Fractionation in the evolution of syntenic homology in *Coffea Arabica*

Zhe Yu

Thesis submitted to the University of Ottawa in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mathematics[1]

Department of 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 Mathematics and Statistics

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.

Based on previous mathematical work on the distribution of gap sizes caused by fractionation in synteny blocks, we studied fractionation in the evolutionary history of the allotetraploid *Coffea arabica* (CA) and its two diploid progenitors, *C. canephora* (CC) and *C. eugenioides* (CE), annotated genome assemblies being provided by the Arabica Coffee Genome Consortium. By taking advantage of synteny blocks produced by SynMap, we studied the fractionation process after speciation and tetraploidization events, including visualization and modelling the distribution of deletion segments, and mechanisms of deletion events. We also expanded the research to eight other plant species to verify the dominance of DNA excision over pseudigenization during the fractionation and other gene loss.

Dedications

This thesis is dedicated to my parents, for their endless love, support and encouragement in my life. It is also dedicated to my grandparents, for their inspiration and unconditional kindness to me.

Acknowledgement

I would first like to express my deepest gratitude to my supervisor, Prof. David Sankoff, for their continuous guidance and encouragement on my research. His patience, motivation, and immense knowledge helped me to choose the right direction and finish this dissertation. The research would not be possible without his great support. It is my fortunate and honour to work with him during my graduate study. I am also very grateful to Prof. Yiqiang Zhao and Monica Nevins for their invaluable comments, suggestions, and feedbacks. My sincere gratitude also goes to Dr. Benoit Dionne. His door is always open and I can always find him when needed and been generous in providing help.

Besides, I would like to thank my fellow colleagues in the laboratory, Chunfang Zheng, Mona Meghdari, Arash Jamshidpey, Yue Zhang, Qiaoji Xu, Daniella Santos Munoz, Eric Lam, Fatemeh Pouryahya, Sindeed Islam, Alma Oladi, Yuji Jeong, Poly Hannah da Silva, Caroline Anne Larlee, Manuel Lafond and Joao Meidanis. I am benefited much from their collaboration and discussion.

Last but not the least, I would like to thank my family: My mother, Jie Tang, the kindest and sweetest woman ever, who was always there cheering me up and who stood by me though the good times and bad. My father, Jixiang Yu, the best role model, who supported me and encouraged me with his best wishes. They gave birth

to me in the first place and supported me emotionally throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

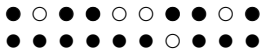
Introduction

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, such as compensating for the energetic, material and structural costs of gene complement expansion and correcting for functional imbalances caused by duplicatior processes.

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 in some cases by unusual patterns of meiosis where more than just one maternal and paternal chromosome are aligned and recombine, processes of sequence divergence, redundant gene loss, and chromosome rearrangement lead to more familiar diploid patterns.

Polyploidization is considered a mutation, and is not part of normal life cycle of

genome replication. A polyploid contains more than one copy of each chromosome in the diploid genome. These homeologous chromosomes diverge in DNA sequence and gene content or order though various processes such as chromosome fissions, inversions, translocations and redundant gene loss. The gene loss process is called fractionation (shown as following and Figure 1.1). The possible mechanisms of fractionation can be schematized as follows (drawn from Chapter 4).



Synteny block on homeologous regions of two chromosomes.
Dark circles indicate retained genes, white circles deleted genes.
There are five retained duplicate gene pairs, four singletons on the lower chromosome and one singleton on the upper chromosome.

Fractionation processes have been surveyed across evolutionarily diverse types of eukaryote organisms [16] as illustrated by the fifteen genomes descending from six whole genome duplication, shown in Figure 1.2.

Terminology

homologous: descending from a common ancestor

orthologous: homologous in two different species

paralogous: homologous within a single species

homeologous: paralogous resulting from genome duplication

Fractionation has been intensively studied in the Sankoff lab since 2008 [15, 29, 25, 17, 2, 30]. Among the preoccupations has been how gene loss affects neighbouring genes on the chromonosome and whether one chromosome is more vulnerable to losing its extra genes than its homeologous chromosome. Another is the distribution of gene runs of fractionated or unfractionated genes. Many models and methods were

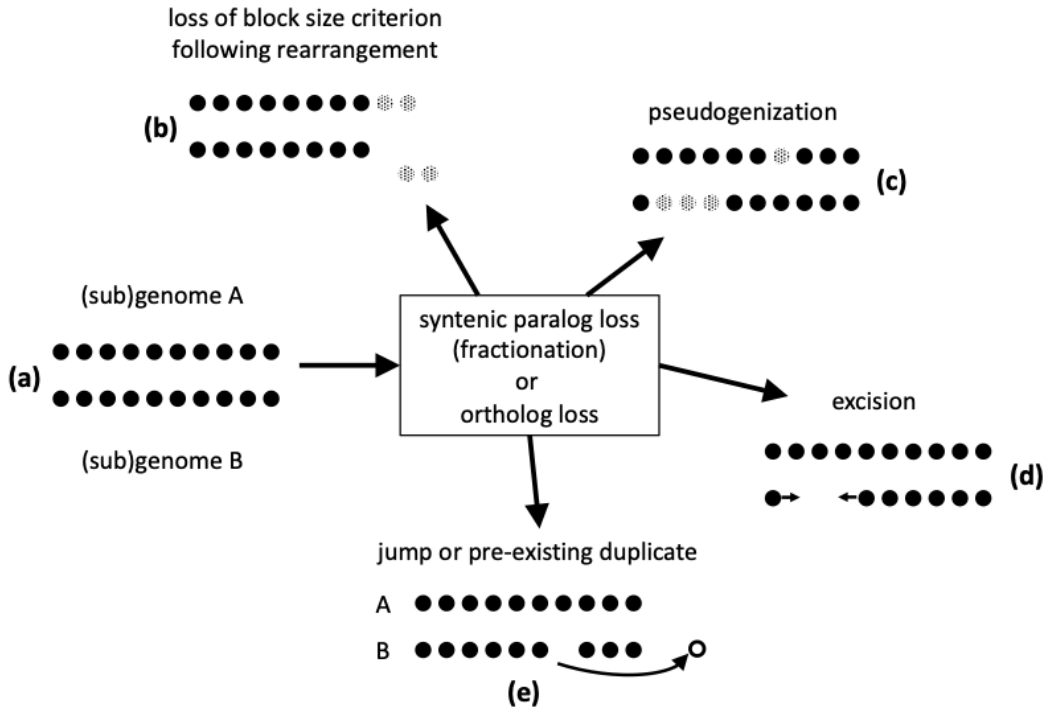


Figure 1.1: Mechanisms of loss of gene pair homology after polyploidization. (a) synteny block made up of co-linear homologous pairs. (b) erosion of synteny block by translocation to a remote chromosomal location of a portion of sub-threshold length. (c) Pseudogenization. Genes rendered inoperable represented by grey dots. (d) Excision of DNA fragment including one or more genes. Arrows represent a new adjacency after the loss of the excised genes. (e) Jump of one member of pair to a different genomic location, or loss of only one of two or more homologs of the same gene.

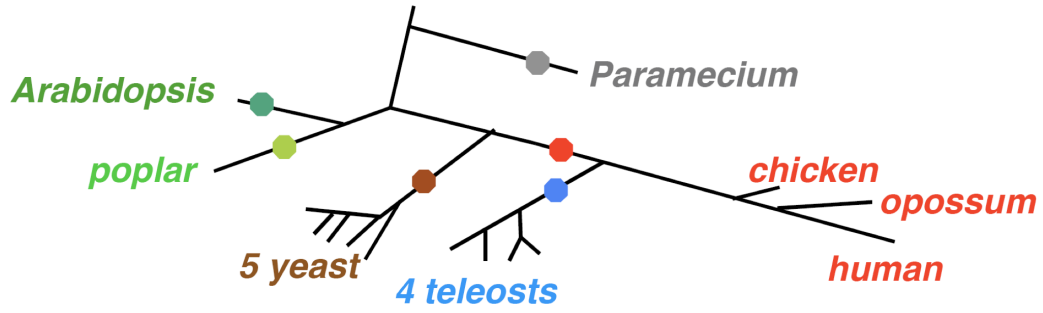


Figure 1.2: Locating Whole Genome Duplication (WGD) events on eukaryotic phylogeny. The WGD events (shown as colored dots) are as signed the following dates: 20, 50, 70, 150, 350, 450 Mya for Paramecium, Arabidopsis, Populus, yeast, teleosts (fish), and higher vertebrates. From [16]

introduced during this time. My masters thesis investigated the distribution question by using a continuous approximation to the original problem, which was discrete.

Five years ago, an international consortium was put together to sequence the arabica coffee genome, and to study its evolution and its population genetics. Our lab, which was previously involved with the consortium that sequenced the robusta coffee genome [3], was asked to participate in the analysis of evolution.

Arabica coffee is the only one of the more than 120 species in the coffee genus which is an ancient tetraploid. Thus it has also undergone fractionation and this became the focus of my thesis.

In this thesis, I introduce numerous new data collection methods, display techniques, statistical genomic analyses, probabilistic models and combinatorial algorithms on fractionation in arabica coffee. In the present introduction, I first sketch some of our previous work on the topic. Then I turn to a brief discussion of the particulars of the arabica coffee genome, and finally an overview of the the rest of the thesis. Much of this work in this thesis has now been published or has been prepared for incorporation in the main arabica coffee genome sequence paper to be

submitted in the next months. The thesis is then an amalgam of published work and new writing.

Papers I have published relevant to this thesis include:

1. A continuous analog of run length distributions reflecting accumulated fractionation events (Z.N. Yu, D. Sankoff), BMC Bioinformatics 17, (Suppl 14):412 (2016).

   This is reviewed in Chapter 1.1.1 below.

2. Gaps and runs in syntenic alignments (Z. Yu, C. Zheng, D. Sankoff), International Conference on Algorithms for Computational Biology, AlCoB 2020: Algorithms for Computational Biology, Lecture Notes in Computer Science 12099: 49-60 (2020)

   This supplies some of the material in Chapter 2 and Chapter 6.

3. Excision dominates pseudogenization in gene loss after speciation and in fractionation after whole genome duplication in plants (Z. Yu, C. Zheng, V.A. Albert, D. Sankoff), Frontiers in Genetics - Evolutionary and Population Genetics (2020).

   This is included in Chapter 4 and Chapter 5.

4. Salojärvi, Jarkko and 60 co-authors (to be submitted). Paper on the genome sequences of *C. arabica*, *C. canephora* and *C. eugenioides*. This paper features text and figures from Chapter 2 to 6 in this thesis, especially material from Chapters 2 and 3.

5. Integrated synteny- and similarity-based inference on the polyploidization-fractionation cycle (Y. Zhang, Z. Yu, C. Zheng, D. Sankoff), to appear in Interface Focus (2021).

   This recent work is discussed in Chapter 1.1.3.

In addition, I presented the material in Chapter 6 as posters and short talks at the Workshop on Comparative Genomics (RECOMB-CG) in Montpellier, France 2019, at the 2020 Annual Conference on Intelligent Systems for Molecular Biology (virtual) and the Canadian Mathematical Society Winter Meeting (2019) in Toronto.

1.1 Previous work

1.1.1 Distribution of deleted segment length

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. Previous work focused on the difficult question of how many overlapping deletion events are responsible for each contiguous deleted region [25, 18, 19] based on the distribution of lengths of contiguous deleted regions, but was not able to account analytically for the dynamics of the process. Figure 1.4 illustrates the distribution of cumulative deletion lengths simulated by repeated application of a deletion process with geometrically distributed (mean μ) deletion lengths. These can be used to to address the original problem of discriminating between the gene-by-gene "functional model" ($\mu = 1$) and the random excision "structural" model ($\mu > 1$) [1, 22]. There have been a number of analytical studies of this problem [29, 25]. To contribute to the understanding of the distribution of lengths of contiguous deleted segments, in

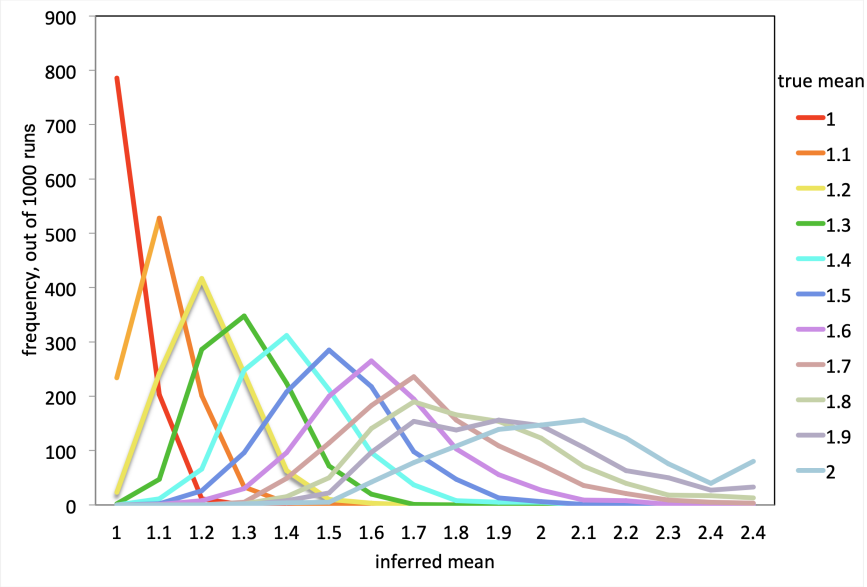


Figure 1.3: Distribution of cumulative deletion lengths simulated by repeated application of a deletion process with geometrically distributed (mean μ) deletion lengths. Colour keys for curves at right. All data involve a proportion of $1 - \theta = 0.20$ deleted genes. Length of genome= 500.

my master's thesis, I proposed a new, continuous model of the fractionation process 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.

To model the fact that successive deletions occur one after another in time, I postulated a number of deletion sweeps of the genome. During the first sweep, illustrated at the top of Figure 1.5 at time (or step) $t = 1$, the first deletion point x_1 is determined by sampling from the exponential distribution

$$\rho(x) = \frac{1}{\nu} e^{-\frac{x}{\nu}}, \quad x \geq 0, \tag{1.1.1}$$

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

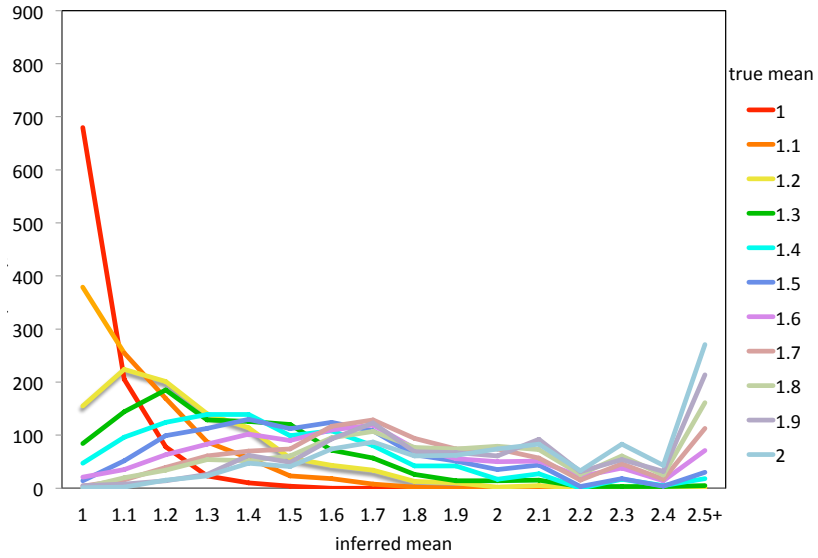


Figure 1.4: Distribution of cumulative deletion lengths simulated by repeated application of a deletion process with geometrically distributed (mean μ) deletion lengths. All data involve a proportion of $1-\theta = 0.20$ deleted genes. Length of genome= 100.

tion

$$\gamma(a) = \frac{1}{\mu} e^{-\frac{a}{\mu}}, \quad a \geq 0, \tag{1.1.2}$$

with mean μ. 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 $\gamma(a)$ 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.

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, as illustrated on the bottom of Figure 1.5.

After a number of analytical results and simulations, we were able to provide a

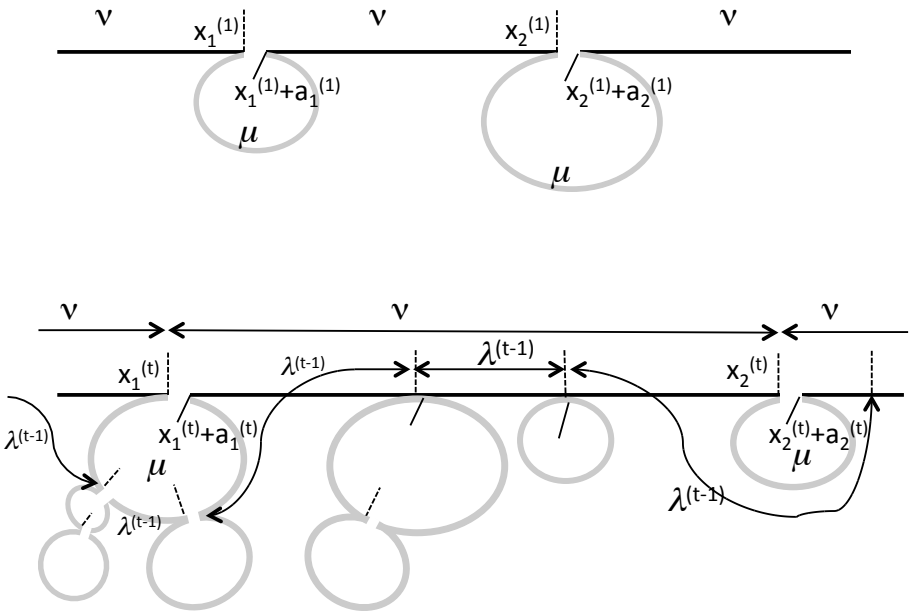


Figure 1.5: Processes pertinent to first sweep and t-th sweep. Solid horizontal bars represent the visible regions of the genome. Grey curves represent invisible regions caused by previous deletions. Dashed markers represent deletion points, solid markers represent end of deletion segments. ν and μ are the means of the deletion point spacing and deletion segment length variables, while $\lambda^{(t-1)}$ is the mean space between visible deletion points after the $t-1$st sweep. From [30]

way to estimate the parameters of the distribution of visible and invisible contiguous segments.

