Fractionation resistance of duplicate genes following whole genome duplication in plants as a function of gene ontology category and expression level

Eric Chun-Hung Chen

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfilment of the requirements
for the Doctorate in Philosophy degree in Biology[1]

Department of Biology
Faculty of Science
University of Ottawa

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

Abstract

With the proliferation of plant genomes being sequenced, assembled, and annotated, duplicate gene loss from whole genome duplication events, also known in plants as fractionation, has shown to have a different pattern from the classic gene duplication models described by Ohno in 1970. Models proposed more recently, the Gene Balance and Gene Dosage hypotheses, try to model this pattern. These models, however, disagree with each other on the relative importance of gene function and gene expression. In this thesis we explore the effects of gene function and gene expression on duplicate gene loss and retention.

We use gene sequence similarity and gene order conservation to construct our gene families. We applied multiple whole genome comparison methods across various plants in rosids, asterids, and Poaceae in looking for a general pattern. We found that there is great consistency across different plant lineages. Genes categorized as metabolic genes with low level of expression have relatively low fractionation resistance, losing duplicate genes readily, while genes categorized as regulation and response genes with high level of expression have relatively high fractionation resistance, retaining more duplicate gene pairs or triples.

Though both gene function and gene expression have important effects on retention pattern, we found that gene function has a bigger effect than gene expression. Our results suggest that both the Gene Balance and Gene Dosage models account to some extent for fractionation resistance.

Préface

Avec la prolifération des génomes des plantes qui sont séquencés, assemblés, et annotés, la perte de gènes dupliqués à partir d'évènements de duplication du génome complet, aussi connu comme fractionnement, a démontré différents patrons comparativement au modèle classique de duplication génique décrit par Ohno en 1970. Des modèles qui ont été proposés plus récemment par les hypothèses équilibré de gènes et de dosage génique essaient de modeler ce patron. Par contre, ces modèles sont en désaccord en ce qui a trait à l'importance relative de la fonction et de l'expression génique. Dans cette thèse nous examinons les effets de la fonction et l'expression génique sur la perte et la rétention des duplications géniques.

Nous utilisons la similarité génique et la conservation de l'ordre génique afin de construire nos familles géniques. Nous avons appliqué des méthodes de comparaison au niveau du génome complet chez diverses plantes des familles de rosids, asterids, and Poaceae pour identifier des patrons généraux. Nous avons découvert qu'il existe une grande uniformité entre les différentes lignées de plantes. Les gènes catégorisés comme jouant un rôle dans le métabolisme avec un faible niveau d'expression ont une résistance fractionnaire relativement basse; perdant les gènes dupliqués rapidement. Par contre les gènes régulateurs et responsables d'une réponse génique ayant un haut niveau d'expression ont généralement une grande résistance de fractionnement; ils retiennent un plus grand nombre de paires ou de triplets de gènes dupliqués.

Même si la fonction et l'expression génique ont des effets importants sur le patron de rétention, nous avons découvert que la fonction génique a un plus gros impact que l'expression génique. Nos résultats suggèrent que les deux modèles; celui de l'hypothèse équilibrée génique et celui du dosage génique, sont en partie responsables de la résistance au fractionnement.

Acknowledgements

I want to dedicate this thesis to my family for all the support they have given me. Also thanks to my supervisor David Sankoff for the guidance and support over the last 5 years.

I also want to thank Yuh-Ing Chen, Annie Morin, and Jean-Hugues Chauchat for their life-saving advice and pointers on statistical questions. Chunfang Zheng taught and explained to me the gene family identification algorithms that are the fundamental part of this project, for which I am very grateful. Aoife McLysaght for her interest and for drawing my attention to the work of Gout *et al.* on *Paramecium*. John Basso for the translation of the abstract into French. Jonathan Ferrier for inviting me to collaborate on various projects. My committee members, Nicolas Corradi, Michel Dumontier, and Stephen Gleddie for their supervision over the years. My examiners for giving me their valuable time to review and critique my work.

I want to thank uOttawa for the funding opportunity to allow me to enroll and complete my doctoral thesis. Finally, I want to thank all the colleagues and friends that I met at uOttawa for making this an enjoyable experience.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Introduction

Novel genes are important for evolving adaptation to new circumstances. But how does a species go about getting genes? While the precise processes and intricacies by which a species gains new genes are still debated, a broad consensus has formed. This includes the importance of gene duplications. In 1970 Ohno [1] proposed the process behind the terms (coined by Force *et al.* [2]) neofunctionalization, subfunctionalization, and nonfunctionalization (now called pseudogenization), in describing possible duplicate gene fates. These models have held up remarkably well with the advent of increasingly sophisticated molecular biology techniques and the rapid advances in whole genome sequencing. One major addition to Ohno's models is the Gene Dosage Model of Papp *et al.* [3] and the

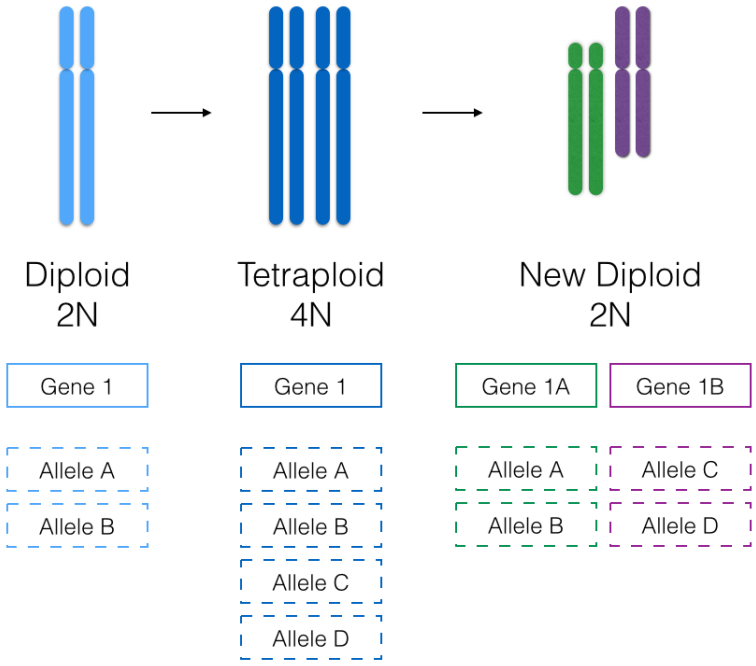


FIGURE 1.1: Whole genome duplication example. A diploid ancestor undergoes tetraploidization to form a tetraploid ancestor. This intermediate ancestor then undergoes fractionation (gene loss from whole genome duplication in plants) and other genome rearrangements to become a present day diploid species.

Gene Balance Hypothesis of Birchler *et al.* [4–7] to address the issue of whole genome duplications.

These new models are increasingly important in study of plant genome evolution because following a number of whole genome sequencing projects, whole genome duplication is now considered to be a frequent process [8, 9].

This thesis focuses on exploring the finer details of these new models of duplicate gene fate from whole genome duplication (Figure 1.1). First I will review the different modes of duplication and the different models of duplicate gene fate following duplication events.

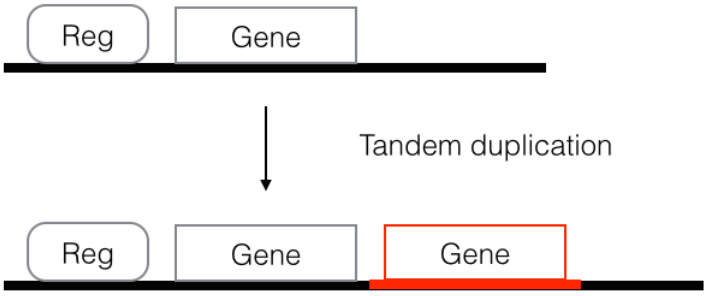


FIGURE 1.2: Tandem duplication example. "Reg" stands for regulatory elements, which may not be duplicated (red) in tandem duplications.

1.2 Modes of duplication

1.2.1 Tandem Duplications

Tandem duplications are local duplications where a single gene, or a few genes that are physically next to each other, are duplicated. These duplications may not duplicate all the regulatory elements controlling the gene as some regulatory elements maybe located far from the regions that is being duplicated; such as trans-regulated-elements (Figure 1.2).

Unlike whole genome duplications, which occur on the scale of millions of years, tandem duplications occur regularly in evolutionary history. Tandem duplications only introduces a few new genes at a time, so changes introduced by one event are small compared to other modes of duplication. Nevertheless such mutations may accumulate over time, resulting in important phenotypic change.

1.2.2 Whole genome duplication

Whole genome duplications can be subdivided into two types. The first type is the autopolyploids, where chromosomes from the same species double. The second type is the allopolyploids, where hybrid polyploids between species are formed. In both cases the original regulatory elements in the genomes are doubled alongside the genes they regulate. The primary difference in fractionation (extensive and rapid gene loss following whole genome duplication in plants) for these two types of polyploids is the biased distribution of gene loss in allopolyploids where one subgenome will dominate over other subgenomes (Figure 1.3 [10, 11]).

In the context of duplicate gene retention in this thesis, we make no distinction between the two types of whole genome duplications for two reasons. The first reason is because the whole genome duplication events of the species that we analyze in this thesis are far too ancient for us to conclusively to distinguish which type of whole genome duplication it is. The second reason is because we will show how consistent the retention pattern is across the various lineages that we analyze, which may or may not have experienced the same types of whole genome duplication. For these reasons, we are confident our result on the impact of functional categories and gene expression levels have on duplicate gene retention does not distinguish between the two types of whole genome duplications.

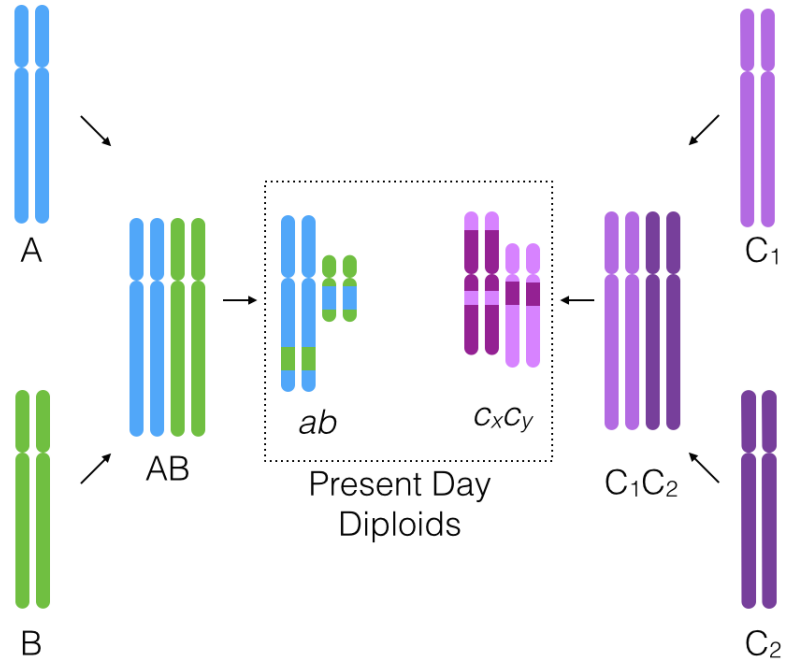


FIGURE 1.3: Allopolyploidy versus autotetraploidy. A simple example of the difference between allopolyploidy and autotetraploidy and the consequence on genome structure. Ancestor A and B hybridizes to form an intermediate allopolyploid ancestor, which then diploidized into a diploid present day genome. For the intermediate allopolyploid ancestor, fractionation is predominately on one of the subgenomes; in this example it is genome B. In contrast, the intermediate autotetraploidy ancestor C_1C_2 of ancestor C, also diploidized into a diploid present day genome but the parental chromosomes are virtually indistinguishable in the present day diploids.

Figure 1.4 show cases where only a chromosome is duplicated or a portion of a chromosome is duplicated, called segmental duplications [6]. These segmental duplications may be similar to tandem duplication, at one extreme, or to whole genome duplications at the other extreme. These genes are hard to identify, we do not dwell on them in this thesis.

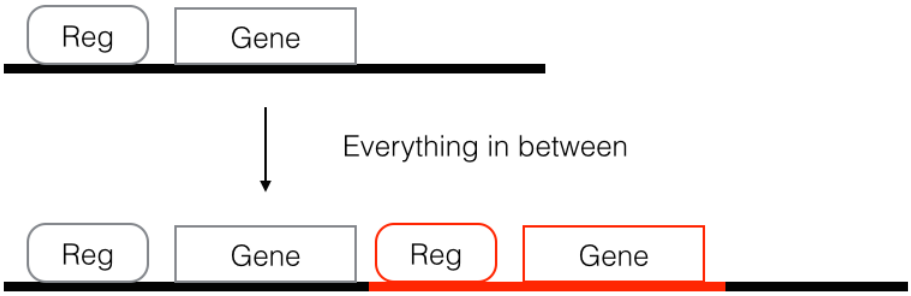


FIGURE 1.4: Segmental duplication. "Reg" stands for regulatory elements. Sometimes referred to as "segmental duplication" [6], this type of duplication can be considered to be tandem duplication that covers a large segment. However, there is no agreed standard on how large of the segment need to be to distinguish between a tandem duplication and a segmental duplication.

1.3 Modes of duplicate gene fate

1.3.1 Classical Models

The classical model of duplicate gene fate includes: the gain of function (neofunctionalization), the loss of the gene, and the retention of both duplicates with divergent function (subfunctionalization) [1, 2]. Force *et al.* [2] explains in great detail the rationale and examples of these models in their 1999 paper. I will briefly introduce the models and the rationale behind the model.

Neofunctionalization

Under the neofunctionalization model, one of the duplicate genes acts as a buffer against loss of function while the other duplicate gene evolves new functions. This allows one

of the duplicate genes to accumulate more mutations or other changes before it poses a significant fitness penalty to the species. Without duplicate genes, the only other sources of novel genes would be the spontaneous formation of new genes in non-coding regions of the genome or pseudogenes (remnants of a gene) or by introduction from another species (xenologs).

Gene loss

Gene loss is by far the most prevalent outcome for duplicate genes with many current genomes' size being much smaller than it should be if the polyploid ancestor genome lost no duplicate genes [12–14]. Duplicate genes can be lost in two ways. One way is by inactivation, where duplicate genes will accumulate enough mutations that it becomes a nonfunctional pseudogene before fully disappearing from the genome. Another way is by DNA excision, where duplicate genes are simply deleted from the genome via recombination. The actual mechanism of gene loss is not important in this thesis.

The specific type of gene loss process this thesis focus on is fractionation. This process, while most visible during the diplodization of the species, is a continuous process so it does not stop even after the species is rediploidized. There is also debate on whether fractionation happens one gene at a time or this happens on blocks of genes at a time [15–17].

Subfunctionalization

Subfunctionalization is the most popular model for retention of tandem duplicate genes [6], it argues that the duplicate genes will each specialize during evolution such that a function that used to take one gene now takes two genes. One of the benefits when duplicate genes subfunctionalize is the increase in regulatory diversity [18]. An example is the duplicated *engrailed* gene in zebrafish where one duplicate is expressed pectoral appendage while the other is expressed in hindbrain/spinal cord [2]. The single copy ortholog present in chicken and mouse are expressed in both organs [2].

Functional enrichment will be a result of some functional classes being more favourable than other functional classes for specialization. As such, the functional enrichment does not distinguish between tandem duplicates from whole genome duplication duplicates.

However, studies of duplicate genes from whole genome duplication present a different pattern of gene retention [3, 19]. The difference is the contrasting patterns of functional enrichment between retained duplicates of tandem duplication and whole genome duplication. Therefore, new models are necessary to explain this different pattern of gene retention.

The subfunctionalization model is not sufficient to explain this difference because the mechanism described by the subfunctionalization model does not distinguish tandem duplicates from whole genome duplication duplicates.

1.3.2 Gene Dosage and Gene Balance

While the dosage of genes has always been considered important in the context of gene duplication [20], how it impacts and influences duplicate gene retention post-whole genome duplication events was first described by Papp *et al.* [3] in 2003. The difference in retention pattern between whole genome duplication duplicates and tandem duplicates has since been observed and described in many model organisms such as yeast [3, 21], *Paramecium* [19, 22], *Arabidopsis* [23], poplar [24], and more [20]. The two main models, Gene Dosage [3] and Gene Balance [7] can be reinterpreted as Relative Dosage and Absolute Dosage [25] to illustrate the fundamental difference between the two models.

Relative Dosage argues that the need to maintain stoichiometry ratios of interacting genes creates a selection pressure against losing copies of the duplicate genes post whole genome duplication [7] (figure 1.5). This idea is similar to the classical view that aneuploidy has a higher cost to fitness than polyploidy [7, 20]. The model predicts genes involved in multi-unit complexes or otherwise involved in complex regulatory systems to be preferentially retained after whole genome duplicate, such as the circadian clock genes in *Brassica rapa* [26].

Absolute Dosage, on the other hand, argues that simply by being highly expressed the gene will be more likely to be retained [27] or that higher copy number is selected for with implied higher expression and higher number of gene products [20]. The model

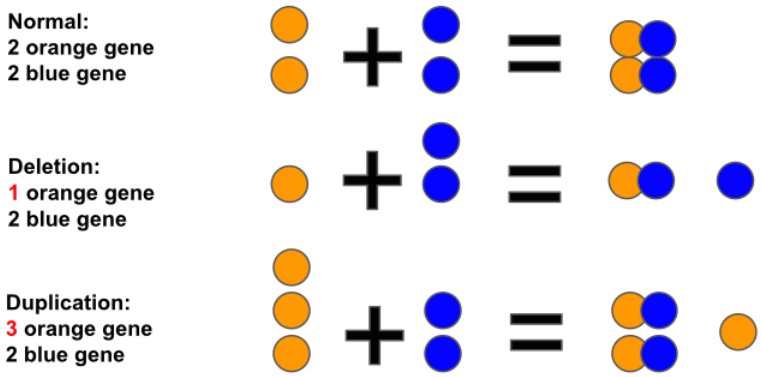


FIGURE 1.5: Relative dosage model. A blue and orange gene product is required to form a functional product. So after whole genome duplication, a loss or gain of any single blue or orange gene would create an imbalance in the stoichiometry of the functional product [7].

explains that high gene expression is costly to the species, therefore these costly genes must contribute in a significant way to the fitness of the species. It follows then, that there is a harsher fitness cost to disrupt the dosage of these high expressing genes than low expressing genes, because high expressing genes are implicit to be more important than low expressing genes [20, 25]. The comprehensive study on expression and duplicate gene retention on the model organism *Paramecium* by Gout *et al.* [27] provides a strong case that expression level is the most important factor in determining duplicate gene fate after whole genome duplication.

In the end, as summarized in the review by Conant *et al.* [20], the two models are not mutually exclusive and it is likely that both models account to some extent for duplicate gene retention after whole genome duplication.

