of the event. We could also have used EMMIX [12] or other mixture of distributions methods to carry this out.

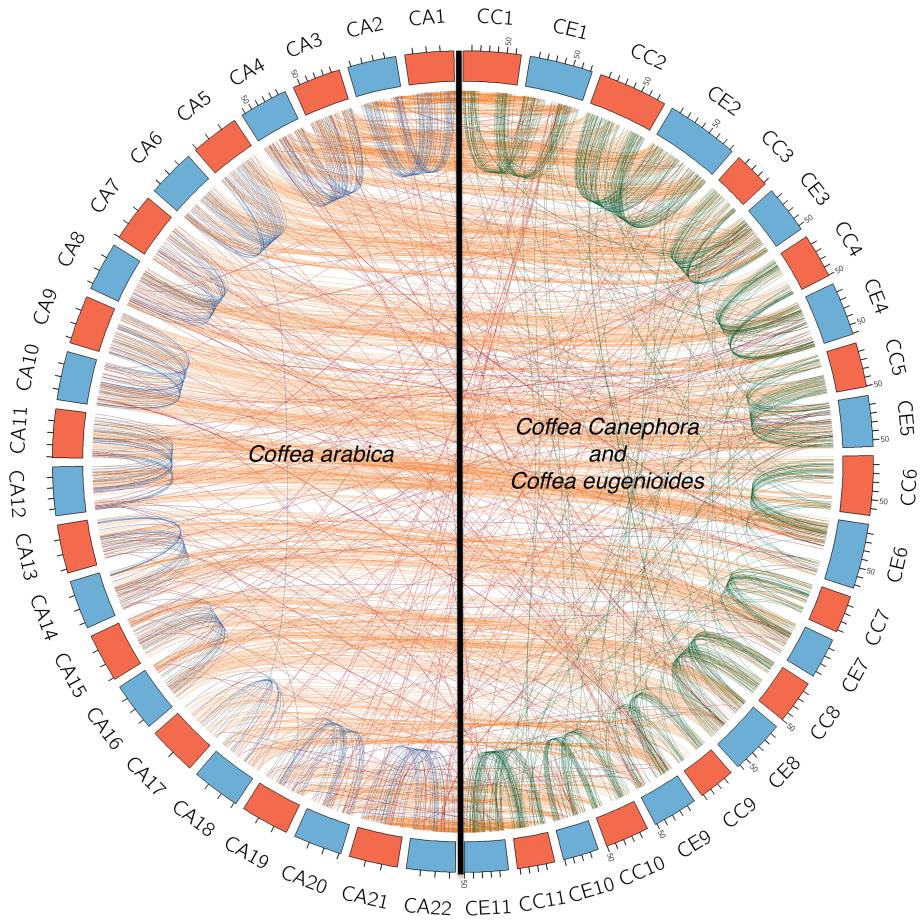
This allowed us to study the evolution of paralogous and orthologous synteny blocks. We considered only genes within the region of the blocks, including gene pairs and singleton genes in each genome that have lost their counterpart in the other genome due to fractionation or other gene loss. We used all four genomes, CC, CE, CA-subCC (denoted just subCC) and CA-subCE (denoted just subCE), producing six comparisons of pairs, and four self-comparisons. We did not look at the whole CA assembly, just the large majority that was successfully separated into the subgenomes.

We studied a number of statistics on the gaps between adjacent pairs of duplicate genes within synteny blocks, the innovative focus of this work, and here report on one of them, the size of gaps between two adjacent duplicate pairs on genes in a block, from 0 (no gap) to a maximum of 10 on either one of the genomes. We make certain operational definitions to allow us to analyze evolution coherently across all evolutionary eras. For example, if we encounter more than 10 genes in a gap on one genome between two adjacent gene pairs, we break up the synteny block into two at that point. This is justified by the regular decrease in frequency in gap sizes from 0, 1, 2, until there are almost none of size 8, 9, or 10, except between neighbouring synteny blocks, which can be separated by large numbers of unpaired genes in either of both of the genomes. We want to study the nature of the distribution of gap size due to fractionation or gene deletion, and this avoids biasing estimates by inclusion of gaps produced by mechanisms other than fractionation. Thus, we use the default parameters of SYNMAP, except for the maximum number of non-duplicate genes interrupting any neighbouring gene pairs, which we set at 10.