1.1.2 Consolidation

In the study of genome rearrangement and genomic distance, fractionation will greatly exaggerate the overall amount of chromosomal rearrangement between the two sister genomes. [17, 2] developed a way to compensate for the resulting inflation of distances. To eliminate the biases, a "Consolidation" algorithm was developed to identify and account for regions of fractionation. This algorithm wipes out almost all biases caused

by fractionation.

A more recent consolidation algorithm solves the more difficult problem of comparing two fractionated genomes while dispensing with a necessity of referencing an unduplicated genome [2].

1.1.3 Fractionation models

A series of branching process models of the fractionation process have been developed in our lab over the last four years [33, 34, 35, 36, 37, 38]. Recently, a major improvement to the kind of data used for estimating the parameters of these models allows for a much broader range of problems that can be studied (Integrated synteny- and similarity-based inference on the polyploidization-fractionation cycle (Y. Zhang, Z. Yu, C. Zheng, D. Sankoff), to appear in Interface Focus (2021)). I provided the methodology for extracting these new data using in Chapters 5 and 6.

1.2 The *Coffea arabica* project by ACGC

This thesis studies various aspects of fractionation in *Coffea arabica* (CA), including the evolutionary history of the allotetraploid CA (with subgenomes deriving from *Coffea canephora* (CC), and *Coffea eugenioides* (CE), denoted subCC and subCE, respectively) and its two diploid progenitors, CC and CE, annotated genome assemblies being provided by the Arabica Coffee Genome Consortium (Figure 1.6) [9].

Arabica (CA) and robusta (or canephora CC) coffee are the two major crop species of *Coffea*. CC and CE are two diverse species within the Africa subclade with observable difference in geography and caffeine content, as outlined by Hamon in [8]. CE is a high-altitude species as is CA, conversely, CC is from low-altitude West and

Central Africa. *C. arabica* originated from those two diploid parents and become the only tetraploid species in genus *Coffea*. The consortium obtained a good quality *C. arabica* genome sequence with clear distinction of the two parental subgenomes [9]. Research on the tetraploidation event in *C. arabica* evolutionary history provides a possible identification on the genetic diversification induced by human intervention, since divergence of two species supports the adaptation to environments and disease resistance.

Figure 1.6: Researchers contributed in Arabica Coffee Genome Consortium.

This history is summarized in Figure 1.7.

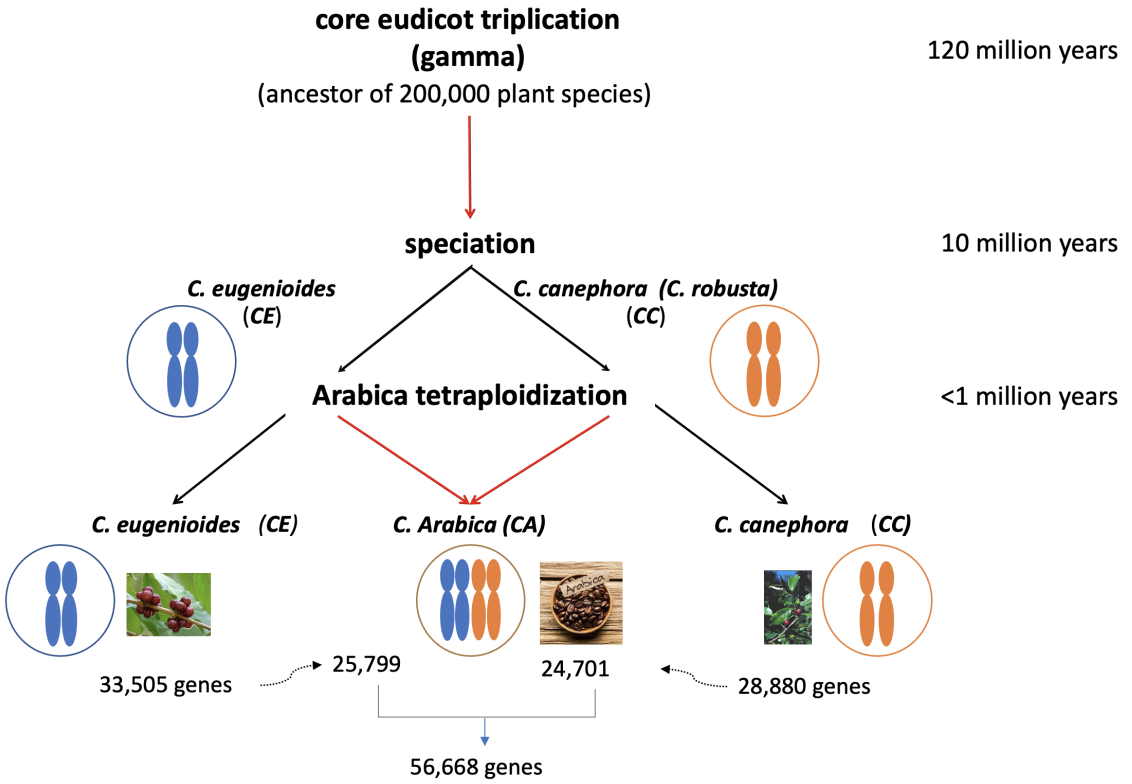


Figure 1.7: *Coffea* phylogeny. Fractionation operates in lineages coloured red.

We survey gene loss in three periods. These are

- loss from the ancestral lineage leading from the γ whole genome tripling event [5] 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 [8], which cannot be due to fractionation since these were independently evolving species, 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.

1.3 Data

Much of our data collection and initial organization of genome statistics is based on the SynMap [12] program on the CoGe platform. This This uses a complex series of algorithms to align genomes and to detect pairs of homologous genes in two genomes, or in self-comparison of a single genome, and organizes them into local "synteny blocks". This includes the calculation of gene pair similarities.

There are 56,668 genes in the assembly of CA, of which 24,701 are in the subCC, and 25,799 are in the subCE, leaving the remaining 6,168 unassigned genes in a construct called "chromosome 0". We also have assembled and annotated sequence data from the diploid "progenitors" CC (28,880 genes) and CE (33,505 genes) genomes.

1.4 Overview of the Thesis

Chapter 2 first studies the distribution of gene pair similarities derived from the comparison of the four genomes and subgenomes. This will serve to confirm the parallels between CC and CE evolution, and between subCC and subCE evolution.

Frequency of gap size in synteny blocks presented in Chapter 3, provides a brief overview of the mechanisms of gene loss in the three periods of evolutionary history. Besides divergence on the whole genome level, the retention rates of syntenic regions explores the pattern of gene loss between the subgenomes in homologous chromosomes. This reveals gene retention difference not only in paralogous but also in orthologous genes.

Chapter 4 investigates the possible mechanisms of gene loss after speciation or polyploidization. The important question is about the mechanisms of fractionation in evolution. It is controversial that whether the loss of genes is a physical excision of part of the chromosomes or by pseudogenization, the loss of function and gradual randomization of the gene sequence. Chapter 5 presents a new approach to identify the relative dominance of the two mechanisms, as well as the application to several polyploidies in plants. I believe that the striking dominance of excision, both for fractionation and diploid gene loss is a novel biological discovery made possible by the methods introduced here.

Chapter 6 then studies 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 [22, 1, 29], 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, Section 6.2 develops an new exact recurrence to calculate the expected number of gaps of each length after a given number of steps. Section 6.3 then provide evidence from the *Coffea* data that demonstrates a systematic departure from this model. In a competing class of models [19, 30], 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.

In the concluding chapter, Chapter 7, we summarize the new approaches to fractionation introduced in Chapter 2 - 6, and the biological implications of the results on arabica coffee as well as other plant.

Chapter 2

The Evolution of Syntenic Homology from Gamma to the Present

2.1 Introduction

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 [10, 11]. This allows us to study the evolution of paralogous and orthologous synteny blocks. We study 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 four genomes, CC, CE, subCC and 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 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.

2.2 Methods

At a general level, we used the "peaks" method [20] to locate the three events that generate duplicate genomes in the evolution of CA: γ hexaploidization, CC/CE speciation and CA tetraploidization (which is effectively a speciation of CC/subCC and of CE/subCE). In this method, the local modal values (peaks) of the distribution of average similarities of the entire set of homologous gene pairs, as calculated by the R function geom_density using default parameters, are estimates of the time of the event. The use of average similarities in a synteny block allows much greater precision in retrieving the date of the evolution event, gamma hexaploidization, CE-CC speciation, CA tetraploidization.

2.3 The sequence of evolutionary events

Amalgamating the gene pairs in all SynMap comparisons produces the Circos plot in Figure 2.1. The Circos plot provides a visualization on the relationship between chromosomes, which marks the identical synteny blocks with lines. The high degree of parallelism of the connecting lines between the genomes reveals the highly similar retention between the homologous chromosomes in CA and between CC and CE.

2.4 The distributions of gene pair similarity

Gene pair similarities are universally understood to decay over time, like radioactive decay. However, there is considerable variability from one gene pair to another since mutation is a random process and there are only hundreds or a few thousand nucleotides in a gene that can mutate. Thus, the use of average similarities in a synteny block allows much greater precision in retrieving the date of the event responsible for the creation of the pairs in the block, reducing the variance of the estimate by at least 60% from that of individual genes. Among the synteny blocks from one of the 10 genome comparisons and self-comparisons we can find synteny blocks dating from as many as three different events clearly distinguished. Because of chromosomal rearrangement processes, only fragments of chromosomes are aligned, so usually several synteny blocks are detected all generated from the same genomic event, i.e., at the same point of time. That the average similarity levels in several synteny blocks are identical or almost so, is confirmatory of their origin in a common event.

All ten self- and pairwise comparisons show a cluster of homologous pairs dating from the early gamma hexaploidization of the core eudicots. Figure 2.2 depicts six of the distributions, two dating from the CC/CE speciation (CC vs CE, subCC

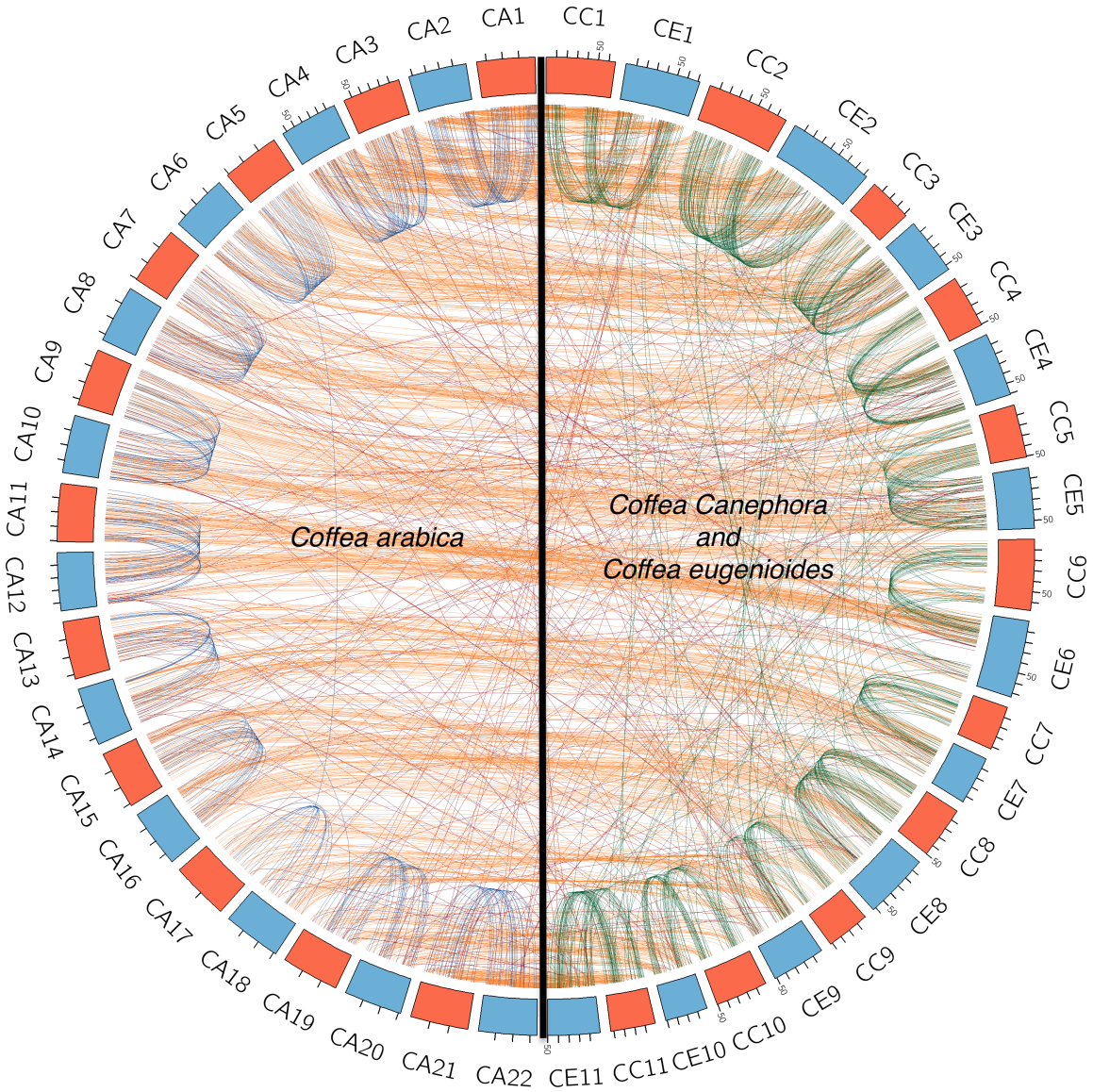


Figure 2.1: Circos plot of pairs of synteny blocks among *Coffea arabica* (CA) and its two diploid progenitors, *Coffea canephora* (CC) and *Coffea eugenioides* (CE), as well as paralogous pairs between homeologous chromosomes in CA. Note that subCC/subCE blocks are about twice as long on average (67 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.

vs subCE, top panel), two from the tetraploidization event (CC vs subCC, CE vs subCE, middle panel), and two from the γ event itself (CC vs CC, CE vs CE, bottom panel). Table 2.1 compares averages over all pairs of synteny blocks with less than 87% similarity, indicating tight clustering of these estimates, in terms of peak gene similarity and synonymous mutation rates (K_s), as well as the average length of these blocks. K_s is a preferred measure of mutation since is is counts only the mutations in "silent" positions in the coding sequence (e.g. most 3rd positions in anticodons).

Table 2.1: Locating the γ hexaploidy in all comparisons; peak of distribution of pairs with less than 87% similarity.

Comparison	Peak of similarity (%)	K_s	# of pairs	Average length
CC vs CE	73.7	2.42	2069	8.28
subCC vs CE	67.8	2.63	1860	8.05
subCE vs CC	68.0	2.60	2105	8.10
subCC vs subCE	67.9	2.69	1922	7.88
CC vs CC	71.5	2.50	1163	8.98
CE vs CE	70.4	2.88	1056	8.28
subCC vs subCC	68.2	2.64	938	8.81
subCE vs subCE	70.1	2.68	1043	8.21
subCC vs CC	68.2	2.71	1949	8.40
subCE vs CE	67.8	2.73	1925	7.89
Mean $\pm$ S.D.	69.4 $\pm$ 2.0	2.65 $\pm$ 0.13	1603 $\pm$484	8.29 $\pm$ 0.36

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 Figure 2.2. Table 2.2 presents the peak similarity and K_s for these comparisons.

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 Figure 2.2. These are depicted in Table 2.3.

Table 2.2: Locating the CC-CE speciation

Comparison	Peak of similarity (%)	K_s	# of pairs	Average length
CC vs CE	99.12	0.0261	17,066	37.84
subCC vs CE	99.08	0.0350	15,985	36.50
subCE vs CC	99.11	0.0307	16,014	63.30
subCC vs subCE	99.05	0.0335	15,318	66.03
Mean ± S.D.	99.09 ± 0.033	0.031 ± 0.0034	16,096 ± 626	50.92 ± 13.79

Table 2.3: Locating the tetraploidization event

Comparison	Peak of similarity (%)	K_s	# of pairs	Average length
CC vs subCC	99.81	0.0150	16,487	80.03
CE vs subCE	99.87	0.0180	17,196	33.26
Mean ± S.D.	99.84 ± 0.043	0.0165 ± 0.0021	16,842 ± 501	56.65 ± 33.07

The current best estimates of γ and CC/CE *Coffea* speciation are of the order of 120 My and 10 My [8], 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.

In addition to the synteny blocks relevant to γ, speciation and tetraploidization, we also noted dozens of synteny blocks of very high similarity level that seemingly indicated a productive process in segmental duplication, especially in CE and subCE. The duplicated segments in all chromosomes are almost all very close to the original (Figure 2.3). Detailed inspection of these regions, however, revealed that they are artifacts of incorrect assembly following the genome sequencing procedure. Informal examination of a range of other plant genomes showed that this artifact shows up frequently for sequences lacking adequate controls on assembly. In the case of CE, this could be traced to the paucity of sampled material from the original plant, which was eventually abandoned in a war zone in East Africa.