1.4 Objective of the thesis

Research on duplicate gene retention has been focused [3, 24–28] on single specific model organisms. This is both due to the limitation of high quality genome assemblies and the need to demonstrate the specific genes or gene networks to support proposed models. The objective of this thesis is to extend the analysis beyond traditional model organisms and to measure if the retention patterns identified in the carefully studied model organisms persist across a wider variety of plant genomes. In particular, the emphasis is on exploring the role of functional categories and gene expression on duplicate gene retention post-whole genome duplication events.

1.4.1 Data

Our genomic data consist of three sets, providing different insights into duplicate gene retention post-whole genome duplication. They are the rosid dataset, the asterid dataset, and the Poaceae dataset. The rosid dataset contains the *Vitis vinifera* [12], *Prunus persica* [29], and *Theobroma cacao* genomes. These rosids separate out early in the evolutionary history from a hexaploid core eudicot ancestor and do not have additional rounds of whole genome duplication. In contrast, the asterid dataset, which consists of *Utricularia gibba* [14], *Solanum lycopersicum* [30], and *Mimulus guttatus* [31], have additional rounds of whole genome duplication. Lastly, the Poaceae dataset, consists of *Sorghum*

bicolour [32], *Oryza sativa* [33], *Setaria italica* [34], and *Brachypodium distachyon* [35]. These Poaceae species are part of another evolutionary lineage (monocots) so we can verify the generality of the results across a broad spectrum of plant species.

Based on the genomic datasets in CoGe [36] we use sequence similarity, gene order conservation, and the "Orthologs for Multiple Genomes" program [37, 38] to produce gene families. We use Gene Ontology [39] for describing gene functions and we use Blast2GO [40] to obtain the functional annotations. The expression dataset is limited to just *Vitis vinifera* [41] for the rosids and *Solanum lycopersicum* [30] for the asterids due to limited high quality RNA-seq data. However, the RNA-seq data that we use covers multiple organs. This provides a fair snapshot of the gene expression profiles of the species.

1.4.2 Overview of the chapters

This thesis is based on three published manuscripts and one in-preparation manuscript. I am the lead author in three of the manuscripts. In the case of multi-author papers the contribution of each author is detailed at the end of the chapter in section title "Authors' contributions". The manuscripts are organized by date and reflect the progression of the research for this thesis.

Chapter 1 covers the general background of the subject matter. Chapter 2 describes,

in greater detail, the method for identifying gene families related by whole genome duplication events. This is also the starting point of this thesis project where we start to investigate the effect of functional categories on duplicate gene retention. Chapter 2 was a collaborative effoer; my contribution was mainly section 2.5 on functional analysis.

In Chapter 3 we utilize the method developed in Chapter 2 to do an in depth analysis of functional categories on duplicate gene retention. We also investigate whether functional bias is present only in gene families that are part of highly conserved syntenic blocks or if the pattern can be found in the gene families that are more loosely defined.

The effect of expression on duplicate gene retention is explored in Chapter 4. This is a preliminary demonstration of the simultaneous effects of expression and functional class. Chapter 5 is a more quantitative assessment of the relative effect of functional category and gene expression on the retention. Finally, in Chapter 6 we summarize our findings and give the concluding statements.

Chapter 2

Ancient eudicot hexaploidy meets ancestral eurosid gene order

Chunfang Zheng, Eric C.H. Chen, Victor A. Albert, Eric Lyons, and David Sankoff

Published: 2013 - BMC genomics 14 (Suppl 7), S3

2.1 Abstract:

Background: A hexaploidation event over 125 Mya underlies the evolutionary lineage of the majority of flowering plants, including very many species of agricultural importance. Half of these belong to the rosid subgrouping, containing several whose genome sequences have been published. Although most duplicate and triplicate genes have been lost in all descendants, clear traces of the original chromosome triples can be discerned,

their internal contiguity highly conserved in some genomes and very fragmented in others. To understand the particular evolutionary patterns of plant genomes, there is a need to systematically survey the fate of the subgenomes of polyploids, including the retention of a small proportion of the duplicate and triplicate genes and the reconstruction of putative ancestral intermediates between the original hexaploid and modern species, in this case the ancestor of the eurosid clade.

Results: We quantitatively trace the fate of gene triples originating in the hexaploidy across seven core eudicot flowering plants, and fit this to a two-stage model, pre- and post-radiation. We also measure the simultaneous dynamics of duplicate orthologous gene loss in three rosids, as influenced by biological functional class. We propose a new protocol for reconstructing ancestral gene order using only gene adjacency data from pairwise genomic analyses, based on repeating MAXIMUM WEIGHT MATCHING at two levels of resolution, an approach designed to transcend limitations on reconstructed contig size, while still avoiding the ambiguities of a multiplicity of solutions. Applied to three high-quality rosid genomes without subsequent polyploidy events, our automated procedure reconstructs the ancestor of the eurosid clade.

Conclusions: The gene loss analysis and the ancestor reconstruction present complementary assessments of post-hexaploidization evolution, the first at the level of individual gene families within and across sister genomes and the second at the chromosome level.

Despite the loss of more than 95% of gene duplicates and triplicates, and despite major structural rearrangement, our reconstructed eurosid ancestor clearly identifies the three regions corresponding to each of the seven original chromosomes of the earlier pre-hexaploid ancestor. Functional analysis confirmed trends reported for more recent plant polyploidy events: genes involved with regulation and responses were retained in multiple copies, while genes involved with metabolic processes were lost.

2.2 Introduction

The publication of the grapevine genome sequence by Jaillon *et al.* in 2007 [12] included the discovery of an ancient hexaploidization event, which also showed clear traces in all the other eudicot genomes sequenced up to that time, but which was absent from the monocot genome of rice. Since then, this event has been confirmed [42] and characterized in most detail with the publication of the cacao genome sequence [43] and in other work of Salse and colleagues [44]. It also has been dated to occur before the radiation of the core eudicots, but after their divergence from more basal eudicots [45]. The original basis for all this work was the observation of seven triples of homeologous regions in each genome. The ancient hexaploid would have combined three identical or closely related subgenomes, each made up of seven chromosomes. In the modern species, each of the three regions (fragmented by post-hexaploid genome rearrangements in some of

the genomes), reflecting the three equivalent chromosomes in the original subgenomes, contains some genes paralogous to genes in one, or occasionally both, of the other two homeologous regions. Meticulous work identified the boundaries of these regions and suggested a credible history of chromosome fusions and other major rearrangements leading to the modern genomes from an ancestral $N = 3 \times 7 = 21$-chromosome hexaploid.

There is a growing literature on gene order reconstruction in the context of flowering plant phylogeny [46–48], although the extensive paralogy induced by hexaploidization, and the subsequent gene order scrambling due to fractionation, i.e., the loss of duplicate genes from one or two of the three homeologous regions, more or less randomly, cause difficulty for the inference of ancestral gene order.

In this paper we follow two separate lines of investigation, the first into the dynamics of fractionation in seven descendants of the hexaploidization and functional constraints on which genes have lost copies, and the second into the reconstruction of an early descendant of the hexaploid, namely the common ancestor of the eurosids. Merging the two approaches combine both an intensional measure – detailed rates of gene loss – and an extensional characterization – changes to chromosomal structure – of the early evolution of the core eudicots.

The seven rosid genomes we analyze include six that have experienced no further polyploidization since their common origin (∼110 Mya): *Vitis vinifera* (grapevine) [12, 49], *Carica papaya* (papaya) [50], *Ricinus communis* (castor bean) [51], *Theobroma cacao*

(cacao) [43] and *Fragaria vesca* (strawberry) [52], *Prunus persica* (peach) [48] as well as one that has undergone a tetraploidization (∼70 Mya) since its lineage diverged from the others: *Populus trichocarpa* (poplar) [53]. Figure 2.1 summarizes the phylogenetic relationships among these species in the context of the core-eudicot clade. We also label four ancestral genomes, the hypothetical 21-chromosome hexaploid ancestor, the rosid ancestor, the eurosid ancestor and the 9-chromosome Rosaceae ancestor reconstructed in [48] and, schematically, in [44].

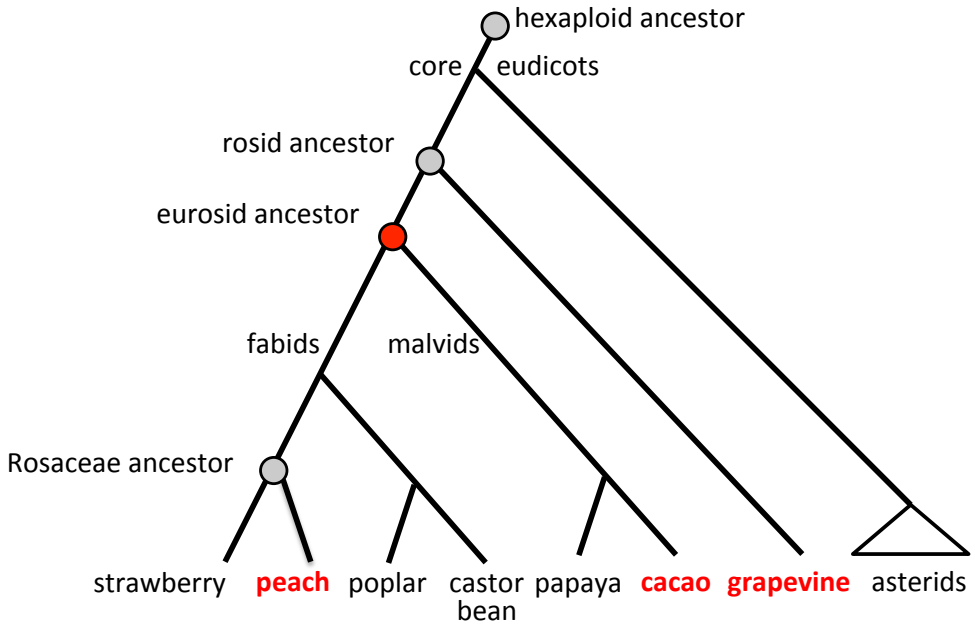


FIGURE 2.1: Lineages of the rosids within the core eudicots. Eurosid ancestor to be reconstructed on the basis of grapevine, cacao and peach, all in red. Only poplar among the genomes shown has had a recent whole genome duplication. All of the published asterid genome sequences and almost all other sequenced rosids not included in this study have undergone one or more polyploidy events since the core eudicot radiation.

For the reconstruction of the eurosid ancestor, we use the grapevine genome, member of the earliest branching rosid order, Vitales, as an outgroup, as well as the two least

rearranged eurosid genomes, cacao and peach. Our entirely automated method for the reconstruction of ancestral gene order

- starts from complete sets of orthologs inferred for pairs of genomes by SynMap in the CoGe platform [36, 54, 55],
- harmonizes them into triples of orthologs across the three genomes by the OMG! technique [37],
- optimally infers ancestral contigs using MAXIMUM WEIGHT MATCHING (MWM) [56] on oriented (same or different DNA strand) gene adjacencies, and then discards any matches not supported by at least two of the genomes,
- replaces all the genes on the genomes by oriented symbols representing the new contigs,
- does another round of MWM on the adjacencies of the new symbols, to form chromosomal fragments,
- merges adjacent fragments if they are not separated by more than a preset number of genes on at least two genomes.

In contrast to genome halving [57] or genome aliquoting [58] methods, we do not attempt to reconstruct the ancestor at the moment of polyploidization. Rather we reconstruct a

presumably rediploidized, partially fractionated, and somewhat rearranged descendant of that polyploid, in an effort to infer recent common ancestor of the extant genomes.

Before this reconstruction of the ancestral genome, which makes no reference whatsoever to hexaploidy or its remnants, we quantitatively analyze the internal paralogies, or homeologies, of the seven genomes, documenting the pervasive pattern of triples of syntenic blocks that are the signature of hexaploidization. We also develop a model for the fractionation process to account for the proportion of gene triples, gene pairs, and single copy genes within each genome. We attribute some of the deviation between model predictions and observations to rate dependence on gene functional class, and undertake to quantify this rate diversity based on known gene ontology in *Arabidopsis* of triplets, duplicates and single-copy of homologous rosid genes.

In the final sections of this paper, we apply the knowledge we have generated about the remnants, in the grapevine, cacao and peach genomes, of the 21 regions produced by the hexaploidization, to try to identify these regions in the ancestral eurosid genome we reconstructed, and find striking coherence between the reconstructed chromosomes and what would be expected from a gross comparison of the genomes at the level of entire regions.

2.3 Gene triplicates resulting from the hexaploidization event

Using SynMap to locate all synteny blocks of minimum size 5 in a self-comparison of each of the seven genomes revealed thousands of syntenic gene pairs, whose sequence similarities are plotted in Figure 2.2. Experience has shown (as reflected in the default values within SynMap) that the threshold of 5 genes anchoring a syntenic block is as low as we can set without admitting large amounts of noise into genomic comparisons. All of the genomes show a clear peak at 70 % ± 3% sequence similarity. In addition, poplar showed a larger peak at 91 %, reflecting its more recent WGD.

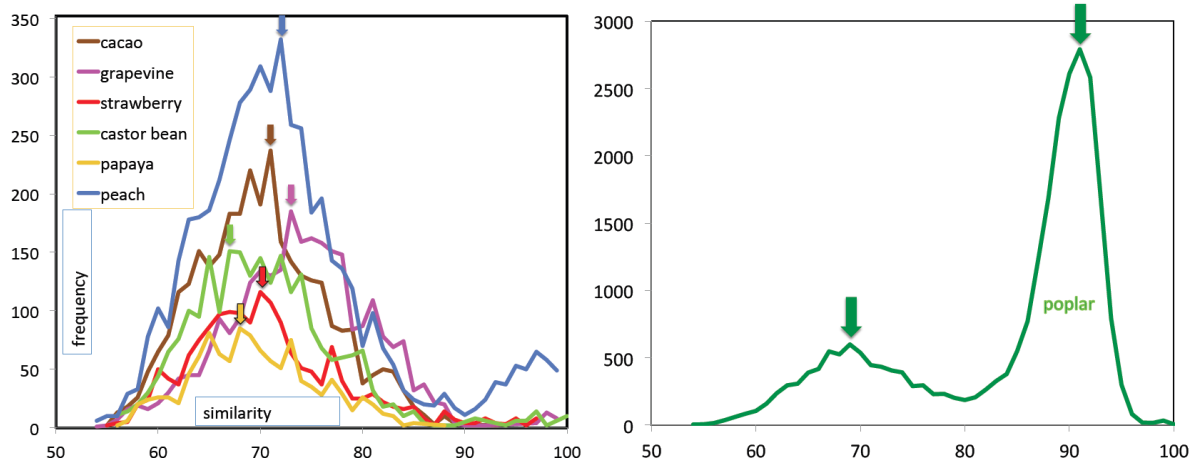


Figure 2.2: Distribution of similarities in gene pairs. Showing triplication peaks between 67 % and 73%. Poplar WGD peak at 91%.

Syntenic dot-plots for the self-comparison of both grapevine and of cacao clearly show a

pattern of pairwise homologies falling largely into 21 groups according to the chromosomal location of the two homologs [12, 43]. This is also true to a lesser extent for peach. Within each of the three genomes, these groups of homologies are distributed among seven triples of large regions, where there are numerous homologies between each of the three *pairs* of regions. There are very few, if any, gene homologies within regions, few between different (non-homeologous) tripled regions, and few between genes in regions and genes outside of all tripled regions.

We adopt the coloring scheme introduced in [12, 43] for the seven triples of regions: red, orange, yellow, green, pale blue, blue and purple. We distinguish between the three homologs of each colour by using dark, medium and light shades of each, ordered according to the number of genes detected in each by our methods, with the darkest containing the most genes.

For grapevine, cacao and peach, the ranges of the triplicated regions were determined by examining the dot plots produced by SynMap. For the other four genomes, the greater degree of chromosomal rearrangement entails more than 21 groups of smaller syntenic regions in the SynMap output. But these regions can still be grouped into triples, with numerous homologies between each of the three pairs of groups of regions and few homologies within groups, between genes in different triples, or between genes in groups and those in no group. Without risk of ambiguity we will refer to these groups

of regions as "regions" as if they were each contiguous regions as in grapevine, cacao and peach.

Figure 2.3 depicts the distribution of the 21 regions created by the hexaploidization event in the 19 chromosomes of grapevine, the 10 chromosomes of cacao and the 8 chromosomes of peach. These regions are slightly extended in each genome by adding genes that are not necessarily adjacent to the original regions, but are identified through their homologies with coloured regions in the other two genomes.

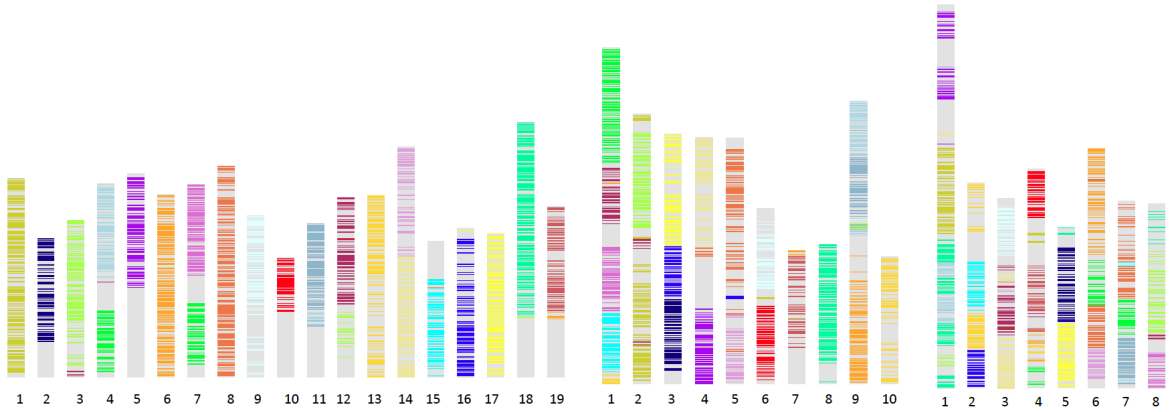


FIGURE 2.3: Triplication in three rosid genomes. Chromosomal distribution of triplicated regions in the grapevine, cacao and peach genomes. Colour intensities represent regions with more (darkest) or fewer (lightest) retained genes. Grey areas represent genes that are not in coloured regions.

Due to rapid fractionation, there are very few *triples* of homologous genes within each genome,as we will discuss in the next section. This loss of iso-functional paralogs results in an *interleaving* [21] pattern of pairwise homologies among three regions.