## 3 Results

### 3.1 The Sequence of Evolutionary Events

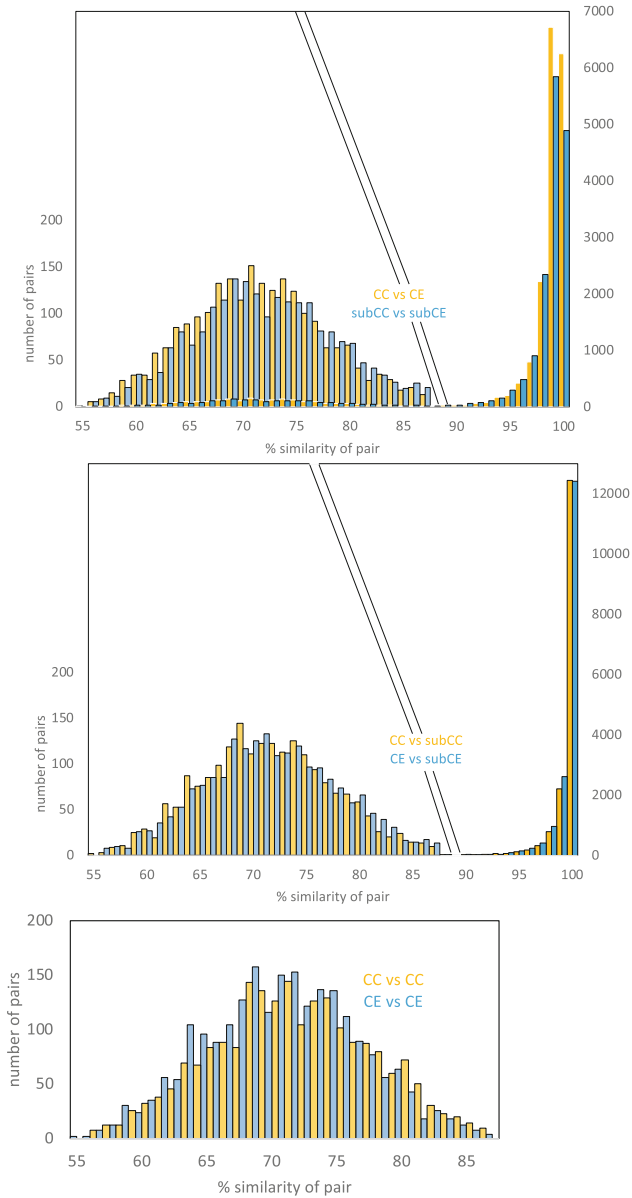
Some 28,800, 33,500 and 56,700 genes were identified in the annotations of CC, CE and CA, respectively, while the subCC and subCE subgenomes identified in CA contained 24,700 and 25,800 genes respectively. Amalgamating the gene pairs in all SYNMAP comparisons produces the CIRCOS plot in Fig. 2.



**Fig. 2.** CIRCOS plot of pairs of syntenic blocks among *Coffea arabica* (CA) and its two diploid progenitors, *Coffea canephora* (CC) and *Coffea eugenioides* (CE), as well as paralogous pairs between homeologous chromosomes in CA. Note that subCC/subCE blocks are about twice as long on average (67 gene pairs) as CC/CE blocks (33 gene pairs), resulting in fewer connecting arcs (about half as many), and a lighter shade apparent in the bundles of connecting arcs on the left of the circle.