2.5 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 asymmetries in γ fractionation rates or evolutionary divergence rates of CC and CE or subgenome dominance play a role, their effects must be relatively small. Based on $17,006$ gene pairs and $21,000$ genes in the synteny block, independent gene loss from CC-CE synteny blocks since speciation, has been up to 25%.

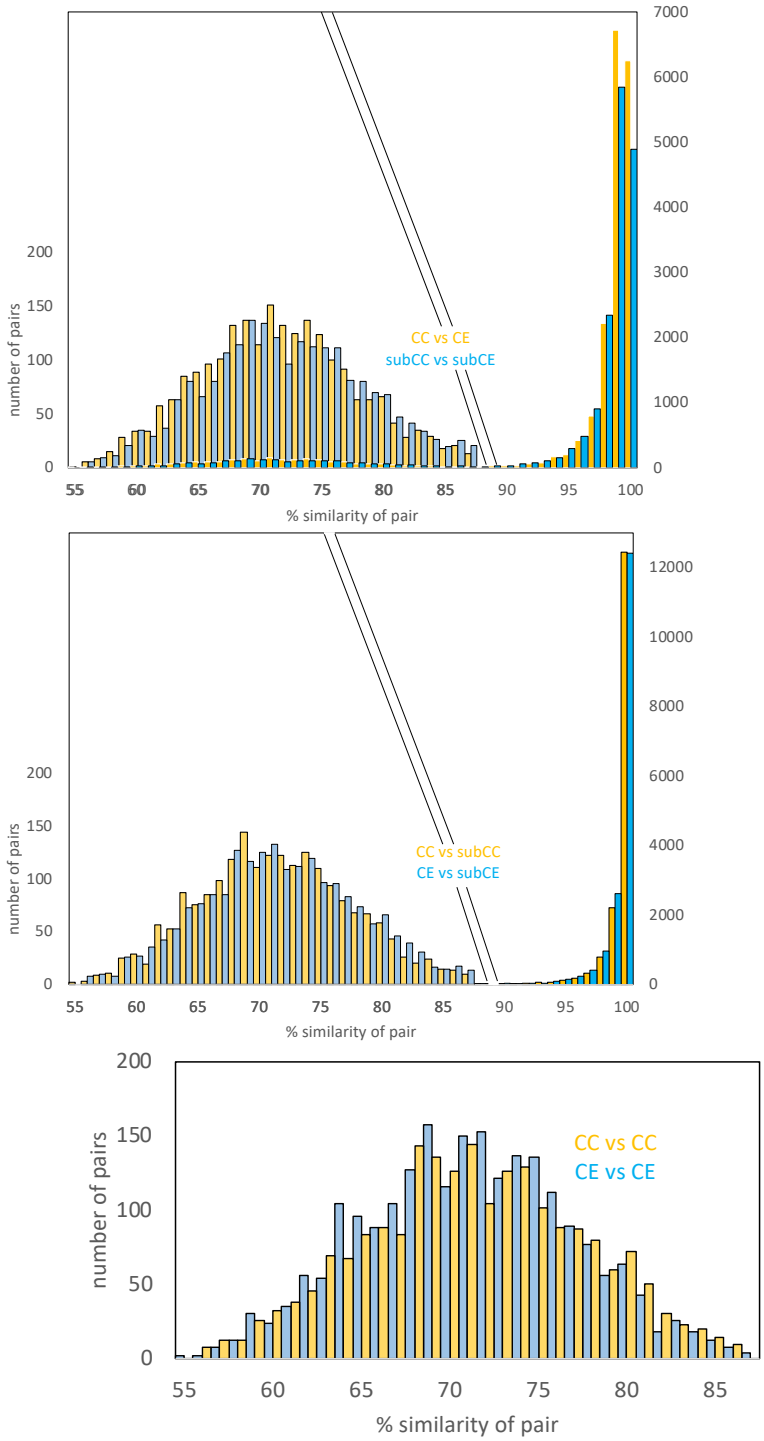


Figure 2.2: Gene pairs originating in speciation (top). Gene pairs originating in tetraploidization (middle). Gene pairs originating in γ event (bottom)

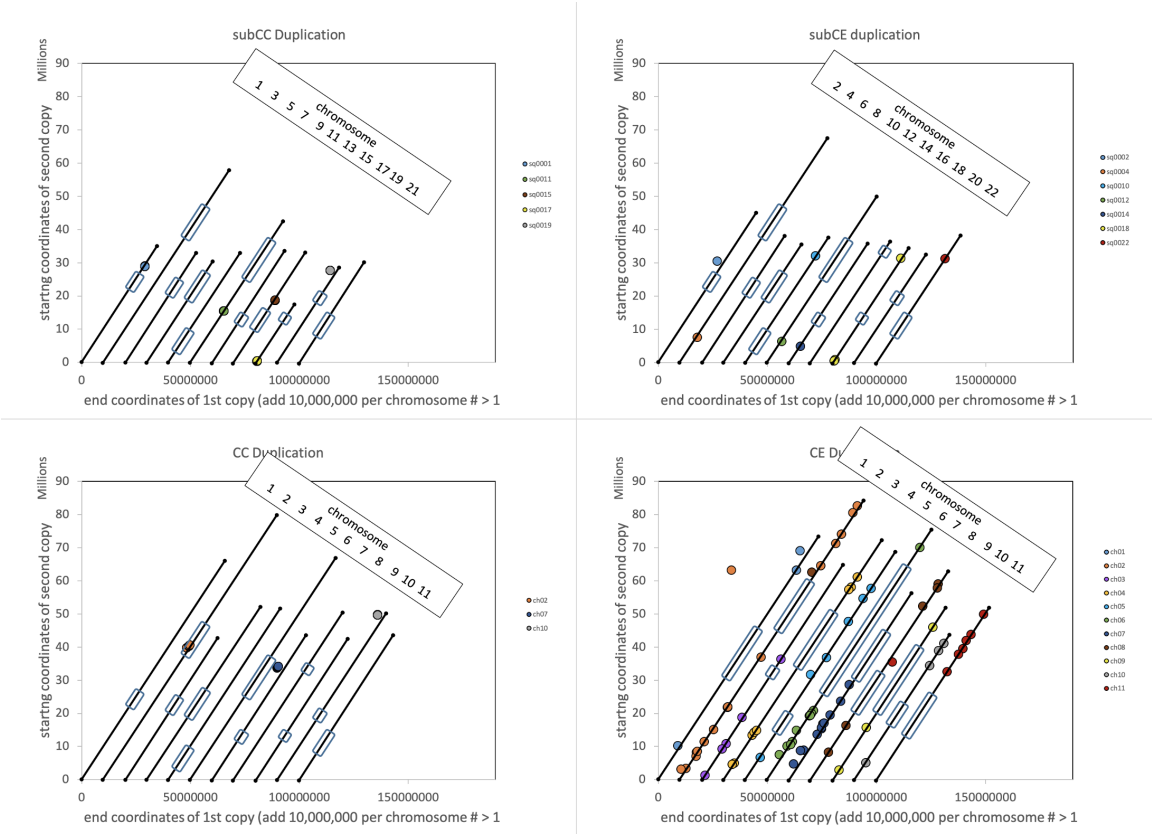


Figure 2.3: Segmental duplication in the chromosomes in subCC (top left), subCE (top right), CC (bottom left), and CE (bottom right). The rounded rectangles are approximate locations of pericentromeric regions, as based on some density graphs of centromeric retroposons

Chapter 3

Gap Distribution and Retention Rate

3.1 Introduction

Whole genome doubling can be considered a single event in evolutionary history that gives rise to duplicate copies of each chromosome – homeologs, containing duplicate copies of every gene, in the same order along the length of both chromosomes. Over time, typically many millions of years, most of the gene pairs – paralogs, in a plant lineage will eventually incur fractionation, loss of one member. Retaining two would be metabolically costly and could involve functional imbalances in the cell, while loss of both copies may not be viable. Thus a "record" of the loss of a gene is maintained by virtue of the retention of its paralog gene in the homeologous chromosome in the same context where the lost gene was originally situated.

This chapter studies a number of statistics on the gaps between adjacent pairs of duplicate genes within synteny blocks in all evolutionary eras, which is 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.

3.2 Fractionation in the *Coffea* ancestors

The self-comparison of CC and the self-comparison of CE, ignoring relatively recent events, are essentially two replicate assessments of the evolution of their common lineage through core eudicot, asterid, Gentianales (order), Rubiaceae (family) and *Coffea* (genus) ancestors.

The two genomes evidence almost identical numbers of gene pairs in the synteny blocks in their self-comparisons, as could be expected, since they are measuring essentially the fractionation process in the same ancestral lineage over 110 My as well as a small amount of additional loss over the recent 8 My of independent evolution. Moreover, the distribution of gap sizes is also almost identical (Figure 3.1). Note that the largest number of pairs are immediately adjacent, but there are still thousands of adjacent pairs interrupted by one or more genes with no homolog.

3.3 Independent gene losses and gains in *C. canephora* and *C. eugenioides*

Comparing the CC and CE genomes allows us to assess the independent losses of genes in the two species since speciation (Figure 3.2). In addition, the CC self-comparison and the CE self-comparison (Figure 3.3) provides a window on any recent events in either of the genomes.

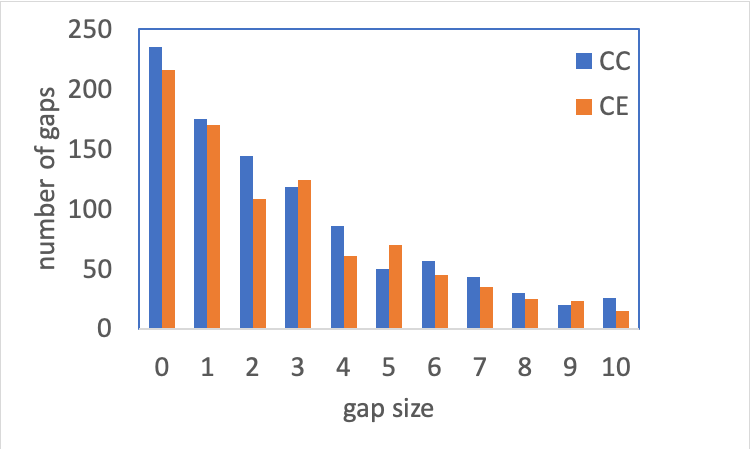


Figure 3.1: With the number 255 of synteny blocks, distribution of gap sizes within synteny blocks generated by the γ hexaploidy, as revealed by the self-comparison of CC and CE. Gap size = 0 indicates adjacent duplicate pairs with no singleton (non-duplicate) genes between them in either genome. Other gap sizes indicate number of intervening non-duplicate genes interrupting neighbouring duplicate pairs.

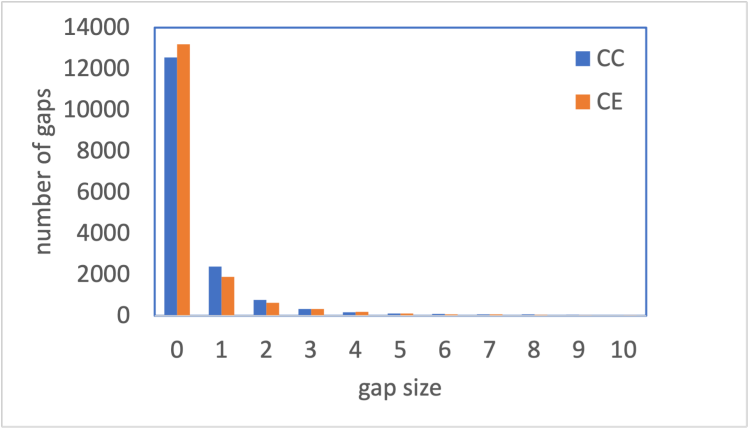


Figure 3.2: Distribution of gap sizes within synteny blocks generated by speciation in comparison between CC and CE. The "CC" values are the number of duplicate pairs whose members are adjacent in CC, but have the indicated (X-axis) gap size in CE. The "CE" values are the number of duplicate pairs whose members are adjacent in CE, but have the indicated gap size in CC. Note that there are far more gene pairs surviving in the 8-10 million years since speciation that in the 120 million years since γ

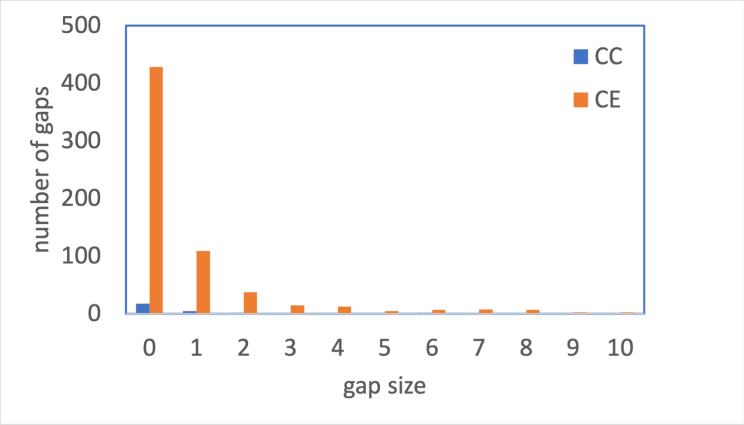


Figure 3.3: Distribution of gap sizes within synteny blocks generated by recent events in CE and CC self-comparisons. The "CC" values are the number of duplicate pairs whose members are adjacent in CC, but have the indicated (X-axis) gap size in CE. The "CE" values are the number of duplicate pairs whose members are adjacent in CE, but have the indicated gap size in CC.

Figure 3.2 shows that there are somewhat more $(13,190 - 12,551 = 639)$, gene pairs with adjacent genes in CE that also have adjacent genes in CC, than gene pairs with adjacent genes in CC that also have adjacent genes in CE. In other words, more genes have been spontaneously lost from CC – not by fractionation since the genomes are in different species – or else inserted into CE, more likely, given our previous observation of recent increases in syntenic paralogy in CE (Figure 2.3).

Figure 3.3 tells a similar story, more dramatically, since CE, compared to itself, contains over a thousand recent gene pairs adjacent, or almost adjacent, to each other in synteny blocks. As we have seen, this is likely artifactual. Indeed is method here becomes more of a way of inferring errors in the assembly laboratory than a way of studying very recent evolution.

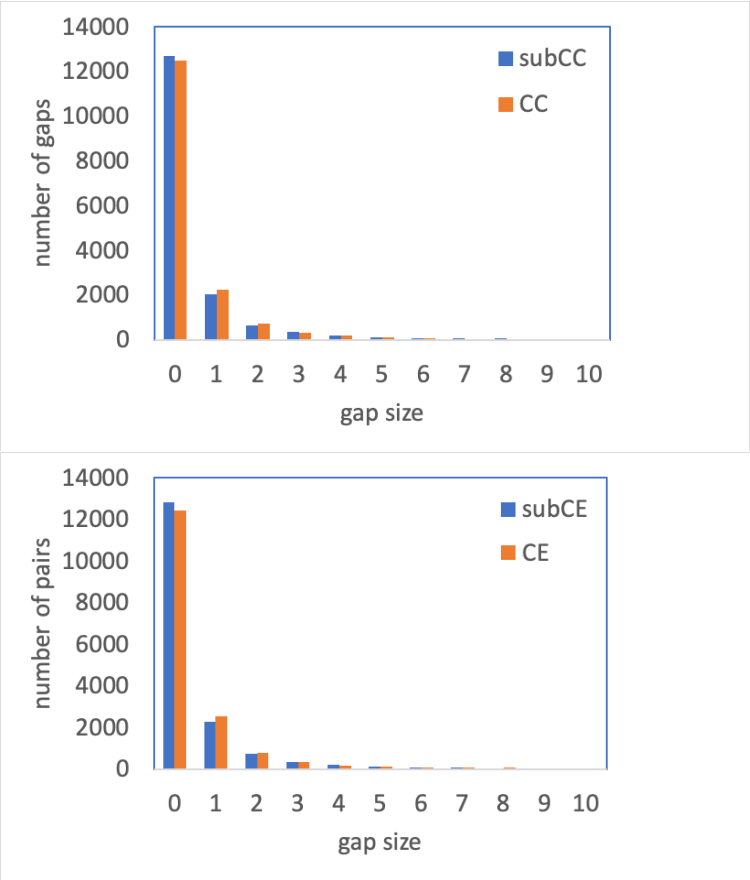


Figure 3.4: Distribution of gap sizes within synteny blocks generated by the CA tetraploidization event in comparison of subCC vs CC (top) and comparison between subCE vs CE (bottom).

3.4 Fractionation in *C. arabica*

In Figure 3.4, the comparison of the CC vs subCC genomes and the comparison of the CE vs subCE genomes provide the clearest view of fractionation since the tetraploidization event. In both diagrams, adjacent genes in the subgenome are interrupted more frequently by non-duplicate genes in the parent genome. There are 211 genes (1.4 %) missing from subCC synteny blocks when compared to CC and (2.7%) missing from subCE blocks when compared to CE.

Even if the distribution of gap size in the comparison of CA subgenomes are almost identical (Figure 3.4), Figure 3.5 shows the divergence in one chromosome of subCC in comparison of the referred chromosome of subCE. Gene loss events occurs more on the tail of the chromosome. The breakpoints on the graph are affected by the region of centromere. The difference of gene retention rates suggests gene fractionation events during tetrapoidization. More retention graphs are shown in the Appendix A.

3.5 Conclusions

Within synteny blocks, only about a quarter of the neighbouring gene pairs generated by γ have no interrupting non-duplicates between them. This suggests that most of the paralogs have been fractionated.

The distribution of gap sizes in synteny 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, although the occasional loss of two or more adjacent gene cannot be ruled out.