TABLE 2.1: Distribution of colours among synteny blocks.

	genome						
	peach	cacao	grapevine	castor bean	strawberry	papaya	poplar
genes in synteny blocks	4326	2878	2435	2147	1576	1098	5202
% genes coloured	83	90	88	90	82	90	86
% genes coloured in coloured blocks	87	91	88	92	84	91	88
number of blocks (most counted 2X)	481	362	247	267	225	160	1155
% one colour	79	96	97	93	85	94	74
% no colour	12	1	1	2	4	1	19
% different colour	9	2	2	5	10	4	7
number of gene pairs (most counted 2X)	5121	3143	2709	2347	1638	1110	7273
% in coloured blocks	93	99	100	98	97	98	98
% same colour, different region	74	84	77	83	71	81	75
% same colour, same region	1	0	0	0	1	1	5
% one gene coloured	9	11	21	12	15	12	12
% neither coloured	7	1	0	2	3	2	2

Grapevine, cacao and peach coloured independently; the ranges of the coloured regions determined by inspection of the self-comparison dot-plots; for the first two, these correspond as closely as can be determined to the ranges displayed in [12] and [43]. The other four genomes are coloured according to homologies with the first three. Peach ranges are more difficult to distinguish than those of grapevine and cacao, accounting for the somewhat elevated number of blocks of different colors, and this ambiguity is projected via the coloring process onto strawberry and poplar.

As we can see in Table 2.1, most of the duplicate genes in synteny blocks (produced by SYNMAP) are coloured. For a very large majority of the coloured blocks, there is only one colour. Importantly, all of the gene pairs with two coloured genes involve two separate regions (different shades in Figure 2.3) of the same colour. These observations can only be explained by a hexaploidization event being the origin of most of the gene pairs within these eudicot genomes.

2.4 Fractionation of duplicates and triplicates

The number of genes in the common ancestor of the seven genomes is unknown; the suggested 10,000 genes in the pre-hexaploidy genome [43] (30,000 in the hexaploid) is likely a severe underestimate; plant genomes generally have more than 20,000. Between hexaploidization (time t_0) and radiation of the core eudicots (time t_1), many duplicates would have been lost. After radiation, loss would have continued *independently* in each lineage. Table 2.2 reports the number of triplets of homologous genes remaining today.

Let p represent the probability that a redundant (duplicate or triplicate) gene would be lost during the time span from t_0 to t_1, from hexaploidization to the radiation, had there been no functional constraints. However, we can assume that the event where all three copies were lost was prohibited. Adapting a derivation for a different scenario of compound polyploidization [59], the probability that:

1. all three genes survived is $\frac{(1-p)^3}{1-p^3}$.
2. two of the three survived is $\frac{3p(1-p)^2}{1-p^3}$, and
3. only one survived is $\frac{3p^2(1-p)}{1-p^3}$.

Similarly, let q_i represent the probability that a redundant (duplicate or triplicate) gene would be lost from species i during the time span from t_1 to t_2, from the radiation to the present, were there no functional constraints. Then the probability that:

TABLE 2.2: Gene family sizes.

	frequencies of gene family sizes						
size	peach	cacao	grapevine	castor bean	strawberry	papaya	poplar
2	1484	1111	945	851	606	474	1278 (3-4)
3	256	172	150	119	57	34	177 (5-6)
≥ 4	21	5	12	2	9	1	16 (> 6)

Numbers of triples and pairs after fractionation

4. a triplet would still survive is $\frac{(1-p)^3}{1-p^3}\frac{(1-q_i)^3}{1-q_i^3}$.

5. an original triplet would manifest as a pair is $\frac{(1-p)^3}{1-p^3}\frac{3q_i(1-q_i)^2}{1-q_i^3} + \frac{3p(1-p)^2}{1-p^3}\frac{(1-q_i)^2}{1-q_i^2}$, and

6. an original triplet would be reduced to a single copy is $\frac{(1-p)^3}{1-p^3}\frac{3q_i^2(1-q_i)}{1-q_i^3} + \frac{3p(1-p)^2}{1-p^3}\frac{2q_i(1-q_i)}{1-q_i^2} + \frac{3p^2(1-p)}{1-p^3}$.

Using equations 4 and 5 to predict the numbers of triplicates and duplicates, assuming various numbers of original triplets, and maximum likelihood estimates of p and lineage-specific q_i's, produces the curves in Figure 2.4. When the data for the eudicot genomes are plotted, it is clear they behave as if the original number of triplets was much less than 10,000. A partial explanation is that a single parameter p and a single q_i for each lineage is too simple a model. Not only can fractionation proceed at different rates in different lineages, its pace can also differ for different classes of genes [5, 6]. (Cf. [60] for a study of rate inhomogeneities in an amphibian tetraploid.)

Though a general functional classification of all or most genes is not available for any of our rosid genomes, in the next section we initiate a comparative study of fractionation

rates by assuming that genes have the same functions as their *Arabidopsis* homologs, when this homology can be detected.

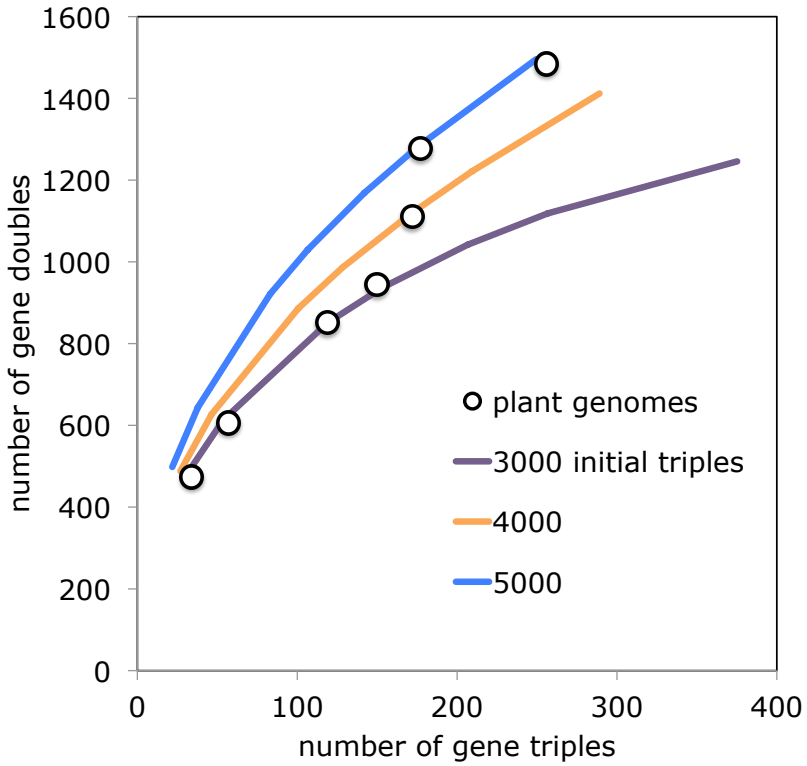


FIGURE 2.4: Model and data. Curves of duplicate-triplicate configurations in model, compared to values for seven rosids.

2.5 Functional analysis

We carried out a search for *Arabidopsis* orthologs for the genes in our grapevine, cacao and peach data by subjected them to the LastZ analysis in CoGe against the TAIR Version 10 database at http://arabidopsis.org/. For each rosid gene we picked the best *Arabidopsis* hit. Only the three genomes were examined because for comparability, our

subsequent analysis required at least one copy of a gene in each genome, and no more than three. We found a total of 7200 such *complete* gene families in the three genomes; including the remaining genomes in the analysis would have drastically diminished this number and compromised the statistics.

Gene ontologies for all the *Arabidopsis* homologs chosen were downloaded from `http://www.geneontology.org/GO.downloads.ontology.shtml`. For 84 % of the 7200 gene families, all the genes mapping to an *Arabidopsis* homolog mapped to the same one. For those families containing members mapping to different homologs, these were virtually always *Arabidopsis* genes having the same functionality.

For the 7200 gene families, we distinguished five categories of multiple-copy gene retention: 1 = gene families consisting of only one copy in all three genomes, 2 = genes that have one copy in two of the genomes and two or three copies in the third genome, 3 = genes that have one copy in one genome and two or three copies in the others, 4 = genes with two copies in each genome, 5 = gene families with at least two copies in each genome, at least one of which contains three copies. Thus genes with lower scores have lost more copies to fractionation and those with higher scores have retained more and are "fractionation-resistant". For each family, we tabulated all the GO terms "hit" by at least one member, and amalgamated these terms for the family. We noted that, not surprisingly, families in higher scoring categories, containing more genes, hit a higher number of terms than families with lower scores. To see if some gene functions were

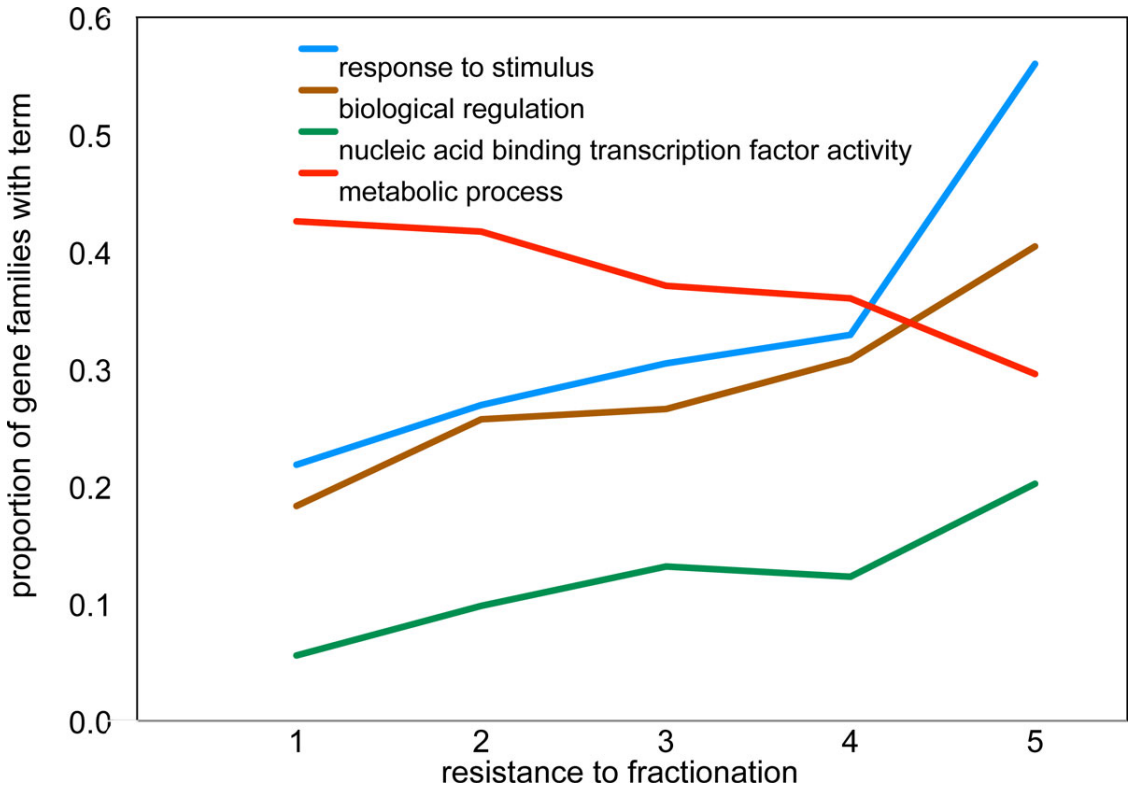


FIGURE 2.5: Fractionation and function. Functional categories of fractionation-resistant and non-resistant gene triplicates.

associated with higher or lower rate of retention, we normalized the proportions of gene families in each retention category showing an association with a term by the proportion of gene families in that retention category showing *any* hits, separately for *Biological function, Cellular component* and *Molecular function*. For example, 3982/5933 = 67% of families in retention category 1 (one copy in each genome) had at least one hit for some *Biological function* term. One of these terms, *Response to stimulus* was hit by 15 % of these families, so that the normalized association value for this GO term and this retention category is 0.15/0.67 = 0.22.

Figure 2.5 shows, out of 67 general functional categories, four with a significant number of hits exhibited a clear trend, either increasing with fractionation retention (or *resistance*) (three of them) or decreasing (one). The association of retention with regulatory and response terms has been shown in connection with more recent polyploidization events [5, 6], as is the tendency for loss of duplicate genes in some metabolic families. Of interest is that this latter decreasing trend is in contrast with the GO term for *Regulation of metabolic process*, which is clearly associated with high retention gene families.

2.6 The reconstruction

2.6.1 Matching

The formal structure we adopt, both for the extant rosid genomes used as input data, and for the ancestral eurosid we aim to reconstruct, is the *oriented gene order*. In graph theoretical terms, these are pairs of graph matchings. The set of vertices of the graphs contains the two ends of all the genes, distinguished either as 5' versus 3', *h*(ead) versus *t*(ail) or "+" versus "-". The first matching simply connects every gene end to its other end. (Because every vertex is connected to another vertex, this is a *complete* matching.) The second connects each gene end to at most one other gene end (usually of a different gene), representing the adjacency of two genes. This is usually not a complete matching;

a vertex representing a gene end at the end of a chromosomes is not connected to any other vertex.

2.6.2 The median problem

The problem of reconstructing a unknown gene order on the basis of three given gene orders from related genomes is called the *median problem*. There are many methods for solving this (e.g. [61–64]); they generally attempt to construct a match as similar as possible, optimizing some measure of similarity under various constraints, to the three given adjacency matchings. The methods are distinguished by fine distinctions among the simplified models of genome rearrangement invoked, assumptions about identical gene complement or lack of duplicate genes, insistence or not on linearity of chromosomes or the number of chromosomes in the solution, relative weights of adjacencies versus chromosome-terminal gene ends, and other considerations.

2.6.3 Non-uniqueness and stepwise procedures

Most gene order median problems are difficult: exact algorithms bog down for realistic instances while heuristics risk somewhat sub-optimal solutions. But of much more concern to biologists than details about optimality and efficiency is that the reliance on

global optimality criteria leads to a severe problem of non-uniqueness of solutions. Reconstruction methods based on gene-by-gene adjacencies do not scale well for gene orders involving tens of thousands of genes. Long reconstructed chromosomes inevitably contain many poorly supported adjacencies, i.e., adjacencies that only occur in one, or even none, of the three data genomes. Though these may be relatively infrequent in terms of overall proportions, for example, 1% of 1000 adjacencies, giving 10 per chromosome, this is still a worrisome number; the reconstructions can often be broken at these points and reassembled in different combinations using other hitherto unused adjacencies, without losing the optimality of the solution. As a result there are an exponential number of solutions, many of them differing greatly at the level of chromosome structure. This is a dismaying situation for a biologist.

This partially explains the pervasiveness, in the field of ancestral genome inference, of reconstructions in two or more steps. The first step reconstructs a large number of small "contigs" or anchors, and the subsequent steps pieces these together into larger chromosome fragments or whole chromosomes, using some other objective function or relying on some other data. Often the ultimate step in reconstruction involves manual intervention, invoking biological intuition or some implicit expectations about the final shape of the solution.

2.6.4 Uniqueness as a priority

In this paper, we also use a multi-step process, but our goals are first, to eliminate as far as possible sources of non-uniqueness in the reconstruction. We make sure that nothing is reconstructed on the basis of a series of arbitrary choices among equally plausible alternatives (a recipe for rampant non-uniqueness), and second, that the whole process be automated and reproducible, with explicit thresholds and other parameters.

Step 1: Generating gene orders on sets of orthologous genes

The input to the reconstruction method is the complete gene order for each of the extant genomes, together with identification of orthologous genes in the different genomes. We construct our gene orders based on SynMap's pairwise comparisons of gene orders: grapevine-cacao, cacao-peach and peach-grapevine, using its rigorous default values to ensure that orthologs are identified not only by their sequence similarities, but also by severe requirements on their syntenic context (i.e., syntenic blocks are identified only if they contain at least 5 pairs of collinear orthologous genes). We impose transitivity on the results where this does not already obtain, e.g., if a is orthologous to b and b is orthologous to c, then a is assumed orthologous to c whether or not this has been detected by SynMap.

For the grapevine, cacao and peach genomes, this initially produced a total of 14,543 orthology sets containing two, three or more genes, for a total of 39,286 genes.

Step 2: Resolving paralogy

The requirement for synteny within SynMap is an effective way of resolving non-tandem paralogy in large majority of instances. Two paralogs in different syntenic contexts are treated separately as different genes by virtue of these contextual distinctions.

Paralogs still show up, however. Genes a_1 and a_2 in genome A and can both register as orthologous to gene b in genome B because of segmental duplication in A or because identical contexts of a_1 and a_2 survived the fractionation process intact since the early polyploidization process. a_1 and a_2 are thus paralogs.

We use the OMG! procedure [37] to judiciously suppress a small number of homology relations to resolve these situations, usually by dividing large orthology sets into smaller ones, allowing only three or two genes to be considered orthologous across the three genomes. The objective is to maximize the sum of the squares of the number of genes in remaining orthology sets (9 or 4, respectively, for three-way and two-way orthology).

For the grapevine, cacao and peach genomes, this produced a final total of 5664 orthology sets containing two genes and 9148 containing three genes, for a total of 14,812 sets containing 38,772 genes.

Step 3: Maximum Weight Matching

Because we will be weighting adjacencies differentially, we do not use any of the dedicated median solvers, but the more flexible approach of Maximum Weight Matching (MWM). This allows us to make a slight weight distinction between two conflicting matches that appear equally often in the three genomes, but in subtly different contexts. We solve the MWM problem on the three sets of adjacencies, one for each genome, determined by the order of the genes output by SynMap and OMG!. The first polynomial-time exact algorithm for MWM — "path, trees and flower" — was introduced by Edmonds [65]. We use Galil's $O(n^3)$ version of the algorithm [56] for which Python code by J. van Rantwijk is available online.

The weighting system we use is displayed in Table 2.3. The first subtlety we incorporate is that if a match is present in a certain number of genomes, the evidence for that match is stronger if one or both of the vertices are absent from the other genomes than if they are both in conflicting adjacencies in those genomes. The second subtlety is that if a match is present in one genome only, but it appears that it is absent from another only because there has been a chromosomal inversion disrupting it, then it should have a higher weight than other single-genome matches.

In the output from the algorithm, the 14,812 orthology sets were organized into 63 contigs. These included all the adjacencies in the input with weights 3.00 (5657 of them),

2.03 (6319), 2.00 (1109) or 1.50 (113), and all except one of weight 1.49 (31 of them). In addition there were 1520 of adjacencies of weight 1.01, 1.00 or 0.99.

Step 4: Eliminating sources of non-uniqueness

We know that it is the adjacencies of weight less than 2.00 that propagate non-uniqueness. Those with weight 2.00 or more are very likely to be in all solutions. In our particular data this extends to the adjacencies of weight 1.50 and almost all of weight 1.49.

If we discard all poorly supported adjacencies from the (MWM) solution, then, there can be little variability in the solutions. This, however, comes at a cost of increasing the number of contigs from 63 to 1583, which must then be assembled into chromosomes by other means.

One incidental benefit of discarding most low-weight adjacencies is that all the circular contigs among the 63 are broken into linear fragments. It is theoretically possible to have circular contigs containing only weight 2.00 adjacencies, but this is not encountered in our data.

In addition, we discard 1368 genes scattered among 934 short contigs (three adjacencies or less). The large majority of these have only one gene, and are likely to represent errors

originating in SynMap or OMG!, or movements of very small genomic fragments, containing little information about the main evolutionary events in these genomes. They also represent possible sources of non-uniqueness in reconstruction of ancestral chromosomes.