### 3.2 The Distributions of Gene Pair Similarity

All ten self- and pairwise comparisons show a cluster of homologous pairs dating from the early hexaploidization of the core eudicots. Figure 3 depicts six of the distributions, two dating from the CC/CE speciation, two from the tetraploidization event and two from the  $\gamma$  event itself. Table 1 comparing averages over all pairs with less than 87% similarity, indicates tight clustering of these estimates, in terms of peak gene similarity (over CDS regions).



**Fig. 3.** Gene pairs originating in speciation (top). Gene pairs originating in tetraploidization (middle). Gene pairs originating in  $\gamma$  event (bottom)

**Table 1.** Locating the  $\gamma$  hexaploidy in all comparisons; peak of distribution of pairs with less than 87% similarity.

Comparison	Peak of similarity (%)	# of pairs
CC vs CE	73.7	2069
subCC vs CE	67.8	1860
subCE vs CC	68.0	2105
subCC vs subCE	67.9	1922
CC vs CC	71.5	1163
CE vs CE	70.4	1056
subCC vs subCC	68.2	938
subCE vs subCE	70.1	1043
subCC vs CC	68.2	1949
subCE vs CE	67.8	1925
Mean $\pm$ S.D.	69.4 $\pm$ 2.0	1603 $\pm$ 484

The speciation of CC and CE generates orthologous gene pairs visible in the CC (or subCC) vs CE (or subCE) comparisons, as can be seen on the right hand side of the top panel in Fig. 3. Table 2 presents the peak similarity for these comparisons.

**Table 2.** Locating the CC-CE speciation

Comparison	Similarity (%)	# of pairs
CC vs CE	99.12	17,066
subCC vs CE	99.08	15,985
subCE vs CC	99.11	16,014
subCC vs subCE	99.05	15,318
Mean $\pm$ S.D	99.09 $\pm$ 0.033	16,096 $\pm$ 626

The CA tetraploidization event, which for our purposes consists of the synchronous speciation of CC/subCC and CE/subCE, is visible as a peak of gene pairs in the CC vs subCC and the CE vs subCE comparisons on the right hand side of the middle panel in Fig. 3. These peaks are listed in Table 3.

The current best estimates of  $\gamma$  and CC/CE *Coffea* speciation are of the order of 120 My and 10 My [3], while the CA tetraploidization is thought to be less than 1 My old. The similarity measures does not correspond well to this timeline. The tetraploidy seems to be 15–20% of the speciation age.

**Table 3.** Locating the tetraploidization event

Comparison	Similarity (%)	# of pairs
CC vs subCC	99.81	16,487
CE vs subCE	99.87	17,196
Mean $\pm$ S.D.	99.84 $\pm$ 0.043	16,842 $\pm$ 501

## 4 One-at-a-Time Model

Consider the following “fractionation” process. We have an array of  $n$  1’s. At the first step, and every subsequent step, we pick a 1 at random and transform it to 0. We stop after a given number of steps  $t \leq n$ .

We prove a recurrence for  $M(t, x)$ , the expected number of runs of 1’s (more precisely, maximal runs) of length  $x$  at time  $t$ .

### Proposition 1

$$\begin{aligned} M(0, n) &= 1 \\ M(0, x) &= 0, \quad \text{for } x \neq n \end{aligned} \quad (1)$$

Thereafter, for  $1 \leq t \leq n - 1$  and  $1 \leq x < n - t + 1$

$$M(t, x) = M(t - 1, x) - \frac{xM(t - 1, x) - 2 \sum_{i>x}^n M(t - 1, i)}{n - t + 1}, \quad (2)$$

*Proof.* The initial values of the process at  $t = 0$  are fixed by definition, and so then are their averages  $M(0, x)$ .