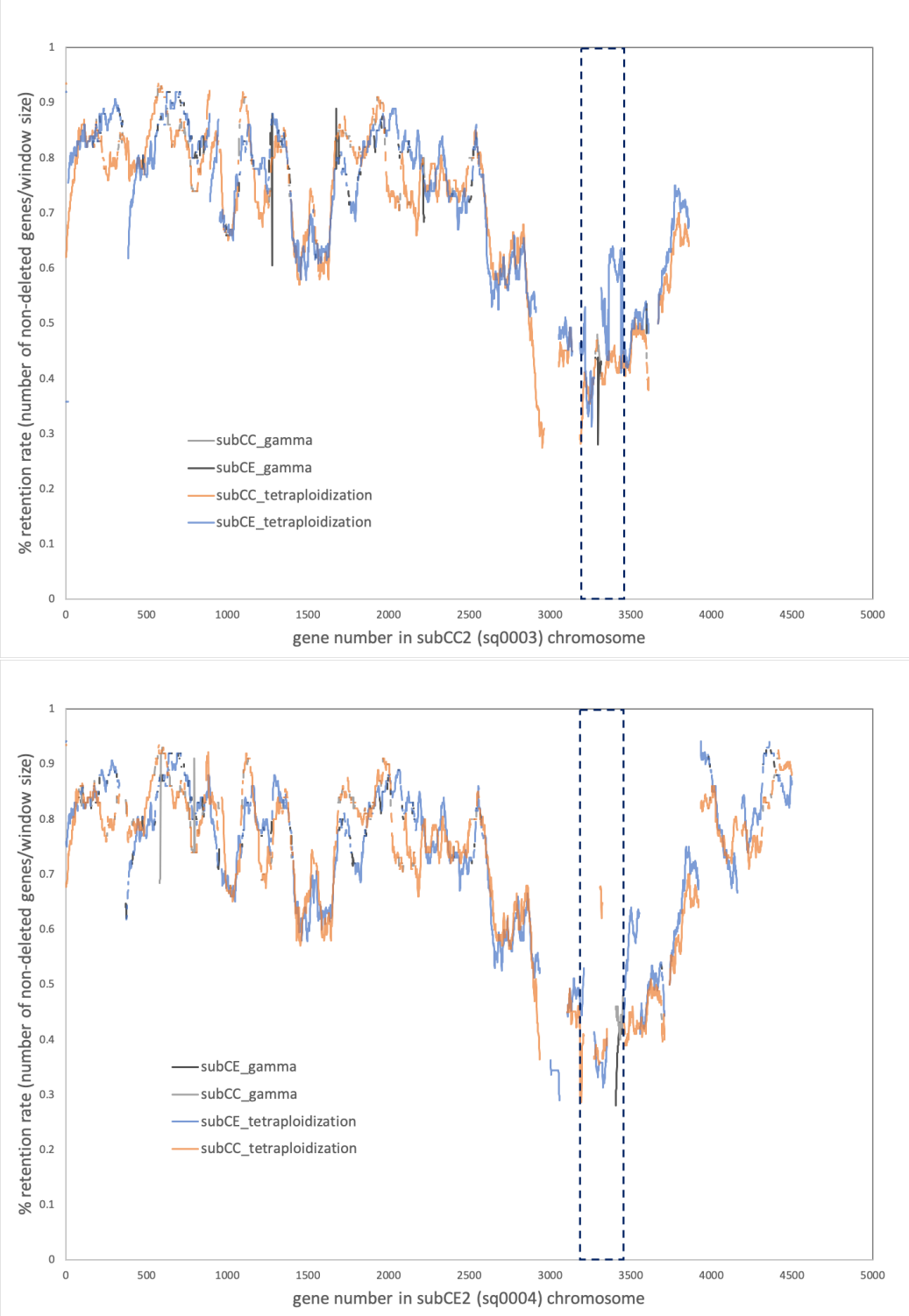


Figure 3.5: Retention rate in syntenic region in comparison of subCC and subCE, arranged on target subCC chromosome 2 (top), and its homologous chromosome 2 of subCE (bottom). Centromere is marked in dashed rectangle.

Chapter 4

Mechanisms of Gene Loss

4.1 Introduction

The alignment of the gene orders of homologous genes in two related genomes, or subgenomes of an (ancient) polyploid, such as that provided by the SynMap program on the CoGe platform [10, 11], is a uniquely reliable first step in the assessment of gene conservation or loss after speciation or polyploidization. The homology of pairs of genes in the chromosomal fragments "synteny blocks" making up such an alignment, is doubly confirmed, first by the common level of sequence similarity of all the gene pairs in the block, and second by the common chromosomal context, namely the common order of the homologous genes in the two fragments, represented as follows:

Synteny block on homeologous regions of two chromosomes.
Dark circles indicate retained genes, white circles deleted genes.
There are five retained duplicate gene pairs, four singletons on the lower chromosome and one singleton on the upper chromosome.

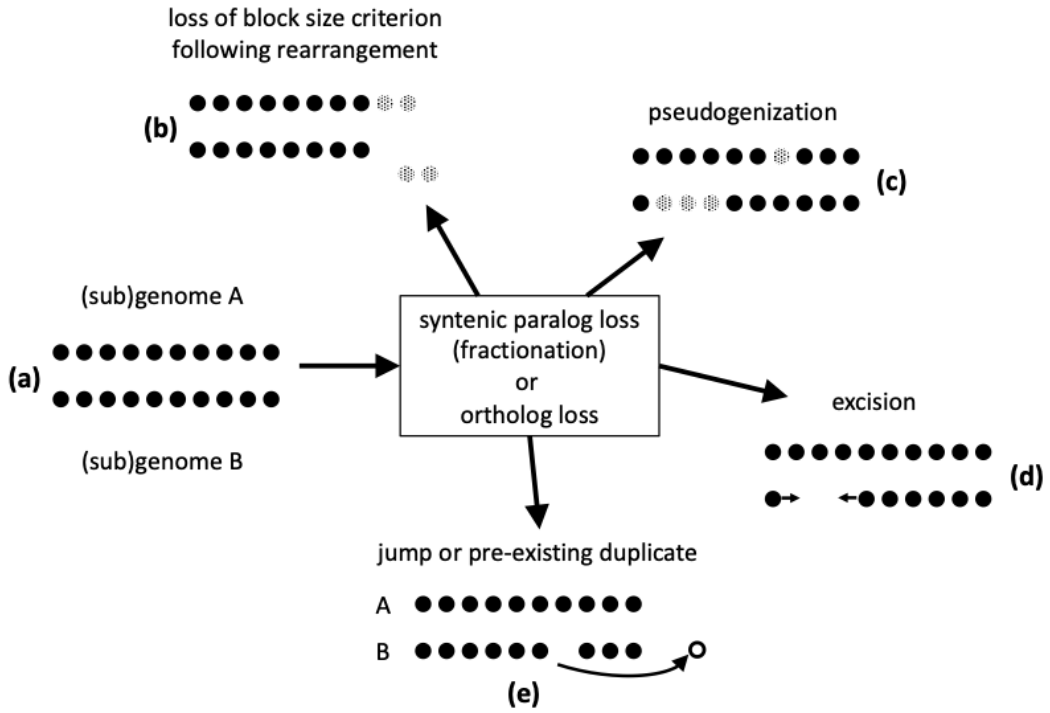


Figure 4.1: Mechanisms of loss of gene pair homology after polyploidization. (a) synteny block made up of co-linear homologous pairs. (b) erosion of synteny block by translocation to a remote chromosomal location of a portion of sub-threshold length. (c) Pseudogenization. Genes rendered inoperable represented by grey dots. (d) Excision of DNA fragment including one or more genes. Arrows represent a new adjacency after the loss of the excised genes. (e) Jump of one member of pair to a different genomic location, or loss of only one of two or more homologs of the same gene.

The stringent criteria, such as a minimum number of contiguous pairs, incorporated in algorithms such as SynMap [12] for constructing synteny blocks generated by speciation or polyploidization tends to excludes some of the homologous gene pairs created by these genomic events (represented in Figure 4.1a), especially after some time has elapsed. Inversions, translocations and other chromosomal rearrangement events in a genome or in either of two related genomes, break synteny blocks into

smaller pieces that may not satisfy the criteria, as illustrated in Figure 4.1b and e.

We have assessed the effect of the default SynMap requirement used in all other chapters in this thesis – at least five closely spaced gene pairs for a synteny block to be identified – by increasing and decreasing this threshold – see Figure 4.2. A slight decrease in the number of genes in blocks when the threshold is increased to 6 is simply due to the elimination of a few blocks of length 5. But as we decrease the threshold to 3, the algorithm starts to capture blocks made up of independently created but coincidentally neighbouring pairs, as well as pairs where one member is already in a larger block, since a gene can be in more than one block. It becomes increasingly difficult to disentangle the behaviour of duplicate gene pairs created by polyploidization from other processes of duplication and loss. Thus we retained the default value, 5.

4.2 Fractionation events

In a synteny block, not all genes are paired. Among other mechanisms, fractionation may remove one gene of a duplicate pair, as illustrated in Figures 4.1(c),(d) and (e). In fact there are 7461 subCC genes in synteny blocks that do not have a subCE counterpart in that synteny block, and 7691 subCE genes in synteny blocks that do not have a subCC counterpart in that synteny block. Part of this is due to the minimum block size (5) criterion of SynMap. If this size is reduced to 2 or 1, the number of unpaired subCC genes drops by 1190 or 2120, and the number of unpaired subCE genes drops by 1240 or 2179. Of course, the validity of the homology of the additional pairs under the relaxed synteny criteria is subject to considerable uncertainty.

Unexpectedly, most of these genes missing from a subgenome were not lost after

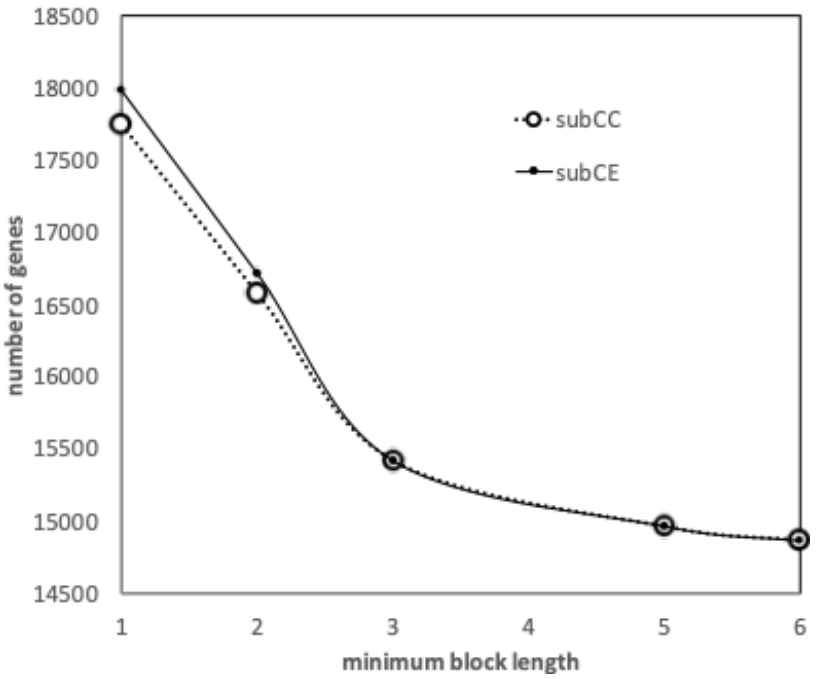


Figure 4.2: Effect of minimum block size on the number of genes incorporated into synteny blocks. Differences in the two genomes stem from gene membership in two or more synteny blocks.

tetraploidization. The great majority were already absent from the expected syntenic position in the parent of this subgenome. Only 835 of the missing subCE genes can be found paired in a subCC-CE syntenic block, and only 950 of the missing subCC genes can be found paired in a subCE-CC syntenic block. Only these are likely to have undergone fractionation after tetraploidization. Thus, 89% of the missing subCE genes were already missing before tetraploidization, as were 88% of the missing subCC genes.

Polyploidization is only one of the processes giving rise to paralogous pairs of genes. Tandem duplicates and other mechanisms endow genomes with many duplicate pairs, some of which may be almost contemporaneous with the genome-wide event under study. Thus, the same gene g may be present in two or more synteny blocks, since it may be homologous to two tandemly-originated copies of a single gene. After fractionation of one of these copies, gene g will be matched in one of the synteny blocks and not the other. This is one of the cases represented by Figure 4.1(e). Thus, aside from the 7461 unmatched subCC genes and 7691 unmatched subCE genes, another 1697 subCC genes and 1580 subCE genes, respectively, were unmatched in one synteny block with the other subgenome but matched in another synteny block (and counted as matched). Most of these, 1567 in subCC and 1479 in subCE, were matched in synteny blocks with CE and CC, respectively, almost all to only one gene. The remaining 130 and 101 genes, unmatched in any synteny block with a parent genome, may have jumped to their current synteny block from some other place in the genome, as in Figure 4.1(e), but this is speculative, lacking other evidence. More importantly 1233 of the 1567 subCC genes and 1242 of the 1479 subCE genes are in matched in exactly one of two synteny blocks with CE and CC respectively, so that their partially fractionated status in the tetraploid is inherited from the parent.

The remaining 324 and 237 genes, matched in only one synteny block with a parent genome, may also be candidates for jump translocation status (Figure 4.1(e)), but this also remains speculation.

4.3 Conclusions

Most of the apparent fractionation in *C. arabica* reflects the difference in gene complement between the parent genomes *C. canephora* and *C. eugenioides* prior to the tetraploidization event.

A substantial proportion of duplicate gene loss from synteny blocks, whether orthologous or paralogous, leaves alternative homological relations in other synteny blocks. The surviving gene originally had multiple homologs.

We continue the further analysis between pseudogenization and excision, Figure 4.1c and d, in the next Chapter 5

Chapter 5

Excision dominates pseudogenization in gene loss

5.1 Introduction

A longstanding biological controversy in evolutionary genomics [22, 1] involves 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, which we have termed the "structural" mechanism, or through targeted (possibly) gene-by gene events such as regulatory epigenetic silencing and pseudogenization, which we call "functional" mechanisms. Because it is often difficult to ascertain whether a single-copy gene is the result of the deletion of a duplicate copy, and because the outcomes of the two kinds of process may appear similar, it is often difficult to discern which one is operating.

Although the basics of polyploidy in plants have been understood for over a century [23], and though this process is well attested across the entire evolutionary

spectrum, from bacteria [7, 21] to pre-mammalian vertebrates [14], the statistical study of conservation and reduction at the genome level originates with the discovery and analysis by Wolfe and Shields of an ancient WGD in the *Saccharomyces cerevisiae* genome sequence [24]. But starting with the first few plant genomes to be sequenced – *Arabidopsis, Oryza, Populus* – the realization has grown that all flowering plants species (except one, *Amborella trichopoda*) are "paleopolyploids", re-diploidized descendants of one or more ancient polyploidization events. It is in the context of the Angiosperm/Magnoliophyte phylum or division that we have attempted to resolve the structure-function controversy [22, 1] using several modeling and statistical approaches [29, 16].

In the present chapter, however, our focus is less on how fractionated gene pairs are organized within synteny blocks, than on what happens to these genes – do they degenerate in place, or are they simply removed from the DNA sequence of the genome?

Our claim is that the overwhelming loss process is the latter: the complete excision of the gene from the genome, the elimination of the sequence of the entire gene. As such, we do not adopt any restrictive definition of a pseudogene or quantification of the various types of pseudogenes in plants, which was done in the recent definitive study of Xie *et al* [28]; here we simply examine whether any DNA, and how much, remains, when a one member of a pair of homeologous genes, as identified by SynMap, is absent from a syntenic block. We will show that in the large majority of cases, there is a drastic loss of DNA, leaving only a small stretch of intergenic sequence, so that no kind of pseudogene, whatever its definition, except for very small fragments of cDNA, can be present. In other words, fractionation, and most gene loss in ancient genomes, does not tend to result in long-lasting full length or part

length degenerate genes, but a relatively complete loss of the DNA. This does not mean that pseudogenes are absent or even rare in these and other genomes. Many of these may persist over many millions of years. Nevertheless, Xie et al. (The Plant Cell 31: 563–578, 2019) found that poplar has almost 25,000 pseudogenes, but less than 1500 of these stem from the Salix whole genome doubling, and most of these are presumably small fragments of coding sequence.

5.2 Methods

5.2.1 Sampling of plant species

In each of four core eudicot plant families, we selected a pair of genomes for which annotated genome sequences are available:

1. *Populus trichocarpa* (poplar) CoGe ID 25127, and *Salix purpurea* (willow) CoGe ID 52439 in the rosid family Salicaceae,

2. *Salvia splendens* (scarlet sage) CoGe ID 55705 and *Tectona grandis* (teak) CoGe ID 55706 in the asterid family Lamiaceae,

3. *Linum usitatissimum* (flax) CoGe ID and *Hevea brasiliensis* (rubber tree) CoGe ID 16772 in the family Linnaceae, also rosids, and

4. *Malus domestica* (apple) CoGe ID 54783 and *Pyrus* × *bretschneideri* (pear) CoGe ID 37224 belonging to the same subtribe Malinae of another rosid family Rosaceae.

All these genomes have undergone at least one whole genome duplication since the ancient whole genome triplication "gamma (γ)" at the origin of the core eudicots.

5.2.2 Construction of synteny blocks

For each of the eight genomes individually we first carried out a self-comparison of the unmasked sequences using the SynMap program on the CoGe platform [10, 11] to construct paralogous syntenic blocks. Based on the distribution of gene pair similarities, also output by SynMap, we retained only those blocks for which the average similarity confirmed that the duplication occurred at the time of the most recent polyploidization event experienced by the genome.