Note that had we attempted to use any of the other sequenced rosids instead of, or in addition to grapevine, cacao and peach, few additional contigs could be expected, because the highly fragmented nature of the subgenomes in these genomes would have resulted in mostly short contigs that would only have been discarded; some existing contigs could have received additional support, but since all except the weight 1.00 adjacencies are already consistent among themselves, this would have been of little benefit.

Step 5: The summary genomes

We use the 649 remaining contigs to create "summary" versions of the original three genomes. Every contig in the solution to the MWM projects a coordinate position and an orientation back to a single chromosome in each of the grapevine, cacao and peach genomes (even though the contig may contain a minority of genes from another chromosome), based on the average coordinates and most frequent orientation of the genes it contains from that genome on that particular chromosome.

Step 6: MWM on summary genomes

We can thus recognize adjacencies between these contig projections in each summary genome and carry out MWM, using the same weights as in Table 2.3 in Step 3 and the deletions in Step 4 to construct a new median consisting of a matching on the contigs. Again we discard adjacencies supported by only one of the three genomes. There remain 177 new contigs.

TABLE 2.3: Weight system for adjacencies.

occurrence of adjacency between gene ends x and y	weight of xy
xy occurs in all three genomes	3.00
xy occurs in two genomes; x or y is absent from the third genome	2.03
xy occurs in two genomes; x and y present in the third genome	2.00
xy and uv occur in one genome; xu and yv in another; uv, but not both x or y, in third	1.50
xy and uv occur in one genome; xu and yv in another; x, y and uv in third	1.49
xy occurs in one genome; one of x or y absent in each of the other genomes	1.01
xy occurs in one genome; x and y both present in one or neither of the other genomes	1.00
xy occurs in one genome; x and y present in other two genomes	0.99

Step 7: Merging contigs

We can then merge new contigs with neighboring projections in two or more genomes, according to a criterion of how many genes may intervene between two candidates for merger. For each pair of new contigs on the same chromosomes in at least two genomes, we do the following:

- decompose each contig into pieces that are internally strictly contiguous on the two genomes,
- find the two pieces, one from each new contig, that are the closest, when appropriately oriented.

We then merge the pair of new contigs that are the closest, as long as they are not more than 500 genes apart on either of the two genomes. The cutoff in this final step, 500 with our data, is critical, but is easy to find since it has a direct and dramatic effect on the result. Lower values, like 400, result in an unrealistically large number of small chromosomes; higher values like 600 produce assemblies including one or two impossibly long chromosomes.

Run times

The pipeline we have described is not computationally intensive. Data preparation Steps 1 and 2 required 1 to 2 minutes in total for the rosid data. The MWM run in Step 3 took 15 minutes. The data manipulations in Steps 4 and 5 required less than a minute. The second MWM executed in 2 seconds.

2.6.5 Validation

The above procedure results a total of 13 chromosomes, as shown in Figure 2.6, where each gene is coloured by its triplication regions in grapevine, cacao and peach.

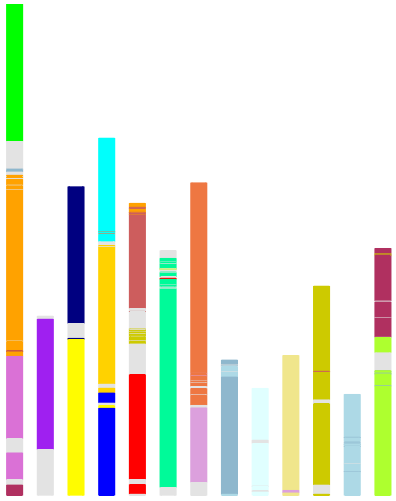


FIGURE 2.6: Ancestral genome. Distribution of triplicated regions in the eurosid ancestor.

This shows that despite the fact that no triplication information whatsoever was used in the reconstruction, most of the 21 triplicated regions are located solely or largely on a single chromosome in the reconstruction. Moreover, as we shall see in the next section, the general structure of the chromosomes corresponds closely to what we would expect from common patterns of region fusions in the three genome.

2.7 Reconstruction by shared fusions

Inspection of the coloured karyotypes in Figure 2.3 is suggestive of some of the evolutionary events which must have transpired in the rosids, without need to reference the individual genes that make up the triplicate regions, in the style of [43, 44]. The main principles we adopt are first, that these genomes are descendants of a hexaploid, and second, that syntenic segments from two ancestral chromosomes on a single modern chromosome, when observed in two or three of the sequenced genomes, is unlikely to have been produced by two or three coincidental events, but is rather most parsimoniously accounted for by a single event at some point in the common evolutionary history of these genomes. This ties in with our search for uniqueness of reconstructions.

This analysis consists of the following hypotheses, observations and inferences, which are schematized on Figure 2.7.

1. The ancestral hexaploid is assumed to contain three identical sets of seven chromosomes, which we label with seven colors, each in three shades from darkest to lightest.

2. All three sequenced genomes have a large light green region adjacent to, or very close to, a small dark red region on the same chromosome, as well as another large dark red region elsewhere. This suggests a non-reciprocal translocation occurring on the tree branch between the hexaploid ancestor and the rosid ancestor, and

accounts for one of the fused chromosomes depicted in the latter. There is also a small light green region at the end of the dark red chromosome in grapevine. This is most likely the result of a chromosomal fission in the tree branch leading to grapevine.

3. There are two chromosomes containing medium green in grapevine. This would have required a fission in the lineage leading to grapevine, but this must have been preceded by a fusion of medium green with medium purple in the common ancestor of grapevine and cacao, since this association is present in both. This accounts for another of the fused chromosomes apparent in the rosid ancestor.

4. All genomes show adjacencies between medium pale blue and dark pale blue, suggesting a fusion in the rosid ancestor as shown, with a further translocation with medium green in grapevine and additional rearrangements in peach.

5. Grapevine and cacao evidence an early fusion of medium blue and medium yellow (the remaining fused chromosome visible in the rosid ancestor), followed by a fission of most of the medium yellow to form a separate chromosome in grapevine.

6. There is an association of dark blue and medium yellow in the two eurosid genomes only, suggesting this event occurred after the divergence from Vitales. However, the inherited association with medium blue is retained in cacao, so that a three-way fusion is is portrayed in the eurosid ancestor. Other exclusively eurosid associations

include fusions of light purple and dark orange, and of light blue and light orange, with much of the light orange fissioning to form a separate chromosome in cacao. In addition, the medium green/medium purple fusion inherited from the rosid ancestor has acquired an association with dark red in the eurosids, so that a three-way fusion seems appropriate there.

These considerations suggest, as summarized in Figure 2.7, that the rosid ancestor had 18 chromosomes, four resulting from fusions among the original hexaploid chromosomes, and 14 consisting of entire chromosomes or parts of single chromosomes from the latter. Grapevine would have undergone two additional fusions and two fissions (at new breakpoints) of chromosomes previously fused in the rosid ancestor.

Similarly the eurosid ancestor can be predicted to have at most 15 chromosomes, derived by four additional fusions and one fission, from which the 10-chromosome cacao and 8-chromosome peach evolved, mostly by further fusions.

2.8 A comparison of gene-based and region-based reconstructions

Although the exercise in the previous section takes no account of the actual gene-order evidence, it does provide a plausibility check on the detailed reconstruction leading to

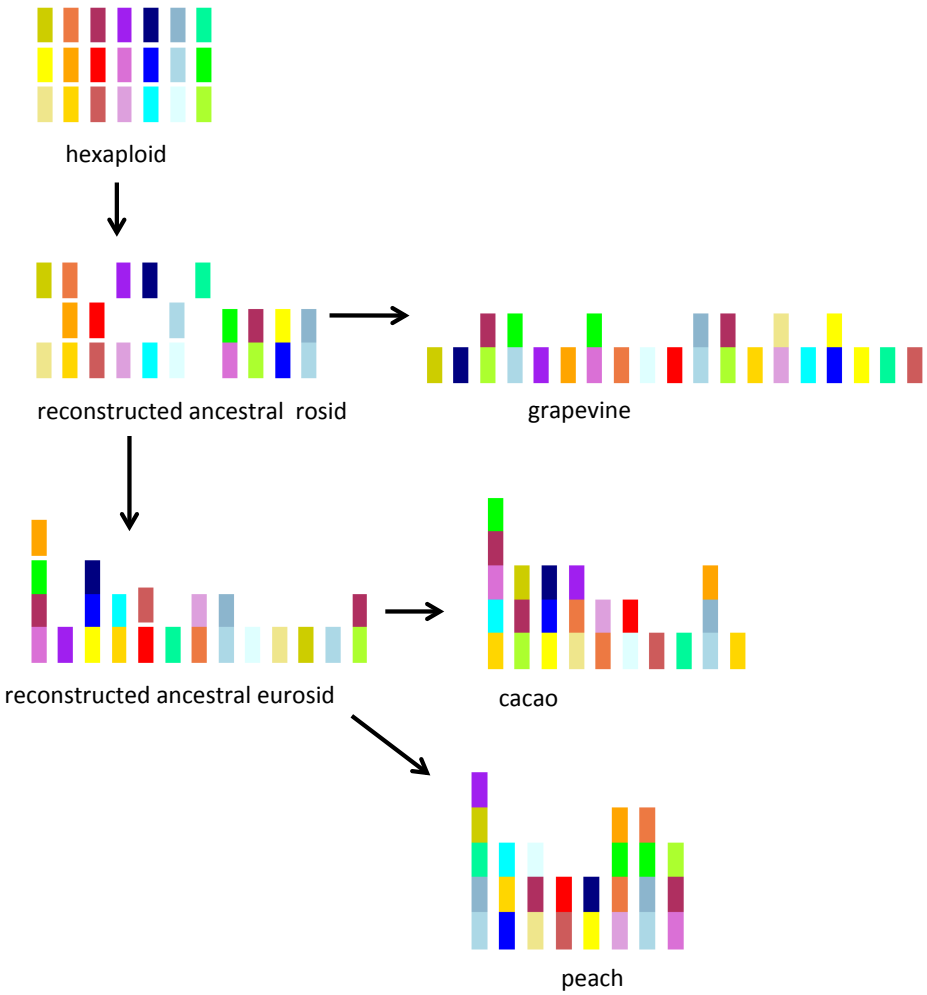


FIGURE 2.7: From ancestors to modern genomes. Schematic attribution of triplicated regions in the rosid and eurosid ancestors.

Figure 2.6. The discrepancy between the 13-chromosome reconstruction in Figure 2.6 and the 15-chromosome version in Figure 2.7 can be accounted for in terms of the fusion of the light orange region with the medium green/medium purple/dark red chromosome in the detailed reconstruction and the fusion of two other red chromosomes. In addition, there is a translocation of the medium blue from its position in cacao (Figure 2.7) to its position in peach (Figure 2.6). The compositions of the four chromosomes affected

in Figure 2.6 are otherwise unchanged, and the remaining nine are identical in the two approaches.

We can conclude that, with allowances for its superficial nature, the "top-bottom" derivation in the previous section gives results which accord well with the detailed reconstruction earlier in this paper. The sole problem is the position of the medium blue region, whose positions in the rosid genomes can only be explained by two coincidental fusions (in cacao and grapevine) or by an unlikely translocation of exactly this region, no more and no less, to another chromosome in peach. Of interest is that in the reconstruction of the 9-chromosome Rosaceae ancestor by Jung *et al.* [48] based on grapevine, peach and strawberry, two of reconstructed chromosomes have the same overall structure as peach 2 and 5; this does not resolve the quandary, but at least suggests that the evolutionary events involved predate the Rosaceae radiation.

2.9 Conclusions

As a prelude to ancestral genome reconstruction, we have quantitatively documented the details of the triplicated regions in the seven rosid eudicots under study, and suggested a model for fractionation in two steps, one pre-radiation and the other post-radiation. Fitting the model to the data from seven rosids requires the assumption of an unrealistically low number of genes involved in the initial hexaploidization, suggesting a dependence of

fractionation rate on gene class, an effect that is confirmed by differential functional associations of highly fractionated versus highly retained gene triples in grapevine, cacao and peach.

A major preoccupation of our reconstruction methodology was to avoid the non-uniqueness of optimal solutions, a problem endemic to one-step reconstructions based even partly on poorly supported adjacencies. The key to this is to discard all such adjacencies from the MWM solution to the median problem. We rely instead on a second application of the algorithm to summary genomes made up of projections back from the high-confidence contigs produced by the initial run.

Our reconstruction procedure requires three parameter settings. One is the weight system, which on one hand distinguishes between conflicting evidence and absent evidence against adjacencies with the same level of support, and on the other hand detects a certain number of poorly supported adjacencies whose contexts in the three genomes suggest that they nevertheless merit inclusion in a solution. The second parameter is a threshold for contig sizes (more than three adjacencies), designed to eliminate noise, due either to errors or insensitivity in the data-generation steps, or to movements of isolated genes to remote locations in the genome. Despite a general tendency to conserve the large triplicated regions, there is much low-level rearrangement occurring in these genomes. They each require from 400 to 600 rearrangements to be generated from the reconstructed ancestral gene order, though much of this is ascribable to fractionation; inevitably a few

of these rearrangements conspire in two genomes to mislead the reconstruction, especially if we place too much faith in the significance of contigs containing only one or two genes. The third setting is the cutoff for merging neighboring fragments after the higher-level run of MWM. In our data, the appropriate value empirically obtained was 500 genes; 400 was too low, unable to detect several valid fusions, while 600 was too high, resulting in an unbalanced genome concentrated largely in a single chromosome. Of the three parameter settings, the first two should be generally applicable, while the third will have to be adjusted (in an automated way if desired, using chromosome number or other karyotype characteristic) to each data set. At least for our data set, all of these settings are indispensable. Not dropping small contigs after the first MWM leads to many unlikely fusions as well as numerous tiny chromosomes that do not fit in with the rest. Using simply weights 1, 2 and 3 instead of those in Table 2.3, or not using the right cutoff lead, both lead to unbalanced genomes and other anomalies.

As more angiosperm genome sequences become available, reconstruction of additional early ancestral genomes will become possible, adding to the Rosaceae ancestor in [48], and the eurosid ancestor studied here, to further elucidate the evolution of large clades of these plants. Additionally, genomes that break up the phylogenetic clades studied here will help resolve conflicting evidence of chromosome fusion and fission events. Paradoxically, although the whole genome doubling (or tripling) that recurs throughout this evolutionary domain has the effect of scrambling local gene order through fractionation,

it also provides a convincing way of validating reconstructions based on gene order, as exemplified by the intact triplicated regions inferred in Figure 2.6.

Authors' contributions

CZ and DS did the research and wrote the paper. EC carried out the functional analysis. VA and EL provided motivation, suggestions and interpretations throughout the project. All authors read and approved the manuscript.

Chapter 3

The dynamics of functional classes of plant genes in rediploidized ancient polyploids

Eric C. H. Chen, Carlos Buen Abad Najar, Chunfang Zheng, Alex Brandts, Eric Lyons, Haibao Tang, Lorenzo Carretero-Paulet, Victor A. Albert and David Sankoff

Published 2013 - BMC bioinformatics 14 (Suppl 15), S19

3.1 Abstract

Introduction: To understand the particular evolutionary patterns of plant genomes, there is a need to systematically survey the fate of the subgenomes of polyploids fixed

as whole genome duplicates, including patterns of retention of duplicate, triplicate, etc. genes.

Results: We measure the simultaneous dynamics of duplicate orthologous gene loss in rosids, in asterids, and in monocots, as influenced by biological functional class. This pan-angiosperm view confirms common tendencies and consistency through time for both ancient and more recent whole genome polyploidization events.

Conclusions: The gene loss analysis represents an assessment of post-polyploidization evolution, at the level of individual gene families within and across sister genomes. Functional analysis confirms universal trends previously reported for more recent plant polyploidy events: genes involved with regulation and responses were retained in multiple copies, while genes involved with metabolic and catalytic processes tended to lose copies, across all three groups of plants.

3.2 Introduction

In whole genome duplication (WGD) or triplication, the entire gene and chromosome structure of a genome undergoes polyploidization followed by rediploidization of the new larger genome and fractionation, or homeologous gene loss, of many or most of the duplicate genes. The doubling or tripling of all the gene contents of a genome has been

hypothesized as an important source of gene innovations and the radiation of species. The inherent variability in these processes may create a variety of beneficial phenotypes similar to those seen due to heterosis [66], and the increase in the diversity of genetic elements may help drive long-term morphological complexity and adaptation to new environments [67].

Studies in a variety of organisms have provided evidence that gene retention during fractionation may differ among functional categories. This has given rise to a number of explanatory models, particularly a reformulation of the classical genetic Gene Balance Hypothesis [68], all of which attempt to explain functional bias in gene retention after WGD.

In a previous study [69] and Chapter 2 [38], we proposed a comparative genome-wide analysis of the descendants of triplicated genes in the ancestor of the core eudicots, focusing on three rosid plants that have not undergone any subsequent whole genome duplication: peach (*Prunus persica*) [29], cacao (*Theobroma cacao*) [43], and grape (*Vitis vinifera*) [12]. We asked whether the genes that have been retained in three or two paralogous copies could be seen to be enriched for certain functional categories. These results, updated to reflect the current study, are illustrated in Figure 3.1.

These results, such as the relatively rapid fractionation of genes labelled "metabolic", and the resistance to fractionation of those labelled "biological regulation", confirmed various accounts in the literature [6, 26, 28, 70].

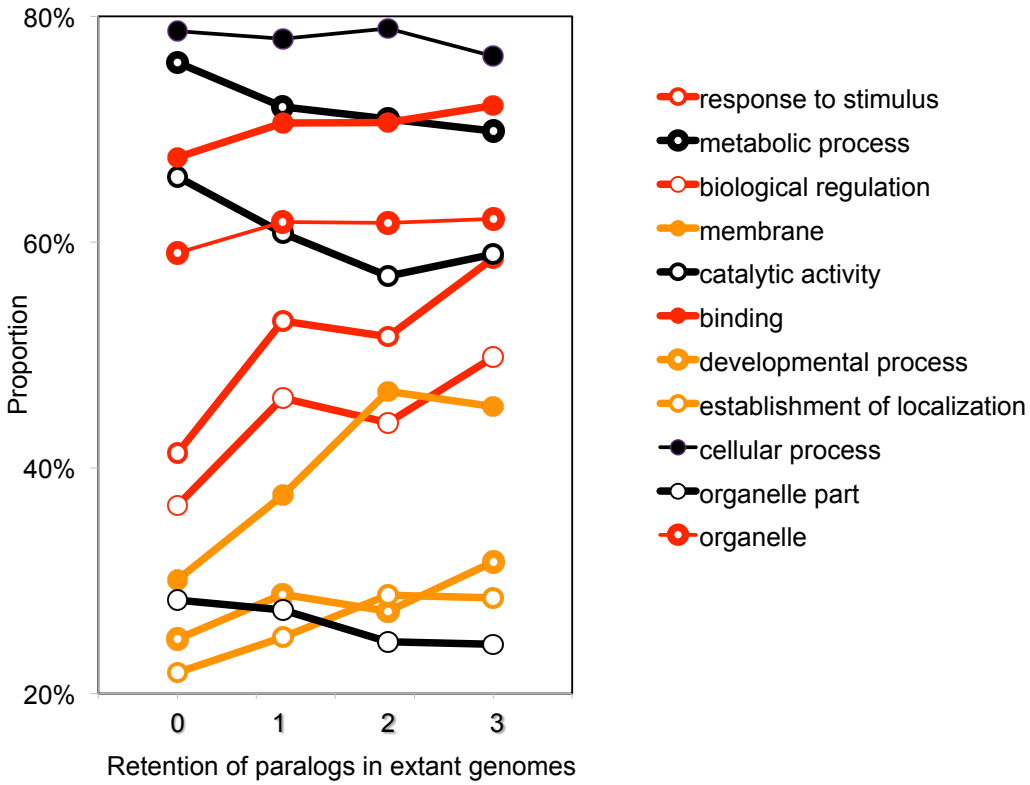


FIGURE 3.1: Proportion of genes with various numbers of extant copies annotated with given high-level functional categorizations. x-axis: number of genomes (out of three: peach, cacao and grape) with more than one copy of a given gene. Score 0 indicates all three genomes contain exactly one copy of the gene; 1 indicates that one of the genomes has two or three copies, 2 means that two of the genomes have two or three copies and 3 means that all three genomes have two or three copies. y-axis: proportion of these gene sets with the indicated annotation. Positive slope indicates fractionation resistance, negative slope indicates fractionation-prone genes. Thick lines indicate statistical significance of non-zero slope.