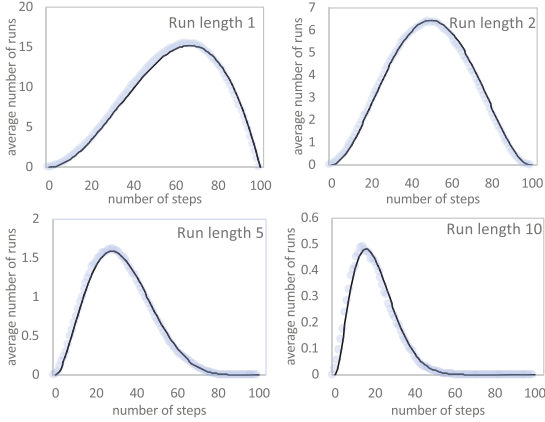
For each  $t > 0$ , in randomly changing one of the  $n - t + 1$  remaining 1’s in the array to 0, there are two mutually exclusive possibilities. An existing run of length  $x$  can be destroyed, for some  $x \geq 1$ , which can happen  $xM(t - 1, x)$  ways. Alternatively a run of length  $x$  can be created. This can occur in exactly two ways in breaking up any remaining run of length greater than  $x$ .

The average change is obtained through division by the total number of cases  $n - t + 1$ .  $\square$

There is a symmetry in the fractionation process, in that the evolution of the number of 1’s, and the probabilistic structure governing the distribution of run sizes, starting from time  $t = 0$ , is identical to the evolution of the number of 0’s, and the probabilistic structure governing the distribution of gap sizes, starting from time  $t = n$ .

To illustrate the the evolution of run lengths, Fig. 4 shows how longer runs only survive at the beginning of the process, and how the number of shorter runs increases until they too are lost to fractionation. Of interest is the case of run length 2, where the symmetry of gaps and runs is clearest.

This process bears much resemblance to the theory of runs [13] in random binary sequences. Given  $n$  Bernoulli trials with a probability of success  $p = t/n$ ,



**Fig. 4.** Evolution of number of runs of 1’s of various sizes as the number of steps  $t$  increases. Solid line: recurrence. Light blue background line: average of 1000 simulations. Genome length  $n = 100$  (Color figure online)

the expected number of successes is  $t$ , and the expected number of runs of length  $x$  is  $M(t, x)$ . However, the variance of the number of successes is non-negligible, whereas it is zero for our process, and the variance of the number of runs of a given length is also greater than our process. Thus our interest in the fractionation process, where the probability of success at each position depends on the total number of successes already achieved.

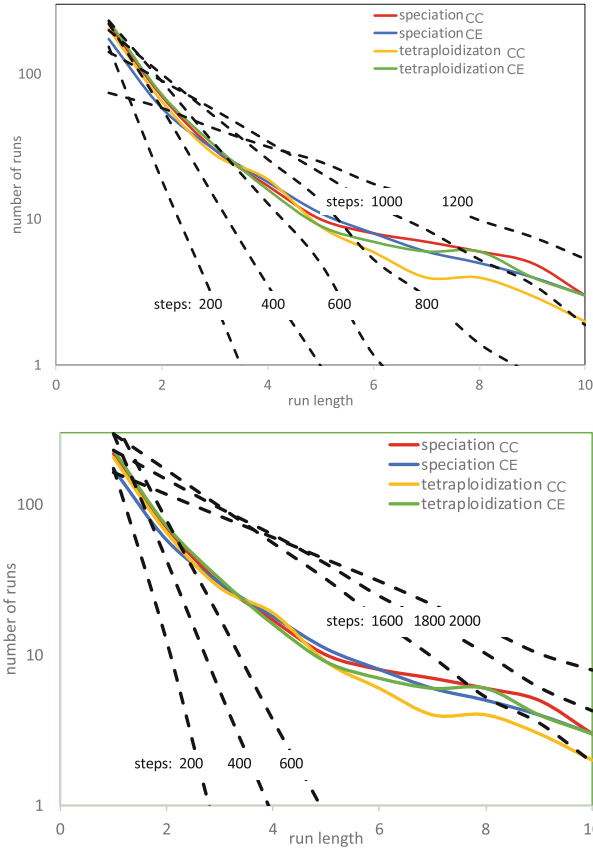
When we compare the behaviour of our “one-at-a-time” model with the gaps in the *Coffea* data in Fig. 5 however, it is clear that the model is inadequate to account for both the simultaneous steep drop-off from gaps of size 1 and the presence of significant number of long gaps.