For each of the four family pairs of genomes, we then used SynMap to compare the two and construct orthologous synteny blocks. We again referred to the distribution of gene pair similarities in selecting only those blocks likely to have been created at the time of the speciation event at the origin of the diverging lineages leading to the two species being studied. We thus aimed to exclude synteny blocks created by polyploidization in the common ancestor of the two, including the gamma triplication, as well as blocks created in either of the two genomes by post-speciation polyploid events.

Since we will be focusing on pseudogenization and excision in our analysis, Figure 4.1c and d, we developed a method that does not favour the identification of one in favour of the other.

5.2.3 Identification of deletion intervals and their lengths

We scanned the output of the retained synteny blocks for homeologous segments on two chromosomes (or two disjoint regions of one chromosome) bounded by one or (usually) more duplicate gene pairs at both ends, where all the genes in one segment – the fractionated side – are absent, i.e., not detected by SynMap. (No gene can be

absent from the other segment – otherwise the ancient gene pair, if it ever existed, would not be visible.) We call the number of contiguous single-copy genes in the unfractionated side of the segment the *length* of the interval. This is the same as the number of genes that are missing from the fractionated side.

For both sides of the segment, we also determine the amount of DNA between the pairs that bound the segment. For the unfractionated side, with all the single-copy genes, this is just the size (in base pairs) of the genes plus the intergenic regions, including the initial region, after one bounding pair, and the final region, before the other bounding pair, in the segment. In the fractionated side, this includes whatever DNA remains between the two bounding pairs, which does not include any genes, according to SynMap.

Two possibilities are represented by Figures 4.1(c) and 4.1 (d). In the former case, pseudogenization, a gene is rendered inoperable, such as by a point mutation that creates a stop codon inside an erstwhile coding region. In the latter, a chromosomal fragment containing one or more genes is simply physically excised. To assess which of these two processes accounts for the data, we note that pseudogenization through acquiring a gene-internal stop codon, or a frameshift, leaving the gene intact, at least initially, does not shorten the length of the chromosomal region it is in. The average length of a pseudogene is roughly half of that of a functional gene [28], but this includes the very numerous short fragments. In contrast, excision of genes, including some or all of the flanking intergenic DNA, will definitely shorten the region, leaving at most a short stretch of non-coding sequence.

5.2.4 The visualization of gene density and pseudogene density

By plotting the average number of base-pairs in the unfractionated, or totally conserved, intervals of a given length against the number of singletons in the interval, we estimate the average size of a gene (plus the following intergenic region). In most cases we expect this plot to be approximately linear, with slope giving the average base-pairs per gene. This is just the inverse of the gene density for that interval. For the fractionated, or totally reduced, side, the number of base pairs per missing gene provides an upper limit (via its inverse) on the number of full-length pseudogenes that may be in the interval. Although most pseudogene tools were developed in the context of human or vertebrate genomes, and have limited applicability for plant genomes [27], Xie *et al* have succeeded in implementing PseudoPipe [31] for surveying pseudogenes in a range of plant species, and their results will be seen to be consistent with ours in the analyses below. PseudoPipe is a method for for detecting and categorizing pseudogenes in a genome.

5.3 Results

5.3.1 Willow and poplar

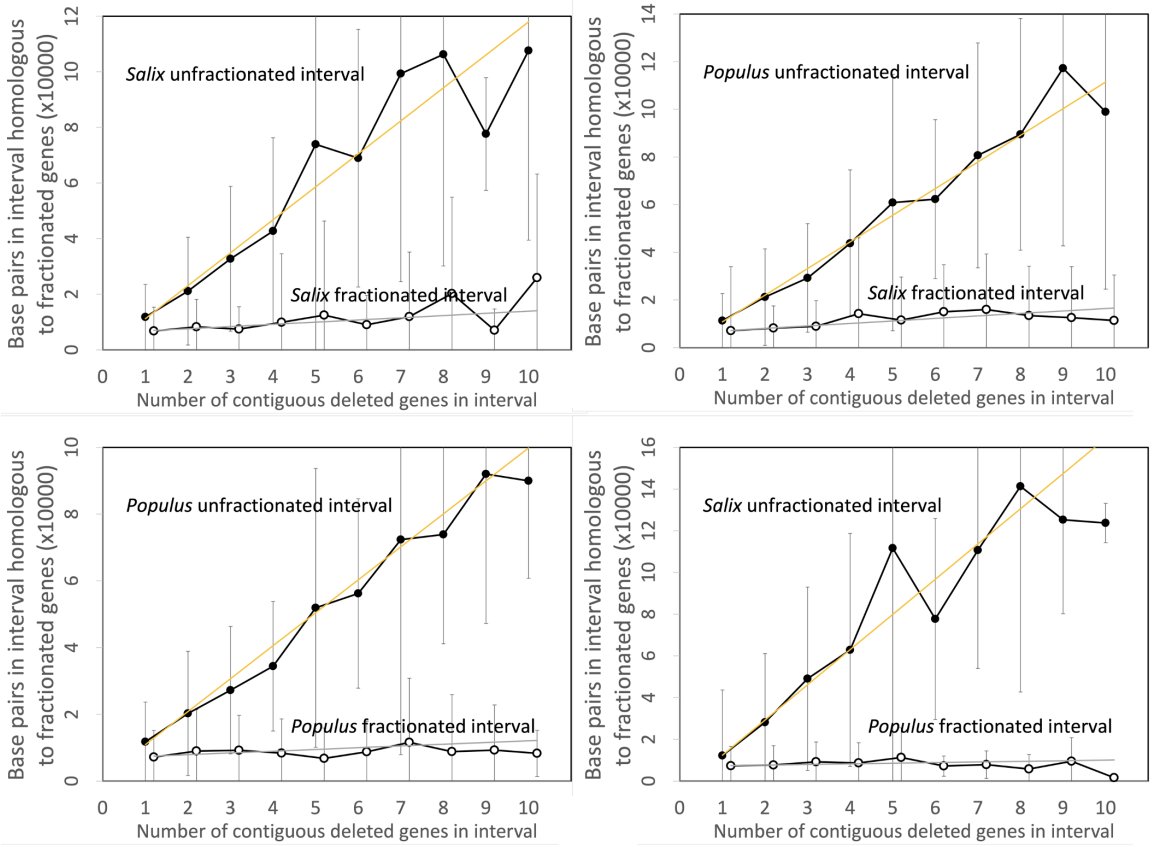


Figure 5.1: Comparison of DNA content in unfractionated and fractionated intervals in the *Salix* and *Populus* genomes. Linear regression fits are indicated. Self-comparisons on the left do not distinguish between subgenomes since these are hard to identify across chromosomes and may are generally mingled due to interchromosomal rearrangements, such as reciprocal translocation and chromosome fission and fusion. The two comparisons between genomes on the right hand side analyze gene loss from each genome separately. We use the terms "fractionated" and "unfractionated" in these two panels to mean "reduced" and "conserved", even though the polyploidization-induced fractionation does not play a role here. Whiskers represent standard deviation. For all values of gap size on the x-axis, sample size is ≥ 10, but for small gaps, sample sizes are in the hundreds or thousands.

Figure 5.1 contains the results of our analysis of the *Salix* and *Populus* genomes. The two panels on the left show the expected approximate linear growth in the number of base pairs in the unfractionated side of the interval. The great variability of the individual regions simply reflects the inhomogeneity of gene density along the length of the chromosome. In contrast, the regions in both *Salix* and *Populus* that have lost annotated genes show zero growth, with relatively little variability, as a function of the number of missing genes; they have lost almost all their DNA sequence. There cannot be significant numbers of pseudogenes, full or reduced, or other relics of the missing genes. This is striking evidence in favour of the predominance of excision. Similar conclusion for gene loss between species, showing by the two panels on the right.

5.3.2 Salvia and teak

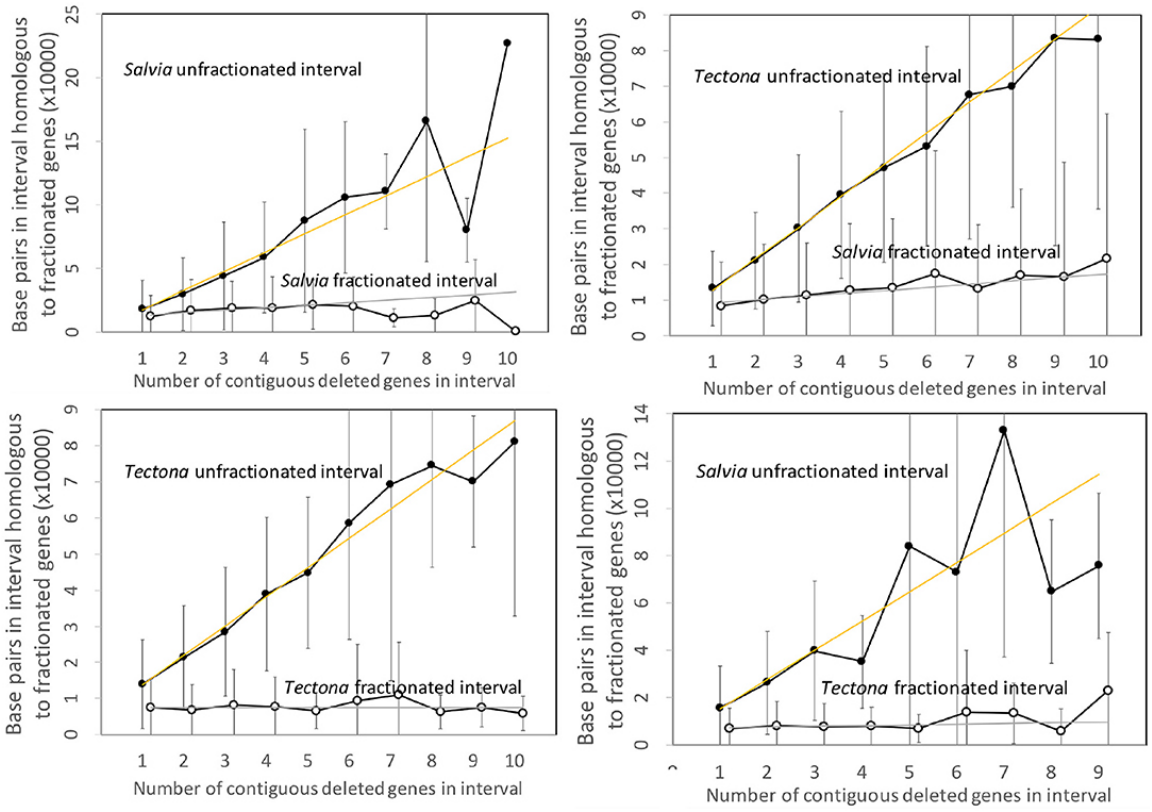


Figure 5.2: Comparison of DNA content in unfractionated and fractionated intervals in the *Salvia* and *Tectona* genomes.

Figure 5.2 contains the results of the corresponding analysis of the *Salvia* and *Tectona* genomes.

The panels in Figure 5.2 are very similar to those from the Salicaceae (Figure 5.1). Some of the curves show great fluctuation of the values for the longer intervals, but this is likely due to smaller sample size. Of interest is that the DNA content of the fractionated (read: "reduced") salvia intervals formed after speciation show a small but steady increase, but still orders of magnitude less than the sizes of the

unfractionated ("conserved") intervals (top-right panel in Figure 5.2).

5.3.3 Flax and rubber

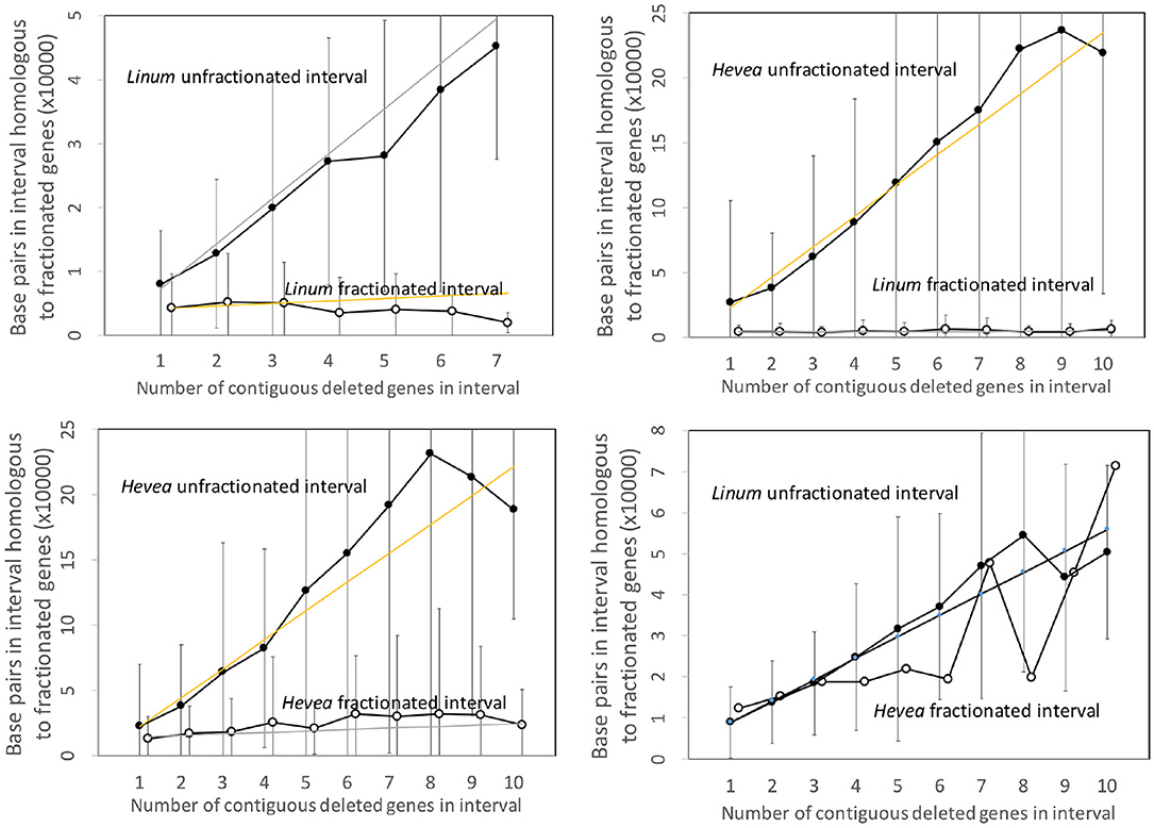


Figure 5.3: Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Linum* and *Hevea* genomes.

Figure 5.3 repeats the same analysis, this time applied to the *Linum* and *Hevea* genomes. The results parallel those of the two other pairs of genomes, except for the apparently anomalous behaviour of the *Hevea* intervals, where the number of base pairs attains the same level as the conserved genes in *Linum* (bottom-right panel in Figure 5.3). This, however, may be seen as an artifact of the disproportionately large

genome of *Hevea* with respect to that of *Linum*. The intergenic space in *Hevea* is four or five times as great as that of *Linum*, and there is much scope for retention or acquisition of repetitive elements and other sequence over the long period since the speciation event, which occurred much earlier than the other events we study.

To put this disproportion in perspective, we can normalize the *Hevea* results by a factor which measures the difference in sizes of the two genomes. This produces the comparisons in Figure 5.4, which better resembles those of the Salicaceae and Lamiaceae.

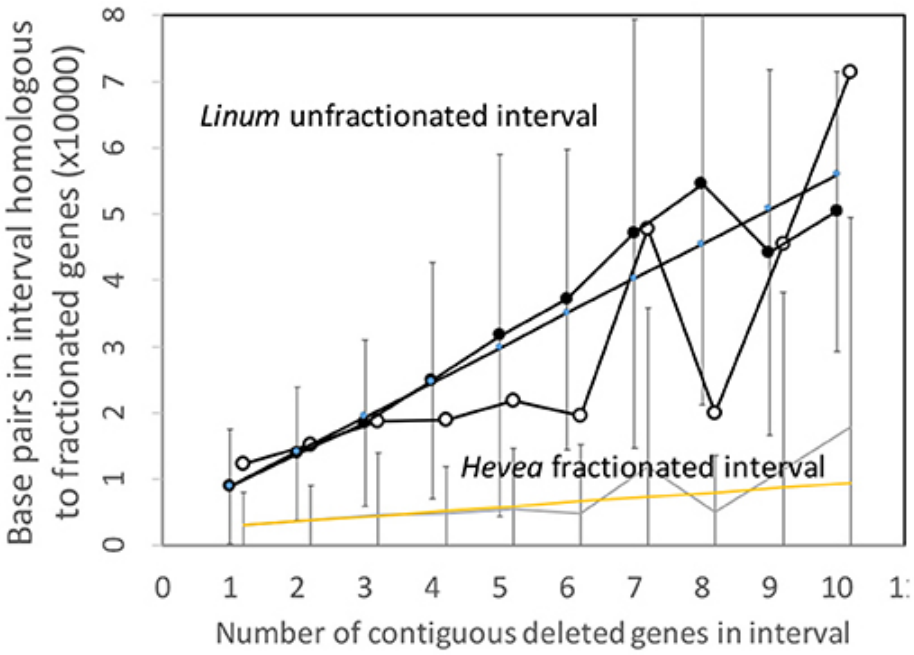


Figure 5.4: Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Linum* and *Hevea* genomes.