This work raised new questions. The first is to what extent the patterns in Figure 3.1 are specific to rosids. Which functional categories have the same tendencies in other flowering plant groups? The second concern has to do with the severe restriction on sets of orthologs and paralogs we included in our data – we required that each rosid manifest at least one copy of the gene in a common syntenic context when compared to the other

rosids, and the same for paralogs within each genome. This requirement attempted to isolate the process of fractionation from other processes of gene family dynamics, such as gene movement and gene family expansion, by focusing on the "natural experiment" created by the core eudicot triplication, whereby each species started with three copies of the same gene in the three identical syntenic contexts, and each then retained either one, two or three of these copies over time. A working assumption was that fractionation would be the dominant process affecting these particular sets of genes. The requirement, however, reduced the total number of "homology sets" considered to about a third of the number of genes in the extant genomes. Do our results therefore pertain only to genes that are syntenically conserved and comprise the "stable genome" for functional reasons, or are they also valid for some additional class of paralog sets, some of whose elements are not detectible using syntenic criteria? The third and final problem is whether the results we traced on high-level functional categories indicate general tendencies within these categories, or simply reflect the preponderance of some subcategories among many others with diverse, and perhaps contrary fractionation and positional stability behaviour.

In this paper we address these three concerns. First, we replicate our original studies on two other groups of plants, one consisting of three lamiid asterid species: tomato (*Solanum lycopersicum*) [30], humped bladderwort (*Utricularia gibba*) [14], and monkey flower (*Mimulus guttatus*) [31], the other, four Poaceae monocot species: rice (*Oryza*

saliva) [33], foxtail millet (*Setaria italica*) [34], sorghum (*Sorghum bicolor*) [32], and purple false brome (*Brachypodium distachyon*) [35], thus compiling the first pan-angiosperm study focused on gene fractionation patterns. The asterids have all individually undergone further WGD since the core eudicot whole genome triplication they share with the rosids. Nevertheless, the principles underlying functionally-influenced fractionation can be assumed to hold just as well after several WGD as after one. The divergence of the asterids from the rosids seems to have occurred within five or ten million years after the triplication event they shared approximately 120 million years ago [71], so that 90 - 95% of their evolution, including much of their fractionation, took place independently in the two groups. The monocot species in our study all share two or three WGD independent of the all polyploidy events in the eudicots. We will show that the independent fractionation patterns of all three diverse groups of angiosperms are highly parallel, following the same patterns of fractionation based on gene functional classes.

Second, we investigate the hypothesis that our results only apply to a non-mobile core of genes that are detected by the syntenic context they share with orthologs in all the genomes in the group. We relax somewhat the requirement that all the extant genomes must contain at least one syntenically validated descendant of the ancestral gene. Even though the additional cohort of genes, part of the more "mobile" genome that is prone to translocation in a genome, is less numerous and thus less conducive to statistical significance of the results, we find that they continue to support the tendencies found

for the functionally "stabilized" part of the genome. We demonstrate that the homology sets we study are enriched for some categories and depleted for others, in comparison with random samples of genes from the entire extant genomes, but hypothesize that this has more to do with lineage-specific expansion of gene families, rather than fractionation dynamics.

Finally, we search for tendencies within subcategories of some of the large functional categories. We find that the negative slope within the catalytic activity category reflects in large measure the consistent behavior across certain classes of enzymes. We resolve the apparent conflict between the trends for the high-level terms "organelle" and "organelle part" seen in Figure 3.1 by de-convoluting lower level terms contained therein.

3.3 Methods and Data

We compared the retention of homologs in six core eudicot species, namely three rosids: peach, cacao and grape, and three asterids: tomato, *Utricularia* and *Mimulus*, and four monocot species: rice, *Setaria*, sorghum and *Brachypodium*, forming three independent data sets.

The data preparation in our approach, illustrated in Figure 3.2, starts with applying the SynMap program in the CoGe platform [36, 54] to selected pairs of genomes stored on the CoGe site. This produces synteny blocks of genes (five or more, in the present work)

likely to be orthologs because they have high sequence similarity and are in the same syntenic context. This includes paralogous genes syntenically mapping to the same ortholog(s). Additional syntenic paralogs derived from polyploidy can be detected through SynMap self-comparisons of genomes. All the genes sharing orthologies and paralogies thus detected, among all the species in each data set are then grouped together yielding "homology sets" representing ancestral pre-WGD genes [72].

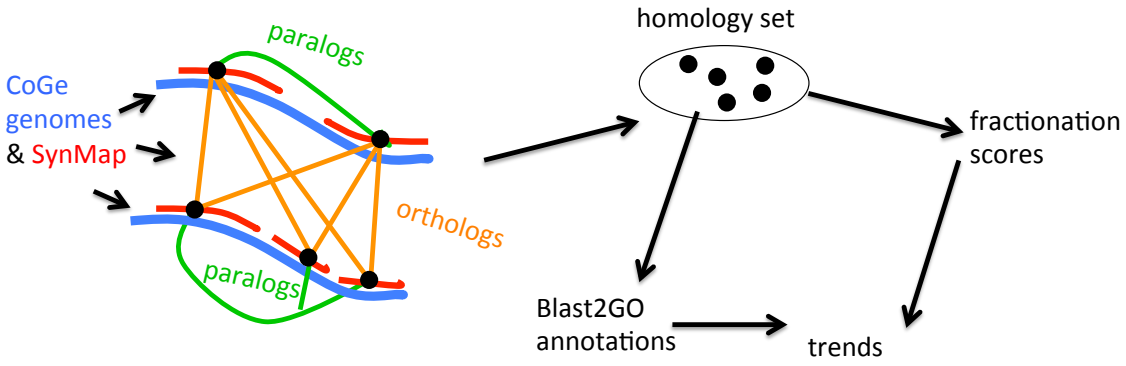


FIGURE 3.2: Pipeline detects functional determinants of fractionation rate. Blue lines represent chromosomes in two genomes; red lines all represent similar syntenic contexts.

The homology sets were first examined to see whether they contained at least one gene from each species in the group. In the first analysis, all sets with no gene in any of these species were excluded from the analysis. (In a later analysis, described below, other homology sets were used.) The remaining sets were classified according to the number of species in which there was more than one copy, so that in the three-species comparisons, the sets could be classified as 0, 1, 2, or 3, and in the four-species set a score of 4 was also possible. We call this number the *fractionation score*.

For each homology set, each of its genes was annotated by submitting it to Blast2GO [40]. Then all the annotations from all the genes in this set were considered as annotations for the set as a whole. No account was taken of the multiplicity of "hits" of a single annotation within the set. Of course, for every annotation, each of the higher-level terms of each hit was also counted as an annotation.

Among all the homology sets we constructed, approximately 90% hit at least one GO term, resulting in 10,688 monocot, 6360 rosid and 4638 asterid homology sets for further analysis.

The GO terms are divided at the highest level into "Biological Process", "Molecular Function" and "Cellular Component" and there are a further 67 terms at the next level, which we call "high-level terms". Homology sets with large fractionation scores, i.e., which contain more than one paralog in all or most genomes, tend to have a higher total number of annotations, simply by virtue of having a larger number of genes. This leads to the artifactual observation that almost all functional categories are more favored by homology sets with high fractionation scores. To correct for this bias, we use a normalized proportion of hits for each term for each fractionation score. This is calculated as the number of hits of the term over all homology sets with this fractionation score, divided by the total number of sets with hits for any terms within the appropriate highest-level term. Thus, if "organelle" received 100 hits in all homology sets with fractionation score

3, and if the number of sets hitting any "Cellular Component" term is 300, the normalized "proportion" is 33.3%.

These normalized proportions could then be plotted against fractionation score as in Figure 3.1. By considering every combination of homology set and functional category as a data point with X-coordinate its fractionation score and its Y-coordinate 1 or 0, depending on whether the homology set was a hit (1) or not (0) for that category, we could then calculate a regression score for the functional category. In Figure 3.1, the functional categories with significant negative slopes are black and those with significant positive score are red or orange.

3.4 Trends for high-level categories

In comparing the fractionation patterns of the three groups of species, we looked for any trends that were statistically significant in at least one of the three (preferably all three) and with similar slopes in all three. Of the 67 high-level terms in the GO hierarchy, eleven that satisfied these conditions are illustrated in Figure 3.3. Another 19 terms also satisfied the conditions but involved numbers of homology sets too small to be informative on the figure. Only three terms, or 5%, were significant in opposite directions in two groups of species.

Thus, the three patterns are surprisingly parallel. Recall that the monocot ancestor and its WGD occurred in a completely different part of the flowering plant phylogeny from the eudicots. And though the two groups of core eudicots descend from the same triplication event, the common ancestor of the asterid lamiids evolved for a lengthy period before radiating into the present-day families. Similarly for the common ancestor of peach and cacao, although grape belongs to an early branching rosid order.

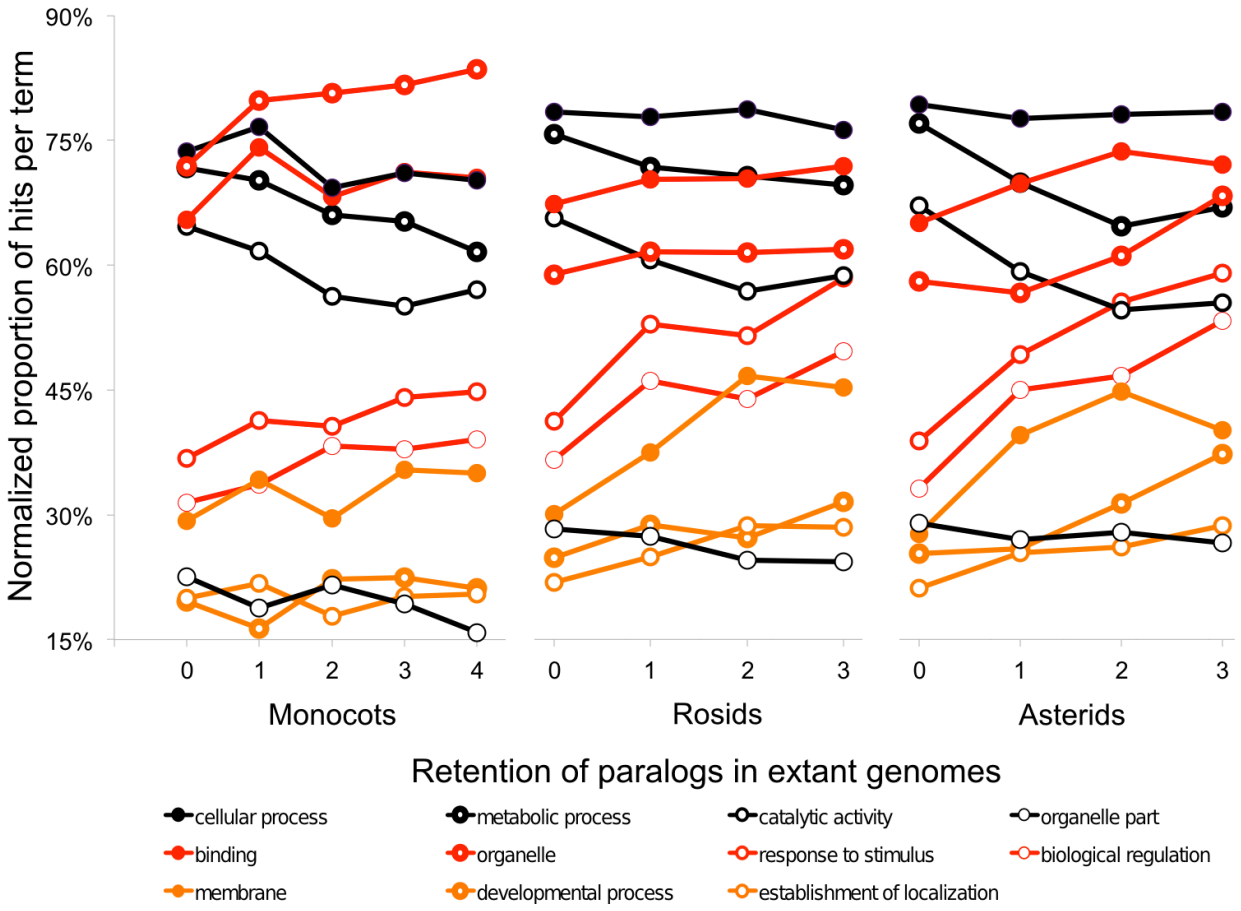


FIGURE 3.3: Fractionation patterns of functional groups in three groups of flowering plants. Axes as in Figure 3.1, except that monocots include four genomes, where rosids and asterids include three each.

The Gene Balance Hypothesis predicts that genes involved in multi-unit protein complexes, or genes involved in cascades in which the downstream genes are involved in multi-unit protein complexes, are more likely to be retained after whole genome duplication. This prediction can in turn predict what GO terms are expected to be fractionation resistant or fractionation prone. The predicted terms are largely the same as the significant in Figure 3.3, especially "binding", "biological regulation", "response to stimulus" and "developmental process".

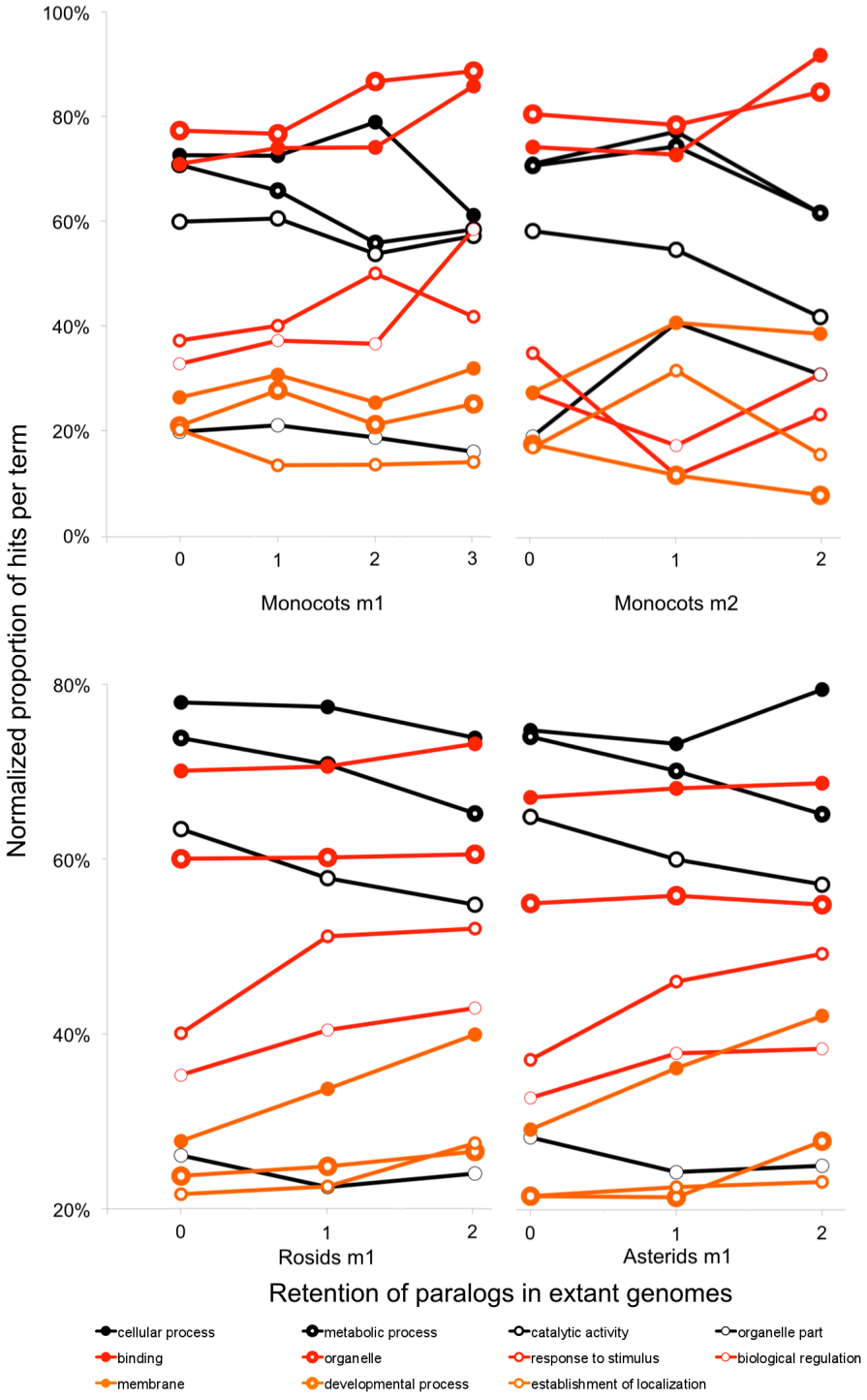
Of the significant terms, two stand out as being fractionation prone: "metabolic process" and "catalytic activity". These two terms describe processes or functions that, as traditionally conceived, involve the interactions of single enzymes with substrates, coenzymes and cofactors that are not themselves proteins. This is in contrast with such processes as gene regulation, which explicitly involve the stoichiometry of two or more distinct proteins. On this basis, the Gene Balance Hypothesis would predict that "metabolic process" and "catalytic activity" be fractionation prone. There is much recent enthusiasm, however, for protein complexes in metabolic and other catalytic reactions, for metabolons [73], protein heteromers and other potential sources of abundance constraints among different proteins, so such predictions might have to be attenuated according to how widespread and how constraining these structures turn out to be (cf. [74, 75] as examples of more moderate opinions).