#### 4.1 The Combined Model

To remedy the poor fit of the one-at-a-time model, we combine it with a gamma distribution component. Whereas the one-at-a-time model involves a single fixed parameter, chromosome size  $n$ , a gamma component adds a shape and a scale parameter, as well as weight parameter to apportion the two components. Fortunately, the optimal gamma component turned out to be a simple geometric distribution, with only one parameter.

To estimate the  $n$ , the geometric parameter  $\lambda$ , and the proportion of steps  $\theta$  allocated to the one-at-a-time model, we compared data on runs of 1’s and runs of 0’s, from the both the speciation event and the tetraploidization event, all taken together. We optimized in terms of a chi-square criterion, when running 50 simulations based on a range of values of  $n$ ,  $\lambda$  and  $\theta$ . (This was after finding that the two parameters of a general gamma distribution did not substantially improve the fit compared to a geometric distribution.) Of importance, however,

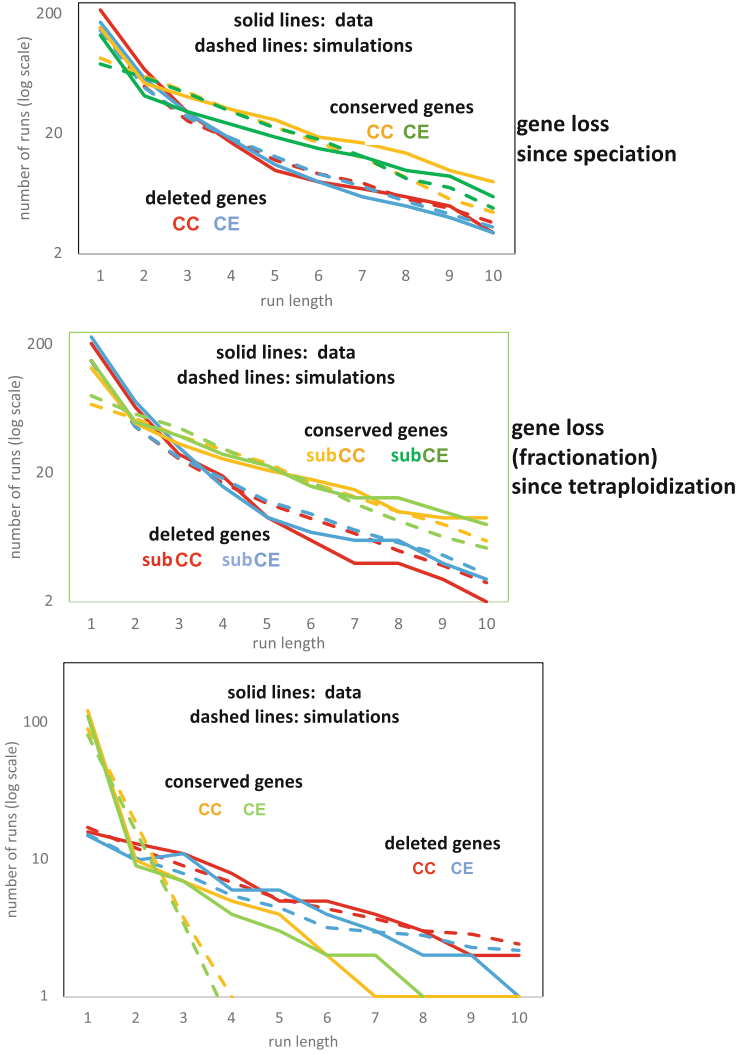




**Fig. 5.** Inability of one-at-a-time model to fit *Coffea* data on runs of zeros (gaps), with chromosome size  $n = 1600$  (top) or  $n = 2800$  (bottom), and at various time intervals (steps)

was that we allowed different numbers of steps in the speciation and tetraploid simulations. The values were 705 for speciation and 590 for tetraploidization, which is coherent with the historical ordering of these two events, and with the mean similarities in Tables 2 and 3. The optimal values were  $\theta = 0.7$ ,  $n = 2800$  and  $\lambda = 2/7$ .