5.3.4 Pear and apple

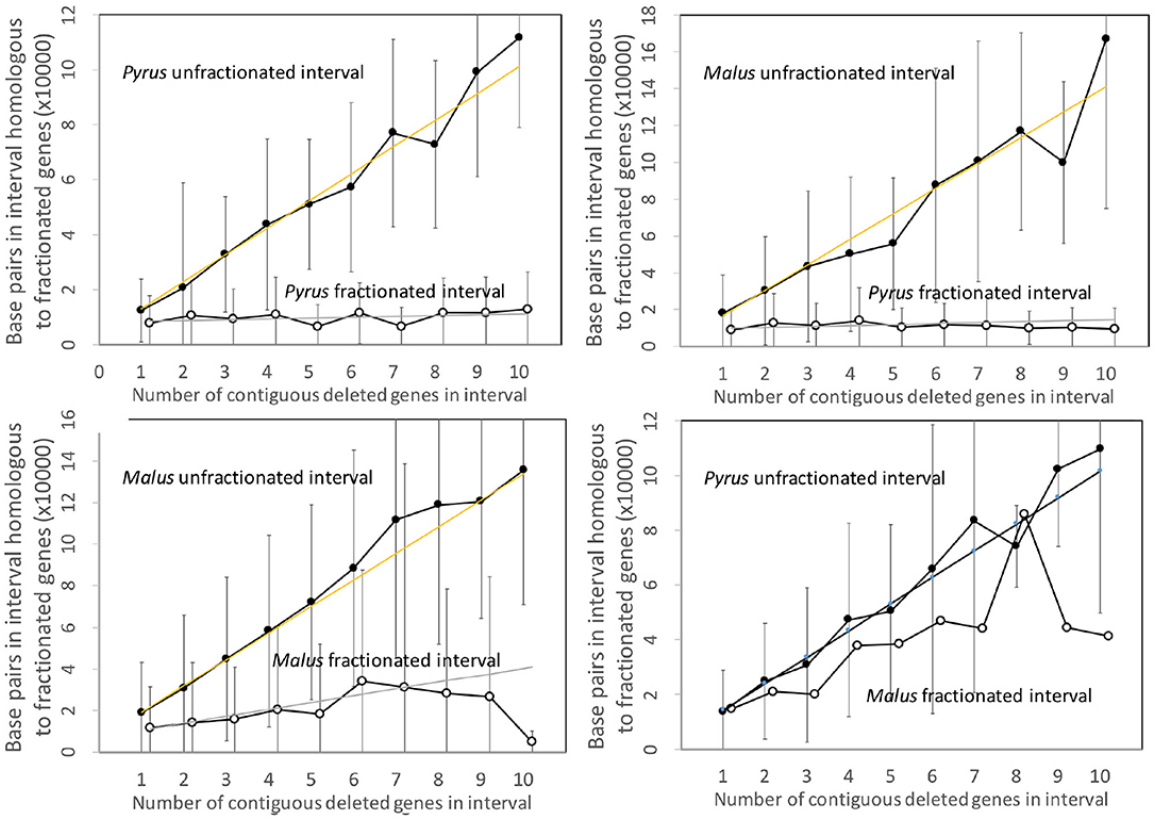


Figure 5.5: Comparison of DNA content in unfractionated and fractionated (conserved and reduced) intervals in the *Pyrus* and *Malus* genomes.

Figure 5.5 shows the analysis of the *Pyrus* and *Malus* genomes.

Here again, we have an anomalous large amount of DNA in the *Malus* reduced gene intervals after speciation. It is true that the *Malus* genome is larger than *Pyrus*, but explaining this through normalization (Figure 5.6) is not completely satisfactory. This is the only trend out of the thirty-two we have presented that departs from our main narrative.

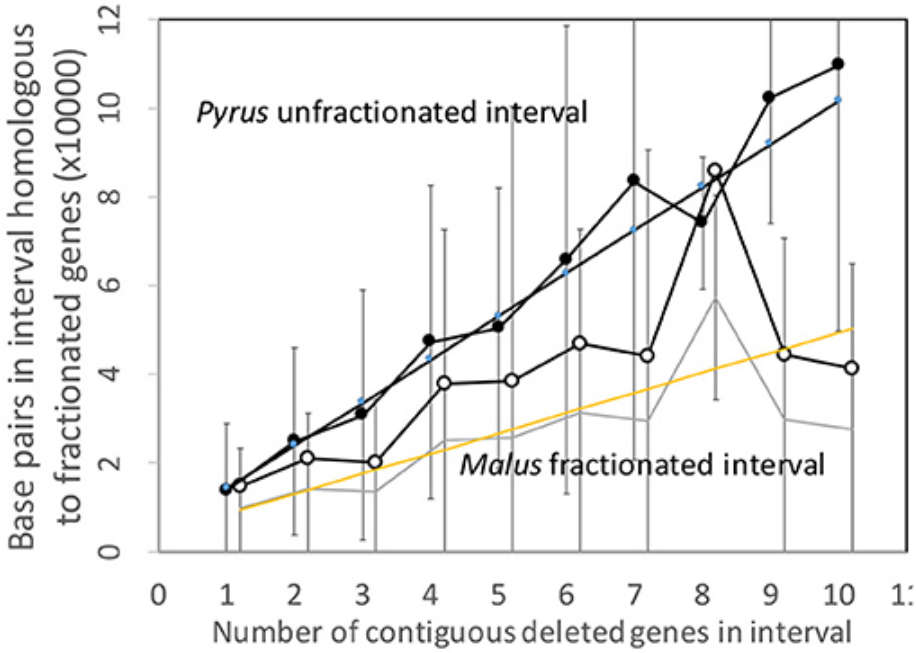


Figure 5.6: Comparison of DNA content in conserved and reduced intervals in the *Pyrus* and *Malus* genomes.

5.3.5 Comparisons across families

To compare the results from the four families, we must take into account the diverse genome sizes, number of genes in a genome, and the resulting gene densities. Figure 5.7 shows that gene density (or rather its inverse: base pairs per length of conserved fragment) in unfractionated and conserved intervals closely tracks the average gene density (or its inverse) for the entire genome. At the same time, the residual sequence length in intervals where fractionation or gene loss has taken place is not sensitive to gene density, it remains very close to zero, as expected from an excision explanation.

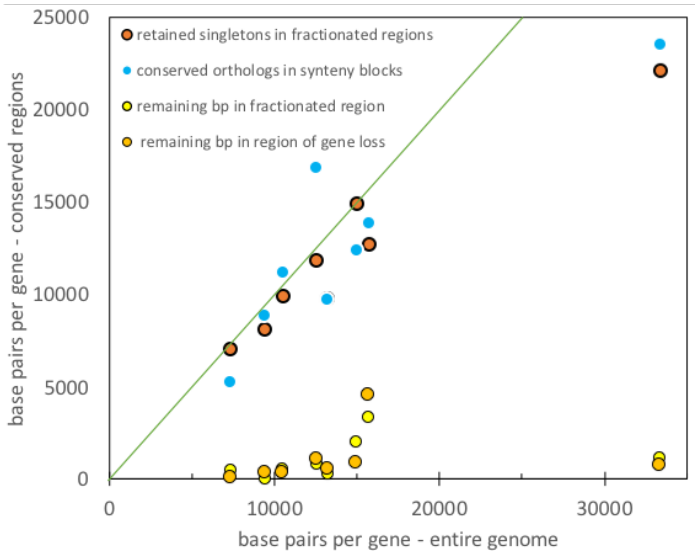


Figure 5.7: Comparison of gene density in unfractionated regions and the whole genome. Diagonal represents equality between the two densities.

We can also report, although it seems superfluous after examining Figures 5.1 to 5.5 and Figure 5.7, that t-tests confirm at a very high level of significance that the slopes of the two regressions in each panel are different.

5.3.6 Occurrences of gene translocation

To exclude other explanations of our syntenic block data, such as that in Figure 4.1e, we looked further into the fate of the fractionated genes in the *Populus-Salix* comparison. By setting the minimum block size to 1 in the SynMap self-comparison, we could detect all pairs of gene duplicates, not only those in synteny blocks. We then searched for pairs to the singletons identified in the original (default 5) construction of synteny blocks that we analyzed in Section 5.3.1 above. Out of 8307 *Salix* singletons, we found only 429, or 5%, that were paired else where in the genome at approximately the expected similarity level. Of the 10737 *Populus* singletons, only 742, or 7%, were

paired elsewhere. Moreover, some or most of the pairs that were identified could have been distinct paralogs that were part of a pre-existing triplet before fractionation – such triplets or higher sets of paralogs are not uncommon. We can conclude that translocation as an alternative explanation to excision can account for only a very small fraction of the gaps in synteny blocks.

There remains the possibility that if the missing genes did not translocate out of the synteny block, the singletons may have migrated in, after the polyploidization or speciation event. The main mechanism for this would be retrotransposition. However, retroposons are generally not annotated as genes in the CoGe database, even in the unmasked genome sequences we studied, and thus would not show up as singletons. Neither are many of the singletons likely to be translocated genes: a large proportion of genes in these genomes are paired, and an equal proportion of the putatively translocated singletons would show up as pairs elsewhere in the genome in the minimum block size 1 analysis. We have already seen that this is not the case.

5.4 The Coffea Genome

To assess which of these two processes accounts for the *C. arabica* data, we note that pseudogenization, leaving the gene intact, does not shorten the length of the chromosomal region it is in. In contrast, excision of genes, including some or all of the flanking intergenic DNA, will shorten the region. Figures 5.8(a) and 5.8(b) show that the length of the intergenic region remaining after the loss of one or more genes is relatively constant at about 10,000 bp, whereas the length of the corresponding region in the other genome, including the genes that were not lost and their intergenic regions, increases by about 20,000 bp per gene in subCC and 15,000 bp in subCE. So

the regions that have lost annotated genes have lost almost all their DNA sequence. This is striking evidence in favour of the predominance of excision. Note that our analysis is confined to the quantitatively most important case where the two genes bordering the deletion are adjacent, excluding the cases where there are missing genes between neighbours on both subgenomes, explained in Chapter 4.

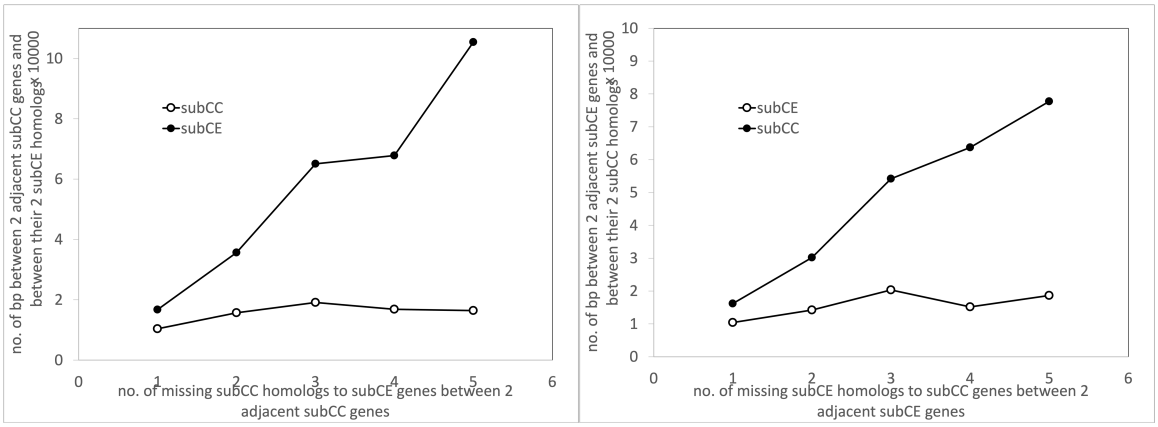


Figure 5.8: Lengths of DNA segment between genes made adjacent by fractionation of intervening genes still present in the opposite subgenome. (a) Fractionation in subCC (left). (b) Fractionation in subCE (right).

The data summarized in Figure 5.8 represent in large measure the unmatched genes that were already absent in the parent genome. To try to distinguish between patterns of loss pre- and post-tetraploidization, we repeated the preceding analysis with restriction to the relatively small subset (11-13%) of genes unmatched in a subCC-subCE synteny block but matched in a synteny block between a subgenome and the opposite parent genome. The results in Figure 5.9 show substantially the same pattern as in Figure 5.8, indicating that the excisions are not just the product of many millions of years of independent evolution, including gene losses, in the parent genomes, but also reflect the fractionation process in the tetraploid.

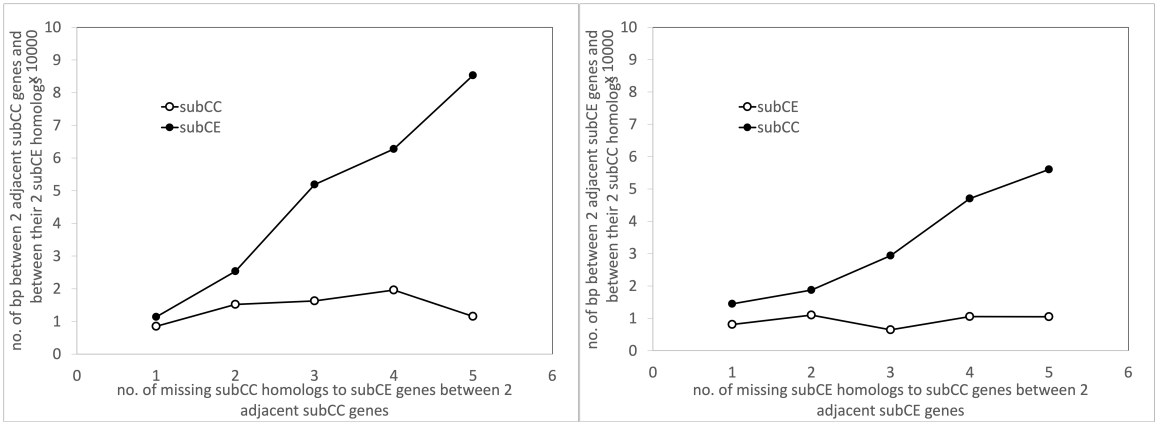


Figure 5.9: Lengths of DNA segment between genes made adjacent by fractionation, *after tetraploidization*, of intervening genes still present in the opposite subgenome. (a) Fractionation in subCC (left). (b) Fractionation in subCE (right).

The aggregate statistics represented in Figures 5.8 and 5.9 do not rule out the possibility of occasional pseudogenization. Unfortunately, the publicly available software for identifying pseudogenes is somewhat dated and was developed in the context of human or vertebrate genomes, and has limited applicability for plant genomes [27]. In our hands, of the three available packages, PseudoPipe [31] was the most productive. It revealed thousands of pseudogenes in the *C. arabica* genome, but these were focused on retrotransposed pseudogenes and highly duplicated pseudogenes, presumably derived from specific classes such as housekeeping genes. There was no specific concern with sets of pseudogenes created almost simultaneously after a polyploidization event. The largest category in the output was that of gene fragments, but these were not controlled for syntenic context. Our search of the full PseudoPipe output found only a handful of pseudogenes in syntenic block gaps, out of many thousands such gaps, that seemed to be derived from a duplicate of unmatched gene in the opposite subgenome.

This prompted us to search directly for fragments of coding DNA in these gaps homologous to parts of unmatched genes in the opposite subgenome side of the synteny block. We required that the level of similarity between these parts and the fragments be greater than 92%, which is the average level of similarity we require for the coding sequence in synteny blocks created by the tetraploidization. (This constraint could be relaxed considerably without any large change in the kinds of fragments found.) Many unmatched genes had different parts connected to disjoint fragments, in which case we used the combined length of the matching parts to characterize the match.

Figure 5.10 shows that fragments of coding sequence, most of them less than 600 bp, and almost all less than 1000 bp – while the average annotated gene contains about 1500 bp – remain in the DNA for a few hundred cases of fractionation. Whether these ever passed through a pseudogene stage is dubious; the original gene seems have been inactivated by a series of deletions of the coding sequence, not necessarily at the 5' or 3' end, but internal to the gene so that several fragments can be matched to different parts of the surviving gene in the opposite subgenome.

5.5 Conclusions

We have developed a graphical based method that can differentiate between excision and pseudogenization. By applying our method to four pairs of genomes, we can argue that excision is the dominant mechanism of gene loss. The statistical evaluation of the massive duplicate gene cohorts created by speciation or polyploidization shows that pseudogenization is either a very rare process or does not result in much stable structure. At the present time, the clear impression is that fractionation simply excises the

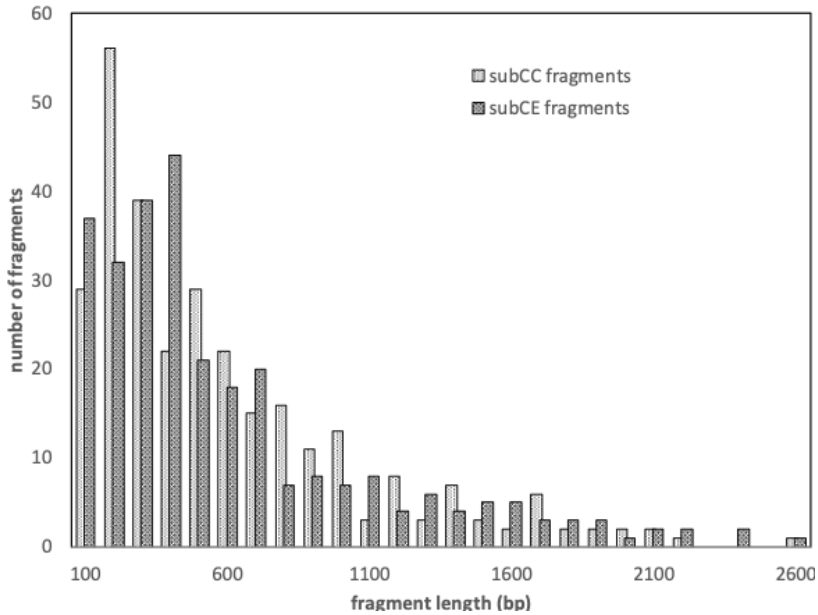


Figure 5.10: Distribution of lengths of fragments matching part(s) of an unmatched gene in the opposite subgenome in a synteny block.