3.5 The "mobile" genome

All homology sets involved in the above analysis contain at least one gene from each species To test whether this requirement biases the analysis towards genes with some special functional properties, we also carried out our analysis on homology sets where no ortholog was detected in one of the genomes. The absence of this gene from a synteny block does not necessarily mean that it is absent from the genome; it may have moved to some other location on the same or different chromosome.

In the case of the monocots, we also constructed a data set where genes were absent from two of the four species, one from the Panicoideae (sorghum or foxtail millet) and one from the so-called BEP clade (rice or brachypodium). These data sets are smaller than the ones composed of full homology sets that we have been analyzing, containing 2532 monocot sets missing one gene, 833 missing two, 4638 rosid sets and 5510 asterids, but their fractionation patterns are remarkably similar as depicted in Figure 3.4. This analysis offers no support for the idea that the stable and mobile parts of the genome fractionate in different ways, though it only pertains to a restricted portion of the mobile genome.

FIGURE 3.4: Fractionation for homology sets containing no gene for one species (m1) or two species (m2).



3.5.1 Comparison of "stable" genome and the general gene complement

Comparison between our rosid homology sets, reflecting genes syntenically conserved from the original polyploid, and a random set of 9000 genes sampled from the three rosid genomes, unconstrained by homology and syntenic context, shows differences with respect to several GO terms (Figure **??**) Terms such as "developmental process", "reproductive process", "biological regulation", "response to stimulus", "establishment of localization" and "cellular component organization or biogenesis" consist of much higher proportion of our homology sets than of the sampled extant genomes. Most of these, though not "response to stimulus", are consistent with the idea of stable gene complement in these areas of reproduction and development, less focused on the interaction of the cell and its external environment. In contrast, the enrichment of the extant genomes with respect to "externally oriented" membrane terms, extracellular region, immune system, metabolic and catalytic terms reflect high rates of gene family expansion, such as by tandem duplication, and other innovations in these categories. Fractionation of ancient syntenic paralogs would thus not play a large role in these differences.

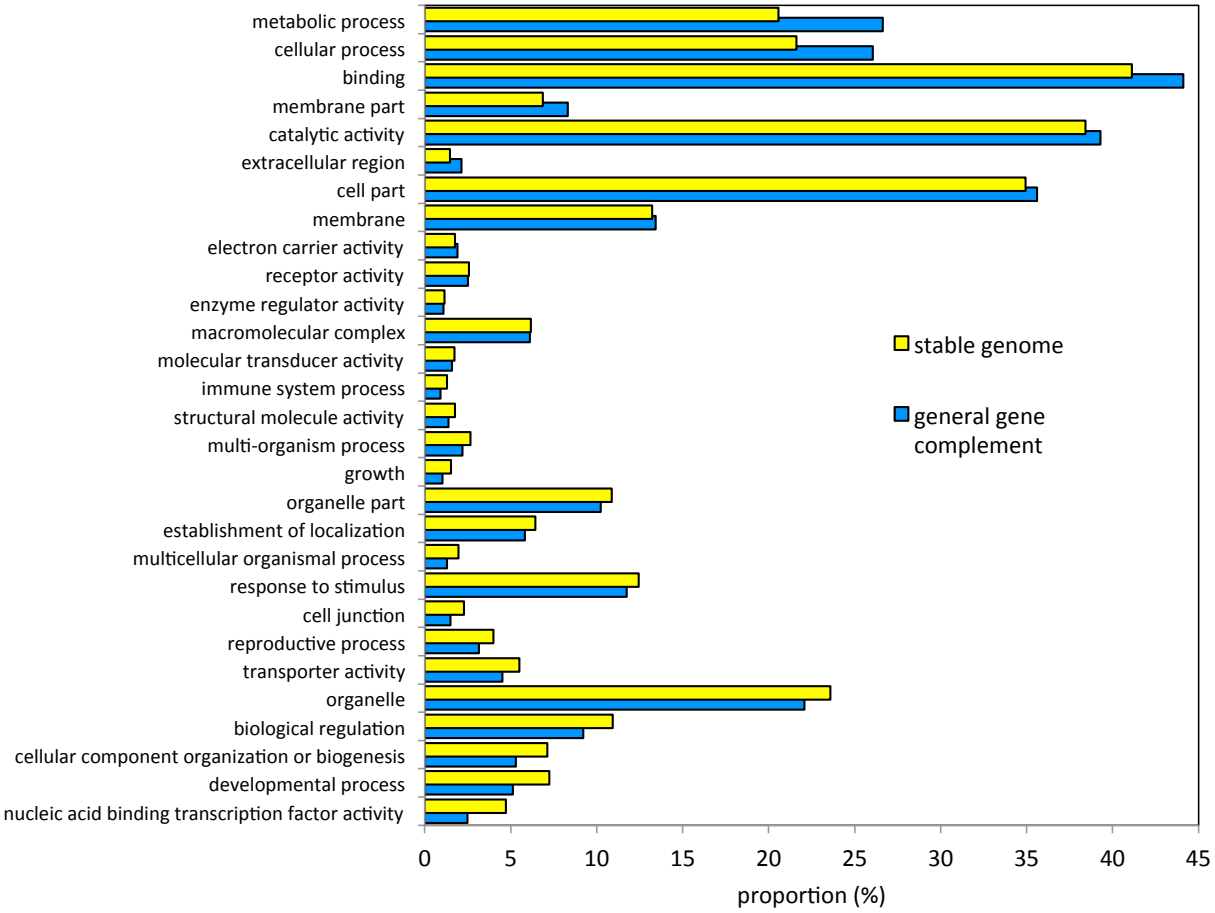


FIGURE 3.5: Comparison of homology sets with sample of three extant genomes, according to hits in functional categories.

3.6 Trends for more specific categories

Finally, we search for tendencies within subcategories of some of the large functional categories. We find that the negative slope within the catalytic activity category reflects in large measure the consistent behaviour, across all three data sets, of isomerase, hydrolase, and some transferase subcategories, while other transferases notably the kinases

and those transferring phosphorus-containing groups, as well as several subcategories of oxidoreductase activity are actually fractionation resistant. The details:

- Oxidoreductase: No overall tendency, but 5 of 88 terms one level lower are fractionation resistant, significantly so in at least one of the data sets:.
  - monooxygenase activity
  - oxidoreductase activity, acting on diphenols and related substances as donors
  - oxidoreductase activity, acting on X-H and Y-H to form an X-Y bond
  - arsenate reductase activity
  - oxidoreductase activity, acting on the aldehyde or oxo group of donors
- Transferase: No overall tendency,
  - (prone) transferase activity, transferring one-carbon groups
  - (prone) methyltransferase activity
  - (resistant) transferase activity, transferring phosphorus-containing groups
  - (resistant) kinase activity
  - (prone) nucleotidyltransferase activity
- Hydrolase: Fractionation prone on the general enzyme class level and 2 of 21 lower terms. “Peptidase activity” is fractionation prone (in conflict with [6]).

  - (prone) hydrolase activity
  - (prone) peptidase activity
  - (prone) hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds

- Isomerase: Fractionation prone on the general enzyme class level and 2 of 14 lower terms.
  - (prone) isomerase activity
  - (prone) cis-trans isomerase activity
  - (prone) intramolecular transferase activity

- Lyase: Has conflicting slopes among the three data sets at the general enzyme class level. No fractionation prone or resistance from 13 lower terms.

- Ligase: No tendency detected

The apparent conflict between the trends for "organelle" and "organelle part" turns out to be due to be the 95% concentration of the latter in chloroplast terms, where only 30% of the former are so annotated.

3.7 Connection between fractionation rates and and paralog retention patterns

In our previous study (Chapter 2) [38] we modeled the duplicate and triplicated data depicted in Table 1 in terms of a single rate of loss of p for the period between hexaploidization and speciation and individual loss rates $q_i, i = 1, \cdots, 6$ for each of the six species starting from an assumed common speciation event. (This assumption is a major biological simplification with, however, no numerical consequences for this particular model.)

TABLE 3.1: Numbers of triples and pairs after fractionation in six rosids.

	frequencies of gene family sizes					
size	peach	cacao	grape	castor bean	strawberry	papaya
2	1484	1111	945	851	606	474
3	256	172	150	119	57	34

Data from [38]

.

In the model, the probability that

1. an original paralogy triplet would survive intact is $\frac{(1-p)^3}{1-p^3}\frac{(1-q_i)^3}{1-q_i^3}$.
2. an original triplet would manifest as a pair is $\frac{(1-p)^3}{1-p^3}\frac{3q_i(1-q_i)^2}{1-q_i^3} + \frac{3p(1-p)^2}{1-p^3}\frac{(1-q_i)^2}{1-q_i^2}$, and
3. an original triplet would be reduced to a single copy is $\frac{(1-p)^3}{1-p^3}\frac{3q_i^2(1-q_i)}{1-q_i^3}+\frac{3p(1-p)^2}{1-p^3}\frac{2q_i(1-q_i)}{1-q_i^2}+\frac{3p^2(1-p)}{1-p^3}$.

Because this model is inspired by concepts of the stable genome, the frequency of single-copy genes, which tend to be part of the mobile genome, are not part of the input data in Table 1, and so the original number of triplicates in the stable moiety, before fractionation, has to be inferred statistically.

It was suggested in Chapter 2 [38], and motivated in part the functional analysis in that research, that a better fit of the model to the data would be obtained by allowing for two or more gene classes with different rates. Thus we have modified the model by dividing the genes into two classes, fractionation-prone and fractionation-resistant in proportions θ and $1-\theta$, with a parameter α linking the rates for the two classes:

$$\text{class } \theta \text{ genes: } p, q_i, i = 1, \cdots, 6$$

$$\text{class } 1-\theta \text{ genes: } \alpha p, \alpha q_i, i = 1, \cdots, 6$$

Based on the data in Table1, with an initial number of pre-triplication genes of 7500, the maximum likelihood fit produced two populations with 75 % and 25 % of the genes, respectively, and with a relative rate parameter $\alpha = 0.60$. For larger initial numbers of genes, only p varied somewhat to compensate, while the estimates of the q_i, θ and α were stable. From these estimates, we could predict the number of genes in each class using the same x-axis as Figures 1 and 3, based on p and q_i for the peach, cacao and grape genomes, and plot the relative proportions of the two gene populations in each class, in Figure 3.6.

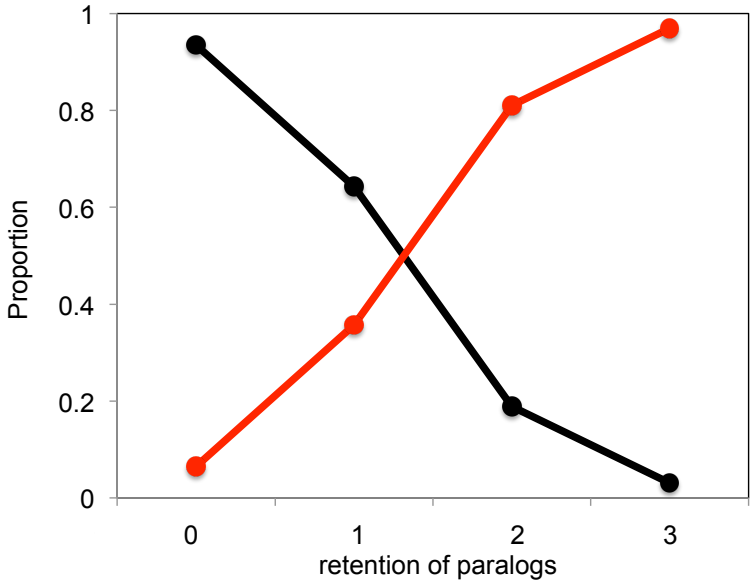


FIGURE 3.6: Predicted paralog retention patterns in rosids for model with two gene classes. The black line represents the fractionation-prone gene population and the red line represents the fractionation-resistant population.

We refrained from trying to model the annotation process, so the results in Figure 3.6 are more contrastive of the two population of genes than the real data. Nevertheless, it is encouraging that the empirical data in Table 1 contain the signal of two classes of genes with different fractionation rates, reminiscent of the patterns in Figures 1 and 3.

3.8 Conclusion

Building on our previous results on the functional determinants of fractionation of rosid genomes, we showed that the independent fractionation patterns of rosids, asterids and

grasses are highly parallel. The same functional categories of genes are preferentially fractionation prone or resistant in all three lineages.

That the asterids demonstrate the same patterns as the rosids and monocots confirm that there is a general trend for paralogs to be fractionation prone or resistant, as influenced by their functional categories, despite the additional WGD events in the asterid lineages. This may be attenuated by changing fractionation patterns for some gene copies from one WGD to the next, as reported recently [76].

Additionally, we tested whether these patterns of fractionation were consistent between the stable and mobile portions of the genome. Even though the additional cohort of genes, part of the more "mobile" genome, is less numerous and thus less conducive to statistical significance of the results, we find that it supports most of the tendencies found for the functionally "stabilized" part of the genome. Our analyses support that there is no fundamental difference between the strictly stable and partly mobile parts of the genome. Rather the same functional categories influence the fractionation fate of genes. Nevertheless, we also demonstrate that the homology sets we study are enriched for some categories and depleted for others, in comparison with random samples of genes from the entire extant genomes. This likely has more to do with lineage-specific expansion of gene families, rather than fractionation dynamics.

Perhaps the most important improvement to be envisaged for our method would be to ensure that the homology sets contain only paralogs created by the initial WGD

event. Tandem duplicates or other duplicates produced more recently than the WGD may constitute a major proportion of all the duplicates in a genome. Although SynMap normally excludes tandem duplicates, some of them undoubtedly remain and may be reintroduced when we combine pairwise genome results to form homology sets. Indeed, for future work, it would be worthwhile to contrast the paralog loss behaviour from WGD fractionation with that from other sources of duplicates. This approach was pioneered in [77] for the *Arabidopsis* genome, with results very similar to our approach, suggesting our comparative approach might bolster these findings.

Our use of general GO terms could conceivably be improved by using a more focused gene ontology database such as Plant Slim developed by the Arabidopsis Information Resource. Indeed, preliminary tests show that our results on high-level terms could be sharpened using this resource, but unfortunately there is relatively little annotation at present using lower level terms, so this avenue is limited for the moment.

It may advance the understanding of functional associations of fractionation to compare the correlations among functional categories enriched for fractionation-prone genes in contrast to the correlations among these categories for fractionation-resistant genes. It is not a methodological flaw that many categories are correlated among themselves — indeed it opens up the opportunity to compare changes in these correlations for genes with differing fractionation scores.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EC and DS planned the research for this article. EC carried out most of the research and writing. EC, DS, EL, VAA, LC-P and HT participated in the interpretation of the results. CZ did the preparation of the genomic data and helped in the computer analysis. AB did the probability modeling and calculations. CBAN assisted in the data analysis and preparation of the manuscript. All authors participated in writing. All authors read and approved the manuscript.

Chapter 4

Gene expression and fractionation resistance

Eric C.H. Chen, David Sankoff

Published: 2014 - BMC Genomics 15 (Suppl 6), S19

4.1 Abstract

Background: Previous work on whole genome doubling in plants established the importance of gene functional category in provoking or suppressing duplicate gene loss, or fractionation. Other studies, particularly in *Paramecium* have correlated levels of gene expression with vulnerability or resistance to duplicate loss.

Results: Here we analyze the simultaneous effect of function category and expression in two plant data sets, rosids and asterids.

Conclusion: We demonstrate they have independent effects, though expression does not play the dominant role it does in *Paramecium*.

4.2 Background

Whole genome doubling (WGD) is a special case of gene duplication in that everything in the genome, including the genes, regulatory elements, and repetitive regions, is doubled or tripled. This process is more common in plant lineages than in other evolutionary domains [78, 79] and is an important source of gene innovations, contributing to diverse morphological and functional complexities in modern plants [67, 80]. The duplicated genes are very vulnerable to loss after the WGD event via excision of chromosomal segments or pseudogenization. These losses are collectively referred as fractionation. Various models have been proposed to explain this process, such as the Gene Dosage Hypothesis [3, 81] and the Gene Balance Hypothesis [7]. These models try to explain the difference in duplicate gene retention pattern based on the traditional models of gene fate: neofunctionalization, subfunctionalization, and pseudogenization, and on the observations on duplicate gene retentions from WGD.

We have shown in Chapter 2 and 3 several groups of plants - rosids, asterids, and monocots - that the functional category of a gene is a major determinant of fractionation resistance, with metabolic genes being fractionation prone, and "response to stimulus" being fractionation resistant [38, 82].

A recurrent theme in works relating to fractionation is the effect of gene expression. In a comprehensive study of fractionation in *Paramecium*, Gout *et al.* [27] identify a clear relationship between high WGD duplicate gene retention rates and high expression level. They also find that within each major gene functional class, higher expression correlates with higher duplicate retention rates; even if the expression levels of each major functional classes differ from each other. They conclude that expression level is the best discriminator for explaining variable resistance to fractionation.

The Gout *et al.* paper [27] is the primary inspiration for this study, where we explored the relationship between of functional class and expression in fractionation resistance in plants. Because we have previously shown that functional class can itself influence the fractionation resistance of the duplicates [38, 82] we wish to consolidate these two kinds of findings into one unified framework.

4.3 Results

4.3.1 Methods

We analyze the genomes of peach [29], grape [12], and cocoa [43], constituting the rosids data set. These selected species have not undergone WGD events since their triplicated last common ancestor at the base of core eudicots, which makes them invaluable for studying long term effects of fractionation. To study the effect of fractionation from comparatively more recent WGD events and the effect of fractionation from multiple WGD events, we survey the genomes of tomato [30], *Utricularia* [14], and *Mimulus* [31], making up our asterids data set. The asterids diverged from the rosids a few million years after the triplication event in the core eudicots ancestor some 120 million years ago [71] and each of the selected species of the asterids has since undergone more recent WGD events [14] (Figure 4.1). We will show that even if individual species of each data sets have evolved and fractionated independent of each other, the overall trend of fractionation remains highly parallel.