Our model breaks down when we use it to simulate fractionation after  $\gamma$ , as can be seen in the bottom panel of Fig. 6. The simulations suggest there should remain no long runs of 1's, but this is likely due to the inability to detect sufficiently long syntenic blocks after extensive fractionation, and possibly some tendency for some regions of neighbouring genes to resist fractionation.



**Fig. 6.** Comparisons of the distributions of run sizes (0's and 1's) with simulations of combined model in the speciation data (top panel), tetraploidization data (middle panel) and  $\gamma$  data (bottom panel).

## 5 Conclusions

It can be noted that in all of our comparisons, there has been a symmetry between CC and CE, and between subCC and subCE. If  $\gamma$  fractionation rates or evolutionary divergence rates of CC and CE or subgenome dominance play a role, their effects must be relatively small.

We have found several indications that the time span since tetraploidization, is almost as long as the period since speciation.

We have investigated the one-at-a-time model in some detail, but it is clearly inadequate to explain the gene loss data, which is surprisingly parallel between post-speciation loss and fractionation. Adding a geometric component, however, allows the model to fit the data quite well.

The distribution of gap sizes in syteny blocks generated by all evolutionary events confirms that gene loss, by fractionation or otherwise, proceeds largely by the loss of one gene at a time, with  $\theta = 0.7$  and further loss from the geometric component.

**Acknowledgments.** 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.

## References

1. De Kochko, A., Crouzillat, D.: Arabica coffee genome consortium: Aims and goals of the Arabica Coffee Genome Consortium (ACGC). In: 12th Solanaceae Conference (2015)
2. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007)
3. Hamon, P., Grover, C.E., et al.: Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (*Coffea*) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Mol. Phylogenet. Evol.* **109**, 351–361 (2017)
4. van Hoek, M.J., Hogeweg, P.: The role of mutational dynamics in genome shrinkage. *Mol. Biol. Evol.* **24**, 2485–2494 (2007)
5. Byrnes, J.K., Morris, G.P., Li, W.H.: Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol. Biol. Evol.* **23**, 1136–1143 (2006)
6. Zheng, C., Wall, P.K., Leebens-Mack, J., dePamphilis, C., Albert, V.A., Sankoff, D.: Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* diploid. *J. Bioinform. Comput. Biol.* **7**, 499–520 (2009)
7. Sankoff, D., Zheng, C., Wang, B., Fernando Buen Abad Najjar, C.: Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling. *BMC Genomics* **15** (2015). <https://doi.org/10.1109/ICCABS.2014.6863915>
8. Yu, Z.N., Sankoff, D.: A continuous analog of run length distributions reflecting accumulated fractionation events. *BMC Bioinform.* **17**(suppl 14), 412 (2016)
9. Lyons, E., Freeling, M.: How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008)
10. Lyons, E., et al.: Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008)

11. Sankoff, D., Zheng, C., Zhang, Y., Meidanis, J., Lyons, E., Tang, H.: Models for similarity distributions of syntenic homologs and applications to phylogenomics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 727–737 (2019)
12. McLachlan, G.J., Peel, D., Basford, K.E., Adams, P.: The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Softw.* **4**, 1–14 (1999)
13. Weisstein, E.: Run. MathWorld-A Wolfram Web Resource. <http://mathworld.wolfram.com/>. Accessed 20 Aug 2019