DNA of a gene or several contiguous genes. Ongoing work to be reported elsewhere suggests that this elimination of sequence does occur piecemeal over 30 million years or even 1 million years. It is of course still possible that once a pseudogene is created, or a gene otherwise silenced, its DNA is immediately vulnerable to repeated small deletions, so that the pseudogene itself would be transient. The distinction between this and some single-event excision becomes a matter of semantics.

More surprising perhaps is that gene loss after speciation, occurring independently in two sister genomes, seems to follow the same trajectory. There is of course no genomic interaction between species pairs like *Salvia* and *Tectona*, but their common origin allows us to use one to track the gene loss pattern in the other. There remain questions of how universal excision is; in the *Salvia-Tectona* and *Poplar-Salix* comparisons it is very clear. Because of the genome size differential, it is is harder to

determine in *Linum-Hevea*, while the case of *Malus*, though fractionation proceeds by excision, further gene loss may involve other mechanisms as well. We note that the role of differential amounts of repetitive sequence and active retroposon activity can impact this type of comparison between species, less so within one species.

Both fractionation in the tetraploid, or gene loss from the parent diploids in *Coffea* genomes, appears to proceed mostly by excision, resulting in a shortening of the genome. Pseudogenization is either very rare, or transient. Some large gene fragments remain after an annotated gene is no longer detected, in a substantial minority of cases, as detected by similarity to the surviving paralog.

Chapter 6

Gaps and Runs in Syntenic Alignments

6.1 Introduction

An open question concerning the dynamics process of fractionation after WGD is whether it occurs by single gene deletion through point mutation and pseudogenization, or by multiple neighboring genes lost through one mutation event, such as unequal crossing over.

In the simplest model, proposed over ten years ago [22, 1, 29], at each step a random gene pair is selected to lose one member. In this assumption, the distribution of gap size is simply a geometric distribution $f(n) = p(1-p)^n$, where n is the observed gap size, and p is the probability of gene conserved [22]. According to this reference, this model does account for many large gaps in the observed distribution.

In the next section, we give a new version of this model, where 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 [19, 30], gene loss is effected by excision of a variable length fragment of a chromosome, often formulated in terms of a gamma distribution. As we will see 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.

6.2 One-at-a-time model

Consider the following "fractionation" process. We have an array of n 1's. n presents the number of genes in the chromosome. 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 step t.

Proposition 6.2.1.

$$\begin{aligned} M(0, n) &= 1 \\ M(0, x) &= 0, \qquad \text{for } x \neq n \end{aligned} \tag{6.2.1}$$

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

$$M(t, x) = M(t-1, x) - \frac{xM(t-1, x) - 2\sum_{i>x}^{n} M(t-1, i)}{n-t+1}, \tag{6.2.2}$$

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 $n - t + 1$. ∎

We simulated chromosomes with various number of genes, in 1000 times, all initially marked with value 1. We substituted 0 for a random gene hitherto marked as 1, one at a time until the entire chromosome was marked 0. Simulation was repeated by 1000 times. After step t, we noted the average number of runs of zeroes $n(t, 1)$, $n(t, 2)$, . . . of length 1,2,... displayed the pattern in Figure 6.1. The axes in the figure are labelled in units of proportion of total chromosome length n, because the curves so normalized seem to approach limiting shapes as n increases.

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, Figure 6.2 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.

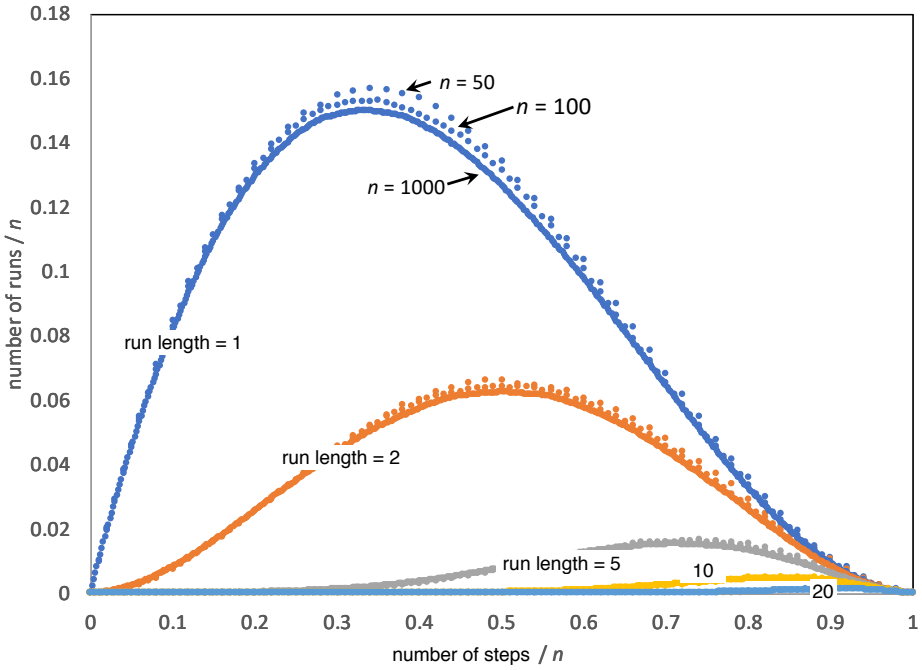


Figure 6.1: Average number of runs of lengths 1, 2, 5, 10 and 20 in a chromosome of size $n = 50$, 100 and 1000.

This process bears much resemblance to the theory of runs [39] in random binary sequences. Given n Bernoulli trials with a probability of success $p = t/n$, 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 is not independent of what has happened in other positions, but 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 Figure 6.3, however, no matter how many steps are posited (giving the different dotted lines) the empirical data from the Coffea genomes. It is clear that the model is inadequate to account for both the simultaneous steep drop-off from gaps

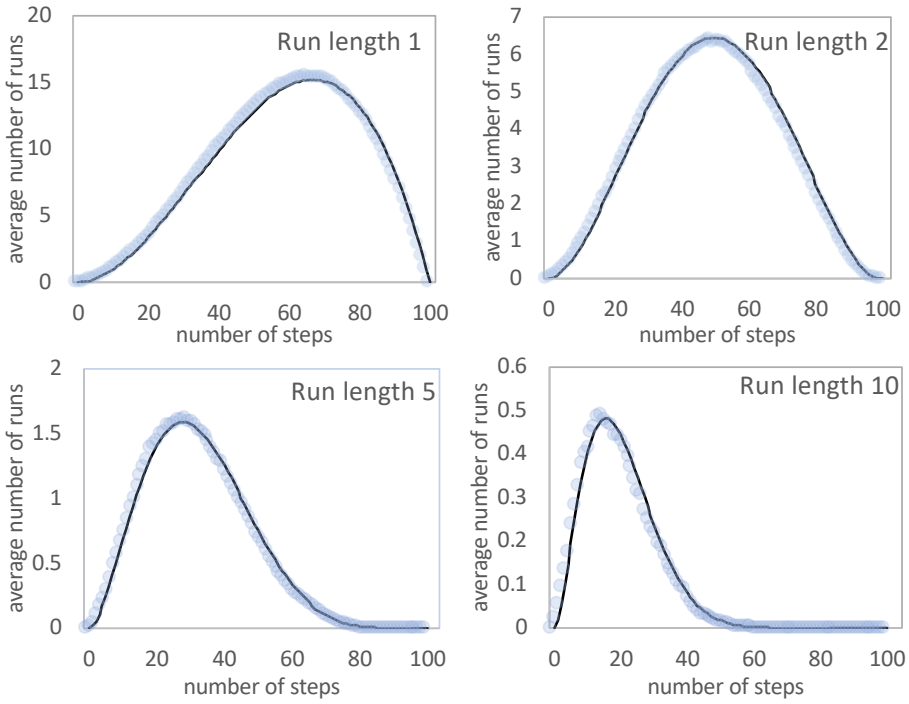


Figure 6.2: Evolution of number of runs of 1's of various sizes as the number of steps t increases. Solid line: recurrence. Light blue background line: average of 1000 simulations. Genome length $n = 100$.

of size 1 and the presence of significant number of long gaps.

6.3 The combined model

To remedy the poor fit of the one-at-a-time model, we combine it with a gamma distribution component. As shown in Chapter 1.1.1, the length of deleted genes are inclined to be gamma distributed. 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 θ to apportion the two components. Fortunately, the optimal gamma component turned out to be a simple geometric

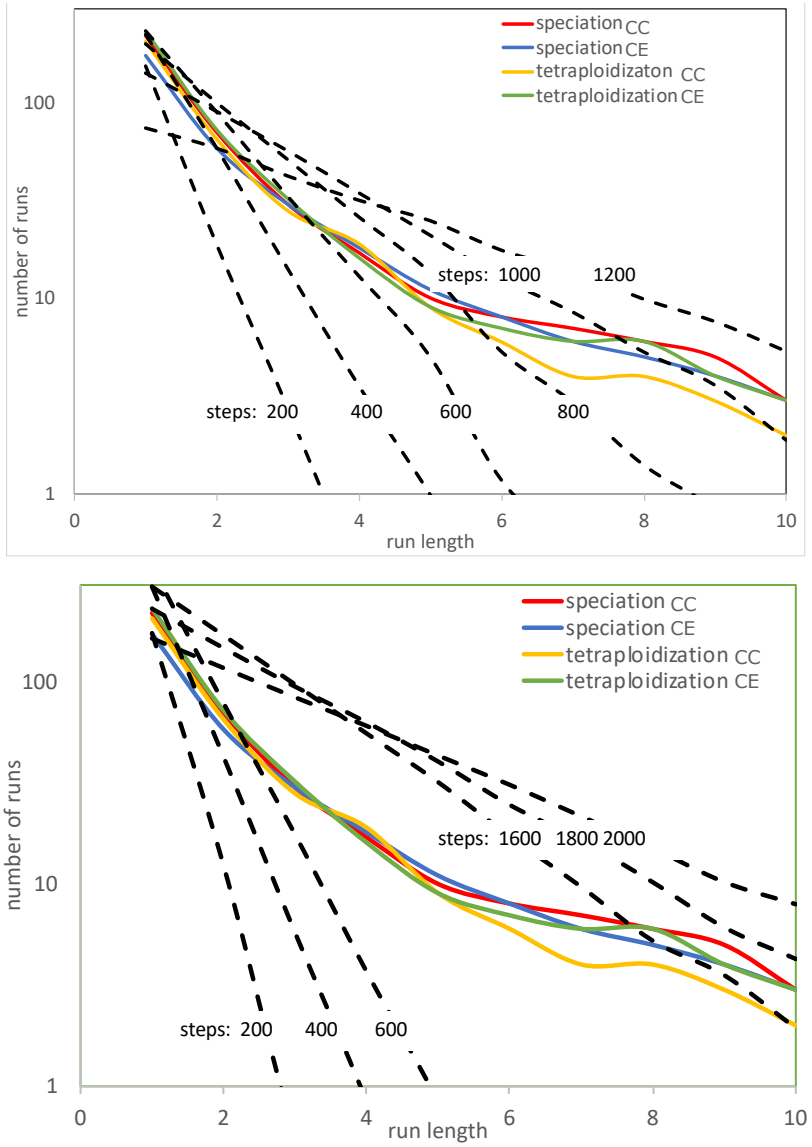


Figure 6.3: Inability of one-at-a-time model to fit *Coffea* data on runs of zeros (gaps), with chromosome size $n = 1600$ (top) or $n = 2800$ (bottom), and at various time intervals (steps).

distribution, with only one parameter. (The optimal shape is consistently close to 1, suggesting that the process is Poisson.)

To estimate the n, the geometric parameter p, and the proportion of steps θ

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 fit criterion, when running 50 simulations of each combination if a range of values of n, p and θ. (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, 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 2.2 and 2.3. The optimal values were $\theta = 0.7$, $n = 2800$ and $p = 2/7$.

Our model breaks down when we use it to simulate fractionation after γ, as can be seen in the bottom panel of Figure 6.4. The simulations suggest there should remain no long runs of 1's, but this is likely due to the inability to detect sufficiently long synteny blocks after extensive fractionation, and possibly some tendency for some regions of neighbouring genes to resist fractionation.

6.4 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 γ 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 synteny 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.

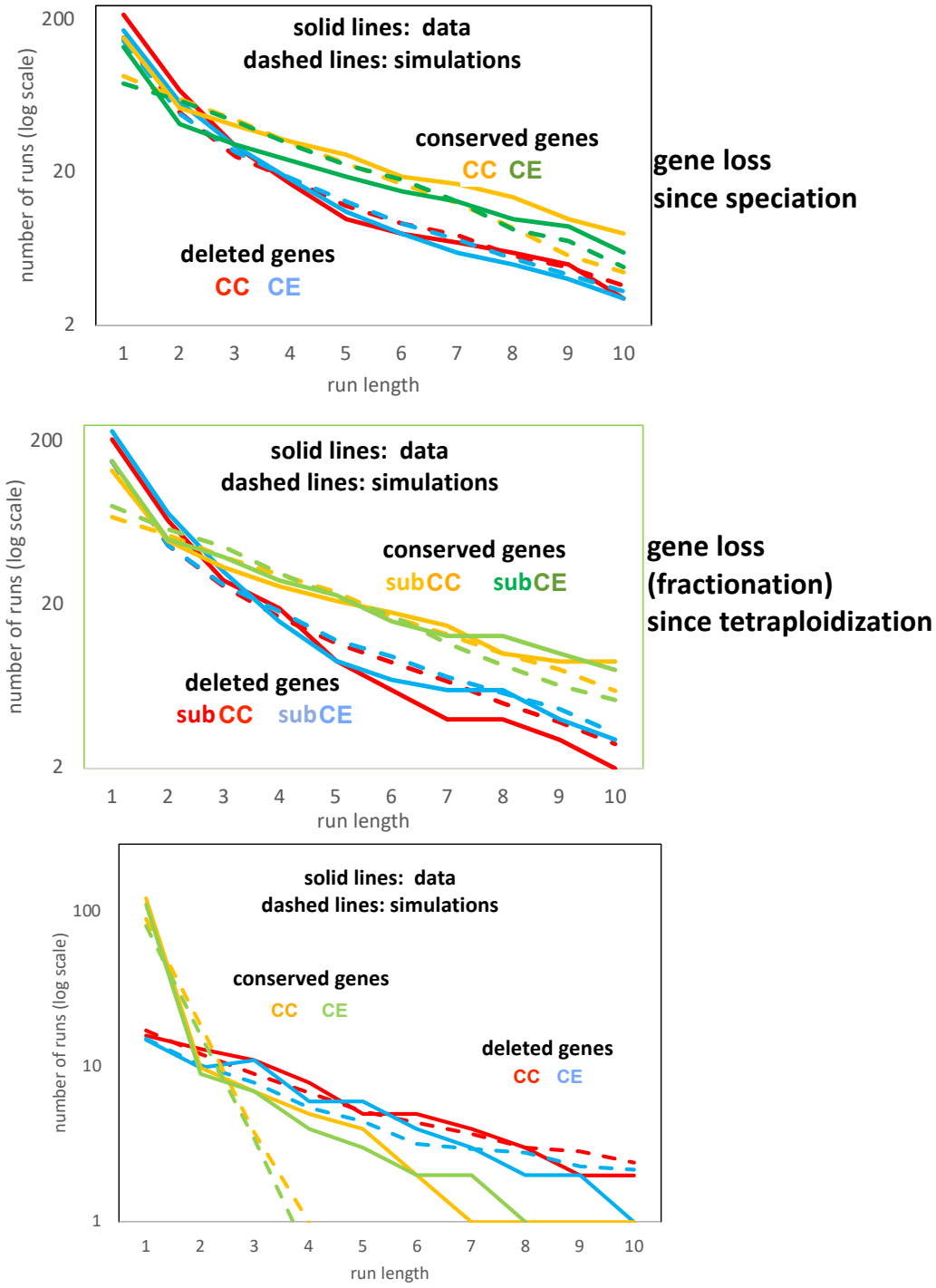


Figure 6.4: Comparisons of the distributions of run sizes (0's and 1's) with simulations of combined model in the speciation data (top panel), tetraploidization data (middle panel) and γ data (bottom panel).

Chapter 7

Conclusion

Fractionation, part of the long-term sequelae to whole genome duplication, has generally been treated as an afterthought, noise in the detection and characterization of ancient polyploidization events. In this thesis, however, I showed how fractionation is a multi-facetted phenomenon of biological import, and studied many aspects of this process prevalent in plant genomic evolution.

Much of my data and analysis pertains to arabica coffee genome, which was accessible through my membership in the ACGC. Many of the analyses benefitted from the availability of the annotated genome not only of the tetraploid CA species, but also the two progenitor species CC and CE, or rather of the contemporaneous diploid descendants of these two.

By comparing the timeline of evolution for CA with the local modes values of the distribution of similarity or synonymous mutation rates (K_s), I studied the statistics of the gaps between adjacent pairs of duplicate genes within synteny blocks in each of the three evolutionary events: gamma hexaploidization, CC-CE speciation, and CA tetraploidization.

I found a high degree of parallelism in the evolution of CC and CE in terms of number of genes lost since speciation, and similar parllelism in the subgenomes subCE and subCC since tetraploidizaiton. I also studied the distribution of conserved genes and discovered this was much stronger in regions of the chromosomes remote from centromere. The sparsity of conserved genes in the pericentromeric regions followed parallel patterns in all chromosomes in both subgenomes and in both progenitors.

I analyzed the retention rate of deletion events and concluded that the frequency distribution of gap lengths (number of deleted genes – not nucleotides) within syntenic blocks could be fit well by 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. As part of this work we derived a new recurrence for the distribution of gap lengths in the one-gene-at-a-time model.