Due to the still scarce availability of high quality expression data, we make use of RNA-seq data from grape [41] to represent expression in the rosids and RNA-seq data from tomato [30] to represent expression in the asterids. Tomato gene expression values are about three to four times higher than grape values because different technology platforms and depth of sequencing. This prevents the meaningful comparisons of absolute

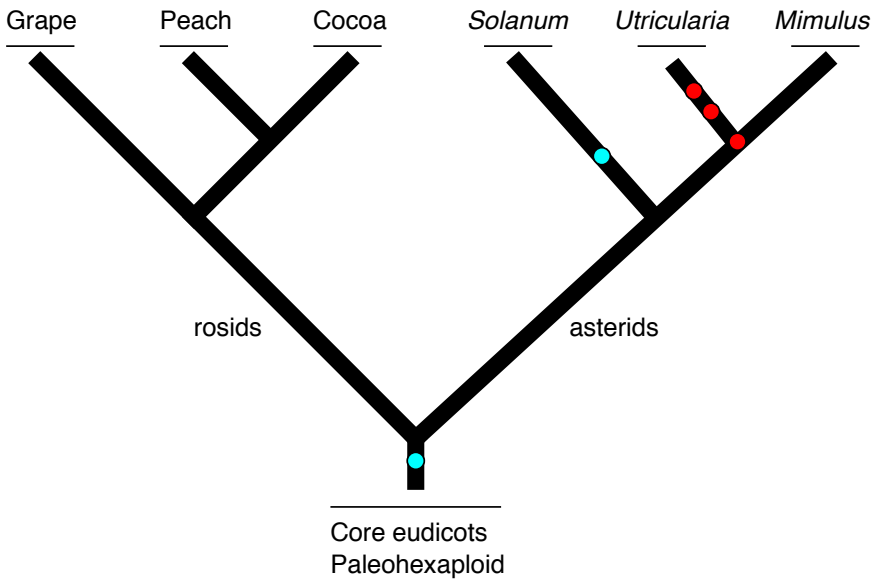


FIGURE 4.1: Phylogeny of the rosids and the asterids data sets. Species in the rosids data set did not have additional WGD events since their paleohexaploid ancestor. Species in the asterids data set did have additional WGD events (red circles) with one whole genome triplication event (blue circle). Branch length are not to scale.

gene expression values between grape and tomato. Normalized comparisons, however, are valid.

We are interested in comparing functional categories, expression levels and fractionation among thousands of genes but the inclusion of a few extremely highly expressed genes could swamp some of these comparisons. Thus we filter out the genes in the top 1% of expression levels. The filter is most pertinent to the more specific GO categories such as individual enzyme classes where the number of genes in each class may be small. Filtering is not necessary for the top level categories since they contain thousands of genes, but for consistency we keep the filter for all the expression analysis.

To take into account varying plant tissues having different expression profiles as well as plant responses to different environments and stimuli, we use the highest reported expression value for any given gene rather than the median or mean. The rationale for this is that the RNA-seq data we are using distinguishes different expression level in different tissues as opposed to responses to a particular stimuli. Many genes are only expressed in specific tissues, so the maximum expression level of a gene is a better indication of its importance in the organism.

4.3.2 Data

In this comparative study we first categorize genes into distinct homology sets, where each set represents the ancestral gene of the rosids or the asterids ancestor just prior to the WGD event(s) [37, 69, 82]. Categorization is based on both gene sequence similarities and positional conservation and is done using SynMap [83] with genomic resources obtained from CoGe [54]. The homology sets are refined using algorithms by Zheng *et al.* [37, 69]. These homology sets then allow for the calculation of levels of fractionation resistance (F). The fractionation resistance of a homology set is determined by the number of species (N) that still have the genes of the set in duplicates in the form: $F = N + 1$. The higher the number of species still retaining WGD in duplicates, the higher the fractionation resistance. A gene that has been returned to singleton in all species has $F = 1$. Each homology set is then annotated with Gene Ontology [39] terms using Blast2GO [40]

to classify their functional class. All the terms associated with any of the genes in a homology set are retained. In addition, by design of the Gene Ontology [39] if a gene is annotated with a particular GO term, it automatically inherits all the parent GO terms which are more general (Figure 4.2). Some of our analysis include individual gene rather than whole homology set; for these we propagate GO annotation by having all genes of a homolog set inheriting particular GO terms if one of the genes is annotated with them.

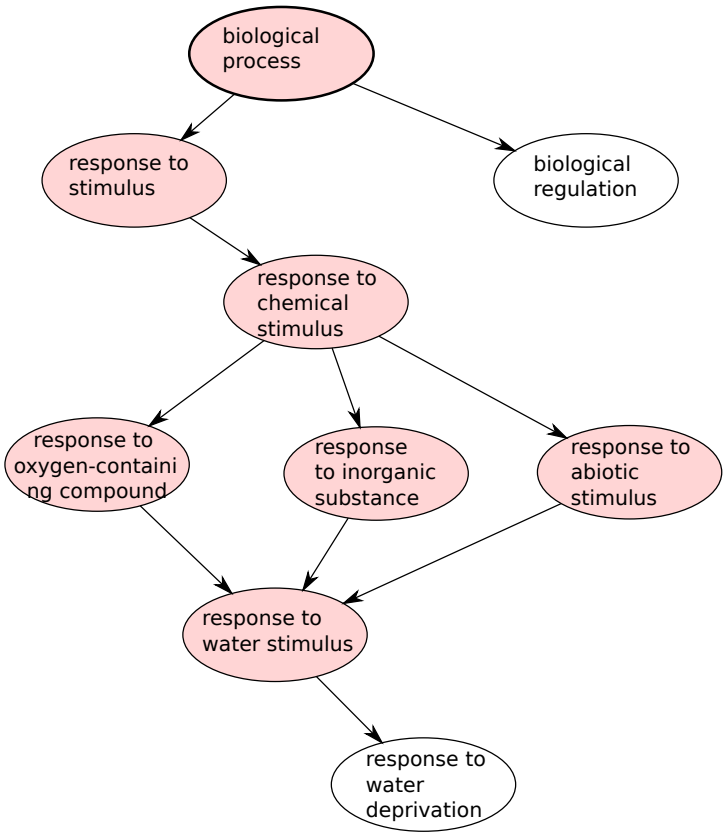


FIGURE 4.2: Gene Ontology. In this example, if "response to water stimulus" is hit by a gene in a homology set then all the terms above it in the GO hierarchy are also hit (coloured term).

GO terms are divided into three domains "biological process", "cellular component", and

"molecular function". A homolog set may not always be annotated in all three domains. Most of our analyses will focus on the mutually exclusive top level categories with each domain. Since most of the homolog sets have fractionated to be single copies in surveyed species, we use normalized proportions (P) to fairly compare enrichment of functional classes. We set $Hit(F, C)$ = the number of "hits" in category C by homology classes with fractionation level F.

$$P(F, C) = Hit(F, C) / \sum_{C'} Hit(F, C')$$

Where the sum is taken over all category C' in the top level domain including C. As we plot $P(F, C)$ against F, we will deem C to be fractionation resistant if $P(F, C)$ increases with increasing F. We deem C to be fractionation prone if the reverse case is observed, where $P(F, C)$ deceases with increase in F.

A similar formula is used to normalize the distribution of gene expression as well. Log transformed expression level of genes is collected into bins of similar expression levels. Functional expression E is a simple average expression of all genes of homology sets of F and C.

$$P_B(F) = Number\ of\ genes\ in\ bin\ B\ from\ F / \sum_{B'} number\ of\ genes\ in\ F$$

$$E(F, C) = AverageExpression[Genes\ of\ Homology\ Set\ in\ (F, C)]$$

4.3.3 Functional Analysis

Figure 4.3 displays the effect of functional class on increase in fractionation resistance. The conclusion is that some top level functional classes are more fractionation resistant (such as "biological regulation", "response to stimulus", "membrane", "developmental process", "establishment of localization", and *etc.*) or more fractionation prone (such as "metabolic process", "catalytic activity", and "cellular process") than others and that the finding is highly parallel across these two different lineages. Despite the shallow slopes of the curves on the 0 - 100% scale of normalized proportion, both the rosids and the asterids data sets contain thousands of homology sets and are statistically significant. All functional classes shown have p value to be orders of magnitude less than 0.05 in linear regressions. The statistical test is a linear regression test based on all the homology sets, not just the few summary points in Figure 4.3. This top level result is in agreement with results reported by Gout *et al.* [27] in the organism *Paramecium*.

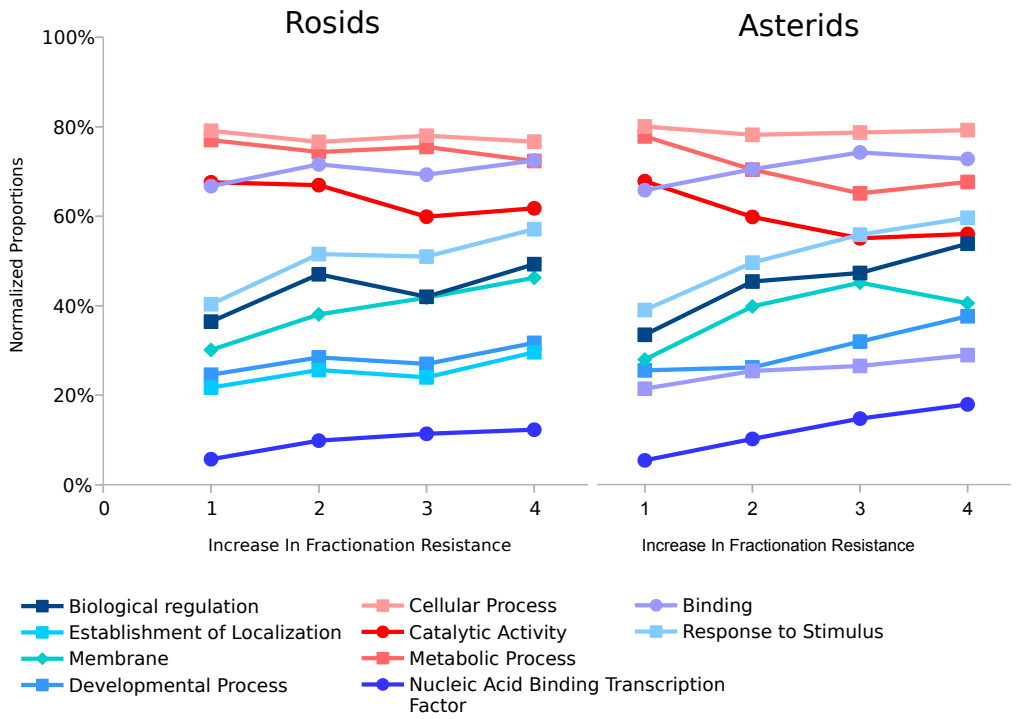


FIGURE 4.3: Biased gene fractionation resistance in different functional class. Updated result from [82]. Only three functional classes in red are fractionation prone.

4.3.4 Expression Analysis

Using means and medians to summarize the expression levels of different fractionation resistant homology sets, which may be represented by one or more genes in grape (or tomato, as the case may be) we also find a trend where more highly expressed sets tend to be more resistant to fractionation. This is true regardless if we plot all genes or plot genes of specific top-level functional class (Figure 4.4). Investigating more specific functional categories such as enzyme classes (Figure 4.5) reveal similar findings where higher expression within each class contribute clearly to increase in fractionation resistance.

This correlation of higher gene expression to fractionation resistance confirms results reported by Gout *et al.* [27], as reproduced in Figure 4.6. The startling parallelism between the rosids and the asterids is also present in Figure 4.4 and 4.5. The trend of increase in expression level with increase in F is the same in both rosids and the asterids. The ranking of functional class by expression to F is also the same in both data sets. The effect of expression on F appears to be universal.

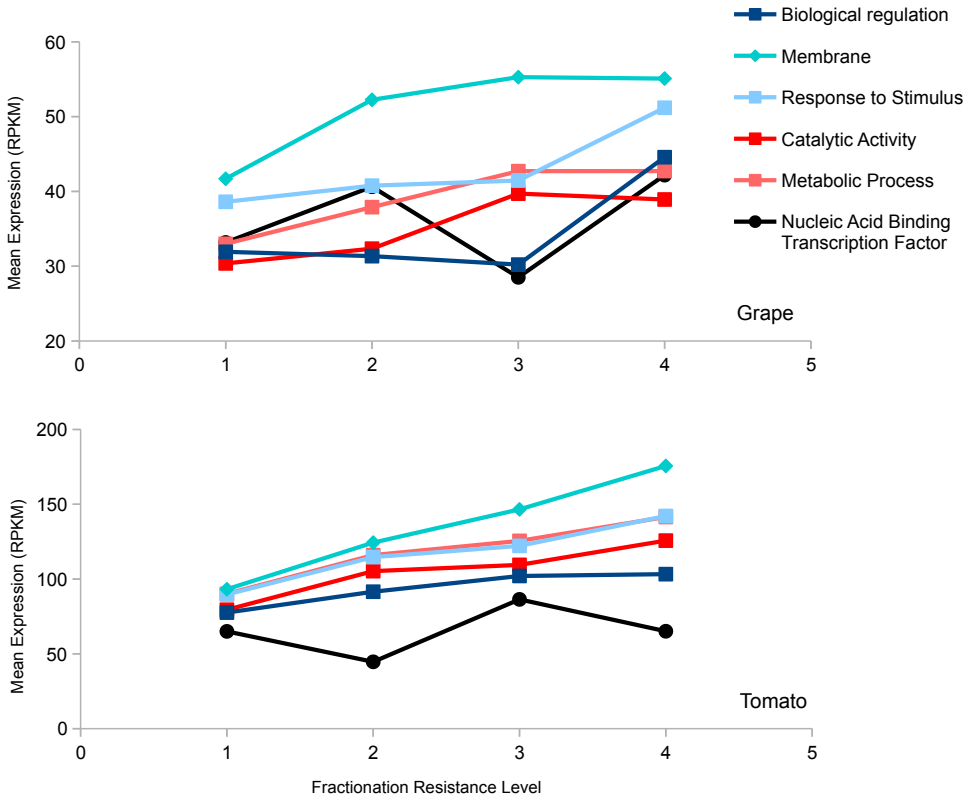


FIGURE 4.4: Gene expression level of different functional class across different fractionation resistance levels. The units in RNA-seq is expressed as number of reads per kilobase per millions reads (RPKM). "Nucleic acid transcription factor binding activity" (in black) is not significant. The statistical test is based on all genes in a category, not just the few summary points in the graph.

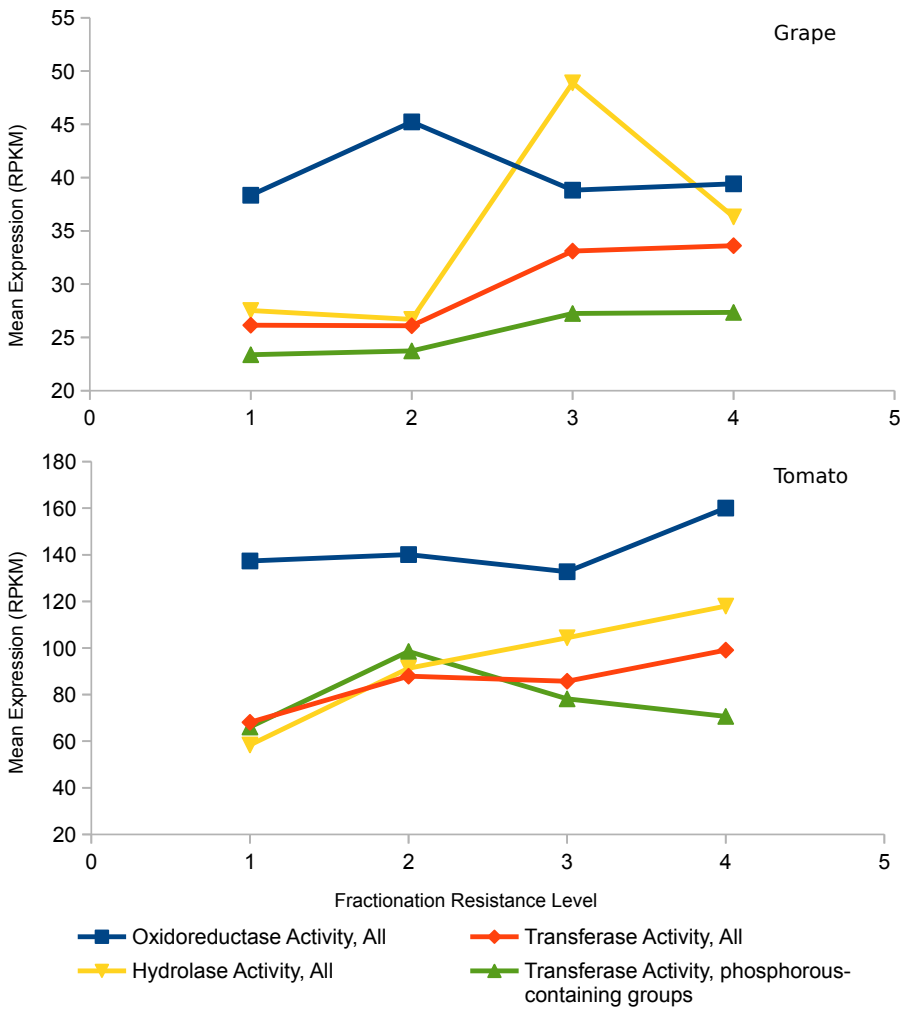


FIGURE 4.5: Expression analysis of major enzyme classes in grape and tomato. Only showing select enzyme classes that with large member sizes (more than 500 members) that are statistically significant. In enzyme classes with small member sizes, the presence of a few highly expressed genes easily distort the averages, even with the 1% filter in place. The statistical test is based on all genes in a category, not just the few summary points in the graph.

However, the change in expression levels of sets from the least fractionation resistant to the most fractionation resistant is smaller in our rosids and asterids data sets than in [27]. Functional classes with lower number of genes ("nucleic acid binding transcription factor activity" has 411 genes in grape and 552 genes in tomato) are still suggestive of the trend

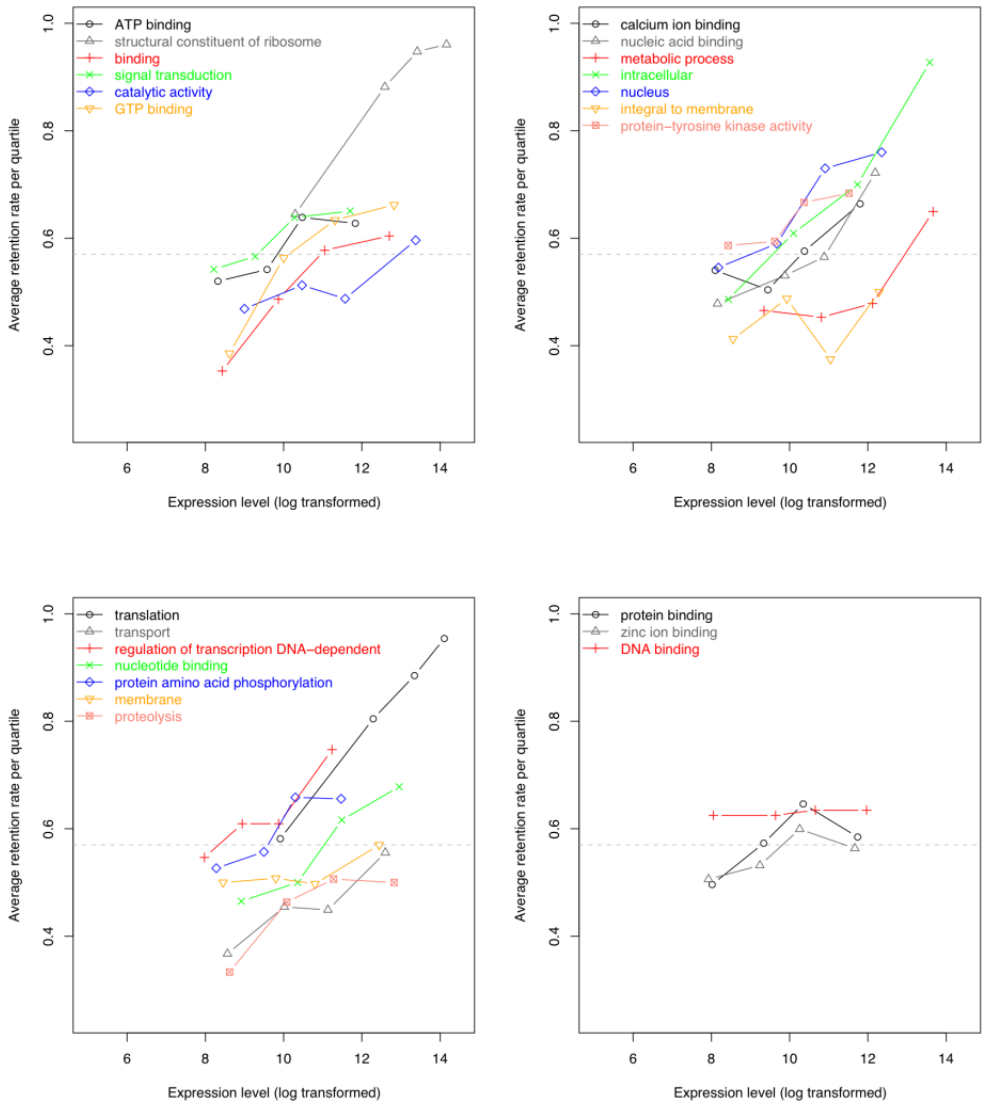


FIGURE 4.6: Taken from Gout *et al.* [27]Figure S3

but are no longer statistically significant in both data sets ($p > 0.05$). These differences may be due to sample size rather than the differences between the protist *Paramecium* and plants.