The detailed syntenic mechanisms of gene loss, especially in fractionation, remain controversial. I focused on fractionation mechanisms of loss of gene pair homology after polyploidization. I found that the region having lost annotated genes have lost almost all their DNA sequence, suggesting that fractionation in the tetraploid, or gene loss from the parent diploids, appears to proceed mostly by excision. This also reveals that pseudogenization is either very rare, or transient. Some large gene fragments remain after an annotated gene is no longer detected, in a substantial minority of cases, as detected by similarity to the surviving paralog.

To ensure that these results reveal a process widespread in the plant world, I replicated it in four pairs of related genomes across the angiosperms. In each of the analyses I proposed and carried out, I discovered biological tendencies not previously documented: the detailed parallelism between the evolution of subgenomes in *C. arabica*, the combination of two processes with account for gap size distribution, and the

great preponderance of excision over pseudogenization. A common trend, completely unexpected, running through all of these is that gene loss through fractionation is not qualitatively different from gene loss in unduplicated (diploid) genomes. Indeed most of the fractionation apparent in subCC and subCE is the result of losses incurred after the speciation of CC and CE, and *before* tetraploidization. the combination of loss models to account for gap sizes in the subgenomes, is the same as for ordinary gene loss in progenitor species. And the preponderance of excision over pseudogenization that characterizes fractionation, applies equally well to gene loss from synteny blocks in independently evolving genomes.

This previously unremarked tendency suggest many lines of further research both in modelling, statistics and in the laboratory, all pertaining to the question of what precisely differentiates the process of fractionation statistically from gene loss unaffected by the presence of close paralogs due to polyploidization.

Appendix A

Chapter 3 Retention Graphs

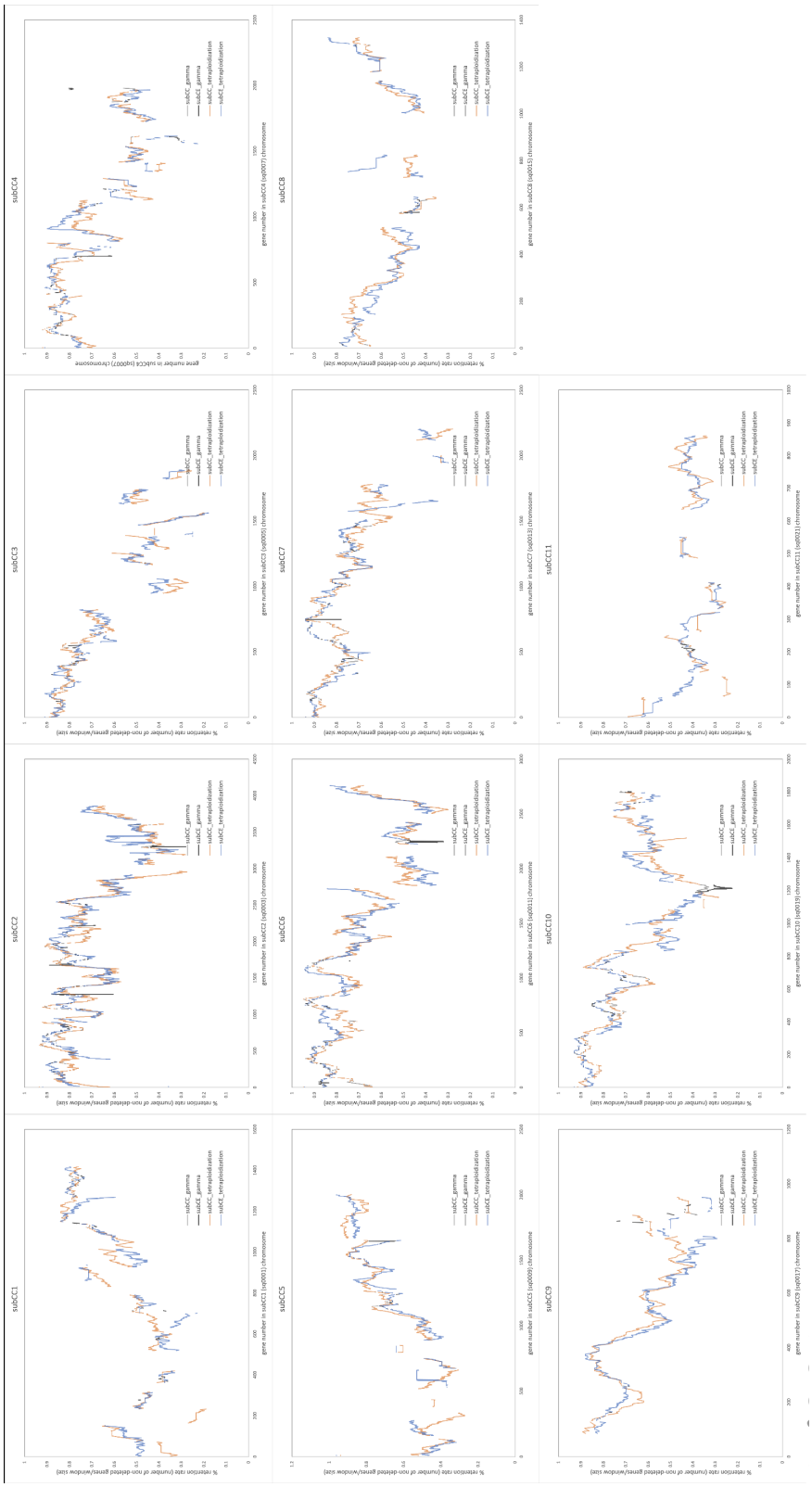


Figure A.1: Gene retention rate targeted on indicated chromosomes on *C. arabica*-*C. canephora* subgenome.

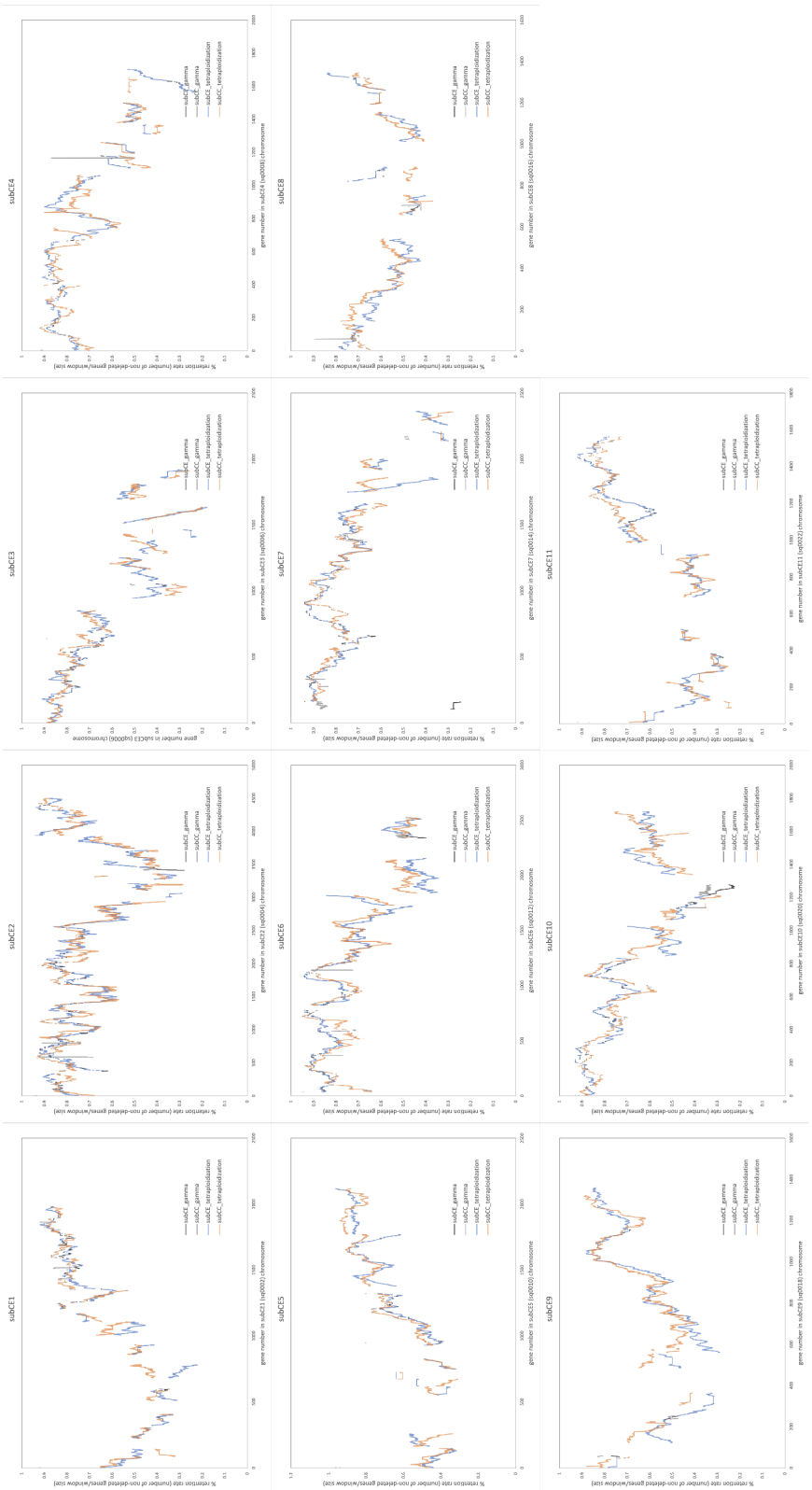


Figure A.2: Gene retention rate targeted on indicated chromosomes on *C. arabica*-*C. eugenioides* subgenome.

Bibliography

[1] Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion (JK Byrnes, GP Morris, WH Li) Molecular Biology and Evolution 23, 1136–1143 (2006)

[2] Fractionation, rearrangement, consolidation, reconstruction (D. Sankoff, C. Zheng), Models and algorithms for genome evolution (C. Chauve, N El-Mabrouk, E. Tannier eds.), Springer, 247-260 (2013)

[3] The coffee genome provides insight into the convergent evolution of caffeine biosynthesis (F. Denoeud, L. Carretero-Paulet, A. Dereeper, G. Droc, R. Gxuyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, JM. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M. Combes, D. Crouzillat, C. Da Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Leclercq, M. Lepelley, T. Leroy, LT. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. VanBuren, Q. Zhang, AC. Andrade, X. Argout, B. Bertrand, A. de Kochko, G. Graziosi, RJ. Henry, Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, VA. Albert, P. Wincker, P. Lashermes) Science 345, 1181-1184 (2014).

[4] A sense of self: the role of DNA sequence elimination in allopolyploidization (NA Eckardt) Plant Cell 13, 1699-1704 (2001)

[5] The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla (O Jaillon, JM Aury, B Noel, A Policriti, C Clepet, *et al.*) Nature 449, 463–467 (2007)

[6] A consolidation algorithm for genomes fractionated after higher order polyploidization (K. Jahn, C. Zheng, Jakub Kováč, D. Sankoff) BMC Bioinformatics 13, S19:S8 (2012)

[7] Multiplicity of genome equivalents in the radiation-resistant bacterium *Micrococcus radiodurans*. (Hansen MT) Journal of Bacteriology 134 (1) 71–75. (1978)

[8] Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species (P Hamon, CE Grover *et al.*) Molecular Phylogenetics and Evolution 109, 351–361 (2017)

[9] Aims and goals of the Arabica Coffee Genome Consortium (ACGC) (A De Kochko, D Crouzillat, Arabica Coffee Genome Consortium) 12th Solanaceae Conference (2015)

[10] How to usefully compare homologous plant genes and chromosomes as DNA sequences (E Lyons, M Freeling), The Plant Journal 53, 661–673 (2008)

[11] Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids (E Lyons, B Pedersen, J Kane, M Alam, R Ming, H Tang *et al.*)

[12] The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids (E Lyons, B Pedersen, J Kane, M Freeling), Tropical Plant Biology 1, 181–190 (2008)

[13] The EMMIX software for the fitting of mixtures of normal and t-components (GJ McLachlan, D Peel, KE Basford, P Adams) Journal of Statistical Software 4, 1–14 (1999)

[14] Evolution by Gene Duplication (Ohno S) Springer (1970)

[15] Internal validation of ancestral gene order reconstruction in angiosperm phylogeny (D. Sankoff, C. Zheng, P.K. Wall, C. dePamphilis, J. Leebens-Mack, V. Albert) Proceedings of the RECOMB 2008 Workshop on Comparative Genomics (S. Vialette, C. Nelson, eds.) Lecture Notes in Computer Science, Springer (2008)

[16] The collapse of gene complement following whole genome duplication (Sankoff D, Zheng C, Zhu Q.) BMC Genomics 11, 313 (2010)

[17] Fractionation, rearrangement and subgenome dominance (D. Sankoff, C. Zheng), Bioinformatics 28, 402-408 (2012)

[18] A model for biased fractionation after whole genome duplication (Sankoff D, Zheng C, Wang B.) BMC Genomics 13, S1—S8 (2012)

[19] Structural *vs.* functional mechanisms of duplicate gene loss following whole genome doubling (D Sankoff, C Zheng, B Wang, C Fernando Buen Abad Najar) BMC Genomics 15, DOI: 10.1109/ICCABS.2014.6863915 (2015)

[20] Models for similarity distributions of syntenic homologs and applications to phylogenomics (D Sankoff, C Zheng, Y Zhang, J Meidanis, E Lyons, H Tang),

IEEE/ACM Transactions in Computational Biology and Bioinformatics 16, 727–737 (2019)

[21] The obligate human pathogen, Neisseria gonorrhoeae, is polyploid. (Tobiason DM, Seifert HS.) PLoS Biol. 2006;4(6):e185. doi:10.1371/journal.pbio.0040185

[22] The role of mutational dynamics in genome shrinkage (MJ van Hoek, P Hogeweg) Molecular Biology and Evolution 24, 2485–2494 (2007)

[23] The chromosomes: Their number and general importance. (Winge Ö) Comptes Rendus des Travaux du Laboratoire Carlsberg 13:131–275. (1917)

[24] Molecular evidence for an ancient duplication of the entire yeast genome. (Wolfe K and Shields D) Nature 387: 708–713 (1997)

[25] Fractionation statistics (Wang B, Zheng C, Sankoff D.) BMC Bioinformation 12, S9—S5 (2011)

[26] Polyploidy Index and Its Implications for the Evolution of Polyploids (J Wang, J Qin, P Sun, X Ma, J Yu, Y Li, S Sun, T Lei, F Meng, C Wei, X Li, H Guo, X Liu, R Xia, L Wang, W Ge, X Song, L Zhang, D Guo, J Wang, S Bao, S Jiang, Y Feng, X Li, AH Paterson., X Wang), Frontiers in Genetics 10, 807 (2019)

[27] Pseudogenes and their genome-wide prediction in plants (J Xiao, MK Sekhwal, P Li, R Ragupathy, S Cloutier, X Wang , FM You) International Journal of Molecular Sciences 17, 1991-2006 (2016)

[28] Evolutionary origins of pseudogenes and their association with regulatory sequences in plants. (Xie J, Li Y, Liu X, Zhao Y, Li B, Ingvarsson PK, Zhang D) The Plant Cell 31(3):563–578. doi:10.1105/tpc.18.00601 (2019)

[29] Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* diploid (C Zheng , PK Wall, J Leebens-Mack, C dePamphilis, VA Albert, D Sankoff) Journal of Bioinformatics and Computational Biology 7, 499–520 (2009)

[30] A continuous analog of run length distributions reflecting accumulated fractionation events (ZN Yu, D Sankoff) BMC Bioinformatics 17 (Suppl 14), 412 (2016).

[31] PseudoPipe: an automated pseudogene identification pipeline (Z Zhang, N Carriero, D Zheng, J Karro, PM Harrison, M Gerstein) Bioinformatics, 1437–1439 (2006)

[32] The effect of massive gene loss following whole genome duplication on the algorithmic reconstruction of the ancestral Populus diploid (C. Zheng, P.K. Wall, J. Leebens-Mack, C. dePamphilis, V.A. Albert, D. Sankoff) Annual International Conference on Computational Systems Bioinformatics CSB (2008)

[33] Evolutionary model for the statistical divergence of paralogous and orthologous gene pairs generated by whole genome duplication and speciation (Y. Zhang, C. Zheng, D. Sankoff),IEEE/ACM Transactions in Computational Biology and Bioinformatics, 1545-5963 (2017).

[34] Pinning down ploidy in paleopolyploid plants (Y. Zhang, C. Zheng D. Sankoff), BMC Genomics 18, 19(Suppl 5): 287 (2018).

[35] Speciation and rate variation in a birth-and-death account of WGD and fractionation; the case of Solanaceae (Y. Zhang, C. Zheng, D. Sankoff), Proceedings of the 16th International Conference, RECOMB-CG 2018: Comparative Genomics (M. Blanchette, A. Ouangraoua, eds.), 146-160 (2018).

[36] A branching process for homology distribution-based inference of polyploidy, speciation and loss (Y. Zhang, C. Zheng, D. Sankoff), Algorithms for Molecular Biology 14, 18 DOI: 10.1186/s13015-019-0153-8 (2019)

[37] Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment (Y. Zhang, C. Zheng, D. Sankoff), BMC Bioinformatics 19, 20 (S20):635 (2019)

[38] Branching out to speciation with a birth-and-death model of fractionation: the Malvaceae (Y. Zhang, C. Zheng, S. Islam, Y.M. Kim, D. Sankoff), IEEE/ACM Transactions in Computational Biology and Bioinformatics 16, in press.

[39] Weisstein, Eric W. "Run." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/Run.html, accessed July 15 (2021)