An important difference between our results and those on *Paramecium* [27] is that a significant number of genes with low expression level are highly retained in the plants whereas in the *Paramecium*, genes with high retention rates are very unlikely to have low expression level. There also exist functional classes that have widely varying expression levels across fractionation resistant class, such as genes from GO term "nucleic acid binding transcription factor activity" (Figure 4.4). However, using the distribution of gene expression to compare the different fractionation categories instead of using means of categories (Figure 4.7) reveals that the highest and the lowest expression bin distorts the general expression pattern of the rest of genes. Therefore, the data still suggests that higher expression levels still results in higher likelihood of being retained in duplicate form.

On the other hand, functional class appears to have greater influence than expression on determining if a homology set is fractionation resistant or not. Both GO terms "metabolic process" and "catalytic activity", which are reported to be very fractionation prone, have similar expression level to GO term "response to stimulus", a very fractionation resistant GO term. The GO term "biological regulation", one of the most fractionation resistant terms, have on average lower expression levels in both of the rosids and the asterids datasets.

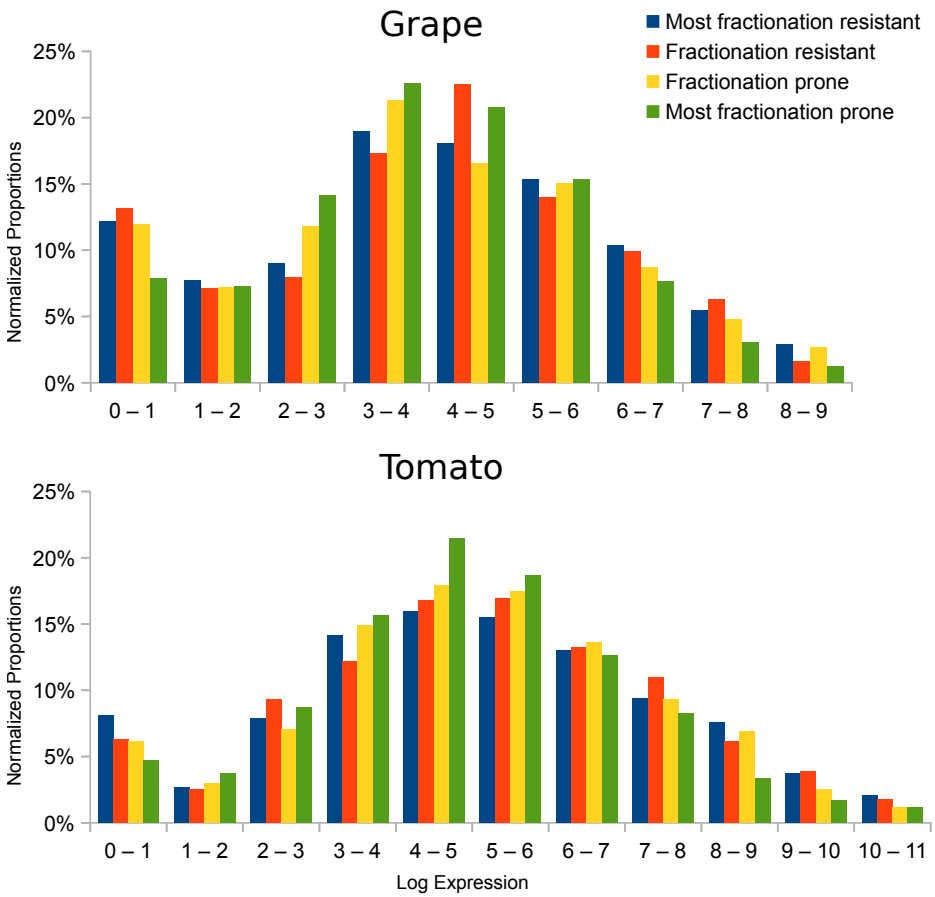


FIGURE 4.7: Gene expression in different fractionation resistance levels. The nature of a RNA-seq experiment means not all possible states of the organism is recorded, therefore the high proportion in the lowest expression block can be explained by the gene not being activated in the condition when the RNA-seq data was collected.

4.4 Discussion and Conclusion

How can we reconcile the relatively small effect of gene expression on fractionation resistance with the claim that gene expression levels are fundamental to copy number variations and fractionation resistance [3, 27, 81, 84–86].

One of the more plausible explanation is rather than just the fitness cost of gene expression controlling fractionation resistance, the fitness cost of disruption of the intended function of the gene or the gene network is a greater contributing factor. Highly connected genes have been reported to be preferentially retained [26, 87] and are predicted to be more retained by both the Gene Balance Hypothesis [7] and the Gene Dosage Hypothesis [3, 81]. As such, gene of a functional class that generally have low expression levels may still have high fractionation resistance level due to the importance of the function or the functional network.

It should be noted that many other factors have been proposed to explain variable fractionation rates. Moghe *et al.* [88] showed that gene sequence features such as longer amino acid length and higher GC3 level, the wobble position in protein translation, contribute to fractionation resistance in *Raphanus raphanistrum*, *Arabidopsis thaliana*, *Arabidopsis lyrata*, and *Brassica rapa* in addition to functional class. They also report that in different WGD events the degree of enrichment from gene sequence features and functional classes may vary, though the directionality of enrichments (be they contributing to fractionation resistance or fractionation proneness) remains mostly the same [88]. We were unable to replicate these results on our data using multiple regression; only expression level was consistently predictive of fractionation.

Of interest, a recent study Makino *et al.* [85] reports the effect of fractionation from ancient vertebrate WGD on the biased distribution of genes with copy-number variations

in humans. This paper claims that retained duplicates suppress changes in copy number in their vicinity.

In conclusion our result agrees with current models that expression does play a role in fractionation resistance although by itself it can not explain the enrichments of functional classes. It is likely that systemic analysis on more genomes will be needed to clarify the role of expression and other sequence feature in explaining fractionation. At the present time a good predictor of fractionation resistance should still contain both expression and functional class and may even include how connected a gene is in the genome.

Author's contributions

Both authors participated in the research, wrote the paper, read and approved the manuscript.

Chapter 5

Statistical analysis of fractionation resistance by functional category and expression

Eric C.H. Chen, and David Sankoff

Manuscript In Prep: 2015

5.1 Abstract

Background: The current literature establishes the importance of gene functional category and expression in promoting or suppressing duplicate gene loss after whole genome

doubling in plants, a process known as fractionation. Inspired by studies that have reported gene expression to be the dominating factor in preventing duplicate gene loss, we analyzed the relative effect of functional category and expression.

Conclusion: Our results suggest that the effect on duplicate gene retention fractionation by functional category and expression are independent and have no interaction. Also, in plants, functional category is the more dominant factor.

5.2 Introduction

Virtually all land plants have a polyploid ancestor [8, 9, 71] with many lineages having additional rounds of whole genome duplication events [30, 71]. These special events in evolutionary history have been linked to increased morphological and genetic diversity [66, 67].

Whole genome duplication, by definition, doubles the genomic content of the species. The doubling of both genes and associated regulatory elements distinguishes whole genome duplication from the more common form of duplication, tandem duplication. Tandem duplication, in contrast, only duplicates one, or perhaps a few, genes at a time, occurs more regularly through evolutionary history, and may not include associated regulatory elements. This distinction results in differential effects of functional class [3, 22, 24] in

fractionation, the significant and rapid gene loss of excess genes after whole genome duplication in plants [8].

The established subfunctionalization model [1, 2] for explaining duplicate gene retention did not attempt to explain this differential effect of functional class between duplicate genes of whole genome duplication and duplicate gene of tandem duplication [3, 7]. Subfunctionalization argues that duplicate genes are maintained by loss of functions in each duplicate such that both duplicates are now required for proper function. A prime example of subfunctionalization is the evolutionary history of hemoglobin α and hemoglobin β [89]. In the subfunctionalization model, no distinction are made between duplicate genes of whole genome duplication and duplicate genes of tandem duplication. But it has been shown that metabolic genes tend to be retained after tandem duplication while regulatory genes are preferentially retained after whole genome duplication. The subfunctionalization model is not adequate to explain the differential effects of functional class a fractionation and on gene loss from tandem duplication.

Two proposed models, "Gene Balance Hypothesis" and "Gene Dosage", offer alternatives to explain this discrepancy [3, 7, 81]. In both models gene expression plays a role in determining duplication gene retention. Both models assume all gene expressions go on to make the corresponding protein products. So the quantity of the protein products can be proxied by gene expression levels. The main difference between the two proposed models is regarding the role of gene expression.

The Gene Balance Hypothesis argues that the penalty of disrupting stoichiometry ratios of gene products is the driving force behind preferential duplicate gene retentions [7]. An example is the circadian clock genes in *Brassica rapa*. The duplicates from whole genome duplication are retained more than expected to maintain the highly delicate dosage balance between the circadian clock genes [26].

Gene Dosage, in contrast, argues that high gene expression itself is driving the retention. The rationale is that high gene expression is a costly endeavour to the species, thus indicating the high importance of these genes to the species. It follows that disruptions to these highly important genes is selected against. The extensive work done on *Paramecium* by Gout *et al.* [27] is an example where it is reported that highly expressed genes tend to be retained in duplicates after whole genome duplication, regardless of functional class.

These two models are not mutually exclusive [20, 90]. By controlling for functional class before examining gene expression and duplicate gene retention, correlation between high expression and high duplicate gene retention is apparent [27, 90].

Both the Gene Balance Hypothesis and the Gene Dosage are needed because each model explains observations that the other model can not fully explain. What remains unclear is which model impacts duplicate gene retention more. In this paper we will show that Gene Balance Hypothesis is the dominant model.

5.3 Methods and Materials

5.3.1 Data

We construct gene families based on the sequence similarity and the conserved gene order between extant species using CoGe [36, 38, 54]. These gene families are pruned into smaller units that are linked by the whole genome duplication in the ancestor using the "Orthologs for Multiple Genomes" program [37].

The species grape [12], peach [29] and cacao [43] form the rosid data set. These species can trace their last common ancestor to the period after the divergence of the asterids, following the core eudicot hexaploid about 120 million years ago [71]. There are no additional rounds of whole genome duplication in the evolutionary paths leading to the these present-day species [12, 29, 43]. Therefore, whole genome comparative analysis of the rosid data set offers insights on the effects of fractionation over long period of time.

The asterid data set provide a different viewpoint of the fractionation process compared to the rosid data set. The asterid last common ancestor diverged five to ten million years after the hexaploid core eudicot ancestor. This early divergence means the fractionation process after the hexaploid ancestor of the asterid data set is mostly independent from the fractionation process in the species of the rosid data set. Furthermore, the species

of the asterid data set, which consists of extant species tomato [30], *Mimulus* [31], and *Utricularia* [14], have additional rounds of whole genome duplication [71].

The asterid data set addresses two potential concerns. The first concern is whether the results of the rosid data set represent a general effect or a clade-specific trend. The second concern is whether the additional rounds of whole genome duplication introduce a different pattern compared to single ancient whole genome duplication event.

We use grape [41] of the rosid data set and tomato [30] of the asterid data set for the expression analysis due to the availability of the high quality RNA-seq expression data. Both grape and tomato's expression data include different organs for a more comprehensive evaluation of gene expression levels. For each gene, we use the highest expression level the gene display across all organ to represent the gene's expression level.

5.3.2 Retention indices

We use retention indices to measure how fractionation resistant or prone gene families are. The retention index of each gene family is calculated by counting in how many species the genes is still maintained in duplicate. For example, if a gene family of the rosid data set is maintained as duplicates only in grapes, then the retention index of that gene family is one. Since there are three species in both the rosid data set and the

asterid data set, retention indices range between zero (gene set reduced to singletons in all species) and three (gene maintained as duplicates in all species).

5.3.3 Expression

In the expression analysis where only one species is present in each data set, we apply additional normalization to the data. For this part of the analysis, we use individual genes instead of gene families for two reasons. The first reason being genes in duplicate families have varying gene expressions that may differ by orders of magnitude. The skewness of the data prevent us from using averages. We also can not just take the highest expressing gene in the gene family similar to how we choose the organ with the highest expression to represent the gene's expression. This is due to reason number two, which is the trend that the more genes a gene family has, the higher the expression of the gene family will be by virtue of having more chance to include a high expressing gene.

5.3.4 Annotations

We use Gene Ontology (GO) [39] terms to classify gene families into functional categories via Blast2GO [40]. GO terms are nested within each other to provide different resolution of annotation (Figure 5.1). We call GO terms that are close to the one of the three "root terms" "high level terms". These high level terms describe general functional categories.

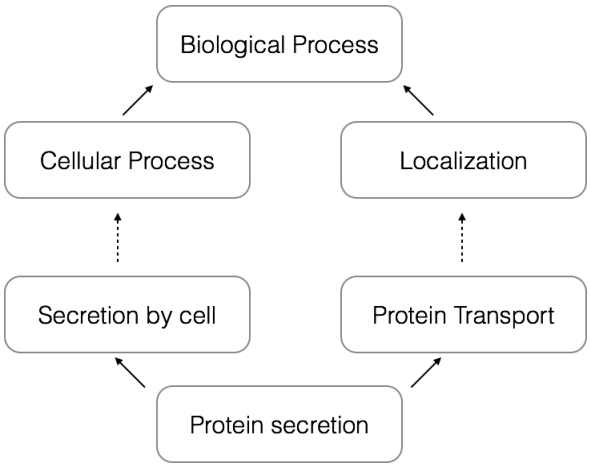


FIGURE 5.1: Example on nested structure of GO terms Starting at a low-level GO term "protein secretion", it is inherited by two slightly higher GO terms "secretion by cell" and "protein transport". After a few more levels of GO terms (represented by dashed lines), the starting GO term is now inheriting two high level terms "cellular process" and "localization". These high level terms are then linked to the root term, "biological process". There are three root terms in gene ontology, they are "biological process", "cellular component", and "molecular function" [39].

As a result, a particular gene may be annotated with multiple high level terms as shown in Figure 5.1.

We designate three high levels GO functional categories (Figure 5.1) that we previously, in Chapter 3 and 4, found to have the highest effect on fractionation [82, 90]. The first category is the "Metabolic process (Z1)", one of the most fractionation-prone. The second category is the "Enzyme class (Z2)". It is also highly fractionation-prone but it includes enzymes involved in signalling pathways so the category as a whole may show increased retention compared to Z1. The third category is the "Regulation and Response" (Z3). This is composed of two most fractionation-resistant GO categories.

Each high level functional categories is further divided into six low-level GO categories to represent more specific and biologically distinct functions. GO terms "secondary metabolic process", "lipid biosynthetic process", "steroid metabolic process", "nucleobase-compound containing metabolic process", "carbohydrate metabolic process", and "protein metabolic process" represent Z1. These six metabolic GO terms are representative of diverse metabolic processes. GO terms "transferase activity", "oxidoreductase activity", "hydrolase activity", "ligase activity", "lyase activity", and "isomerase activity", the six major enzyme classes, represent Z2.

GO terms "regulation of metabolic process", "nucleic acid transcription factor activity", "signal transduction", "response to hormone", "response to temperature", and "response to stress" represent Z3. This is a combination of two highly fractionation-resistant functional categories in "biological regulation" and "response to stimulus" [82] so that there are six low level and biologically distinct GO terms in each high level functional categories (Table 5.1).

5.4 Results

From our previous results in Chapter 3 and 4 [82, 90], we predict Z1 to be the most fractionation-prone, closely followed by Z2, and then Z3. The latter should be significantly different from both Z1 and Z2 by being fractionation-resistant.

Table 5.1: GO terms and number of genes

				Tomato	Grape
Metabolic process (Z1)	Z11	GO.0008610	lipid biosynthetic process	286	397
	Z12	GO.0008202	steroid metabolic process	54	75
	Z13	GO.0006139	nucleobase containing compound metabolic process	655	1055
	Z14	GO.0005975	carbohydrate metabolic process	575	810
	Z15	GO.0019538	protein metabolic process	1109	1389
	Z16	GO.0019748	secondary metabolic process	131	214
Enzyme Class (Z2)	Z21	GO.0016740	transferase activity	962	1227
	Z22	GO.0016491	oxidoreductase activity	529	693
	Z23	GO.0016787	hydrolase activity	878	1254
	Z24	GO.0016874	ligase activity	177	246
	Z25	GO.0016829	lyase activity	119	152
	Z26	GO.0016853	isomerase activity	67	131
Regulation and Response (Z3)	Z31	GO.0019222	regulation of metabolic process	965	1043
	Z32	GO.0001071	nucleic acid binding transcription factor activity	403	324
	Z33	GO.0007165	signal transduction	550	573
	Z34	GO.0009725	response to hormone	492	464
	Z35	GO.0009266	response to temperature stimulus	291	284
	Z36	GO.0006950	response to stress	1032	1301

The inherently different gene count for different functions (Table 5.1) means the categories are not balanced for ANOVA. We sidestep the issue by using the average retention index of each functional category instead of the raw count. This strategy comes at the expense of statistical power since we are now left with just two data points for each low-level functional categories. Still, Figure 5.2 shows the expected result of high expression correlating with high fractionation-resistance.

Figure 5.2 is a visual representation of what the average retention indices are for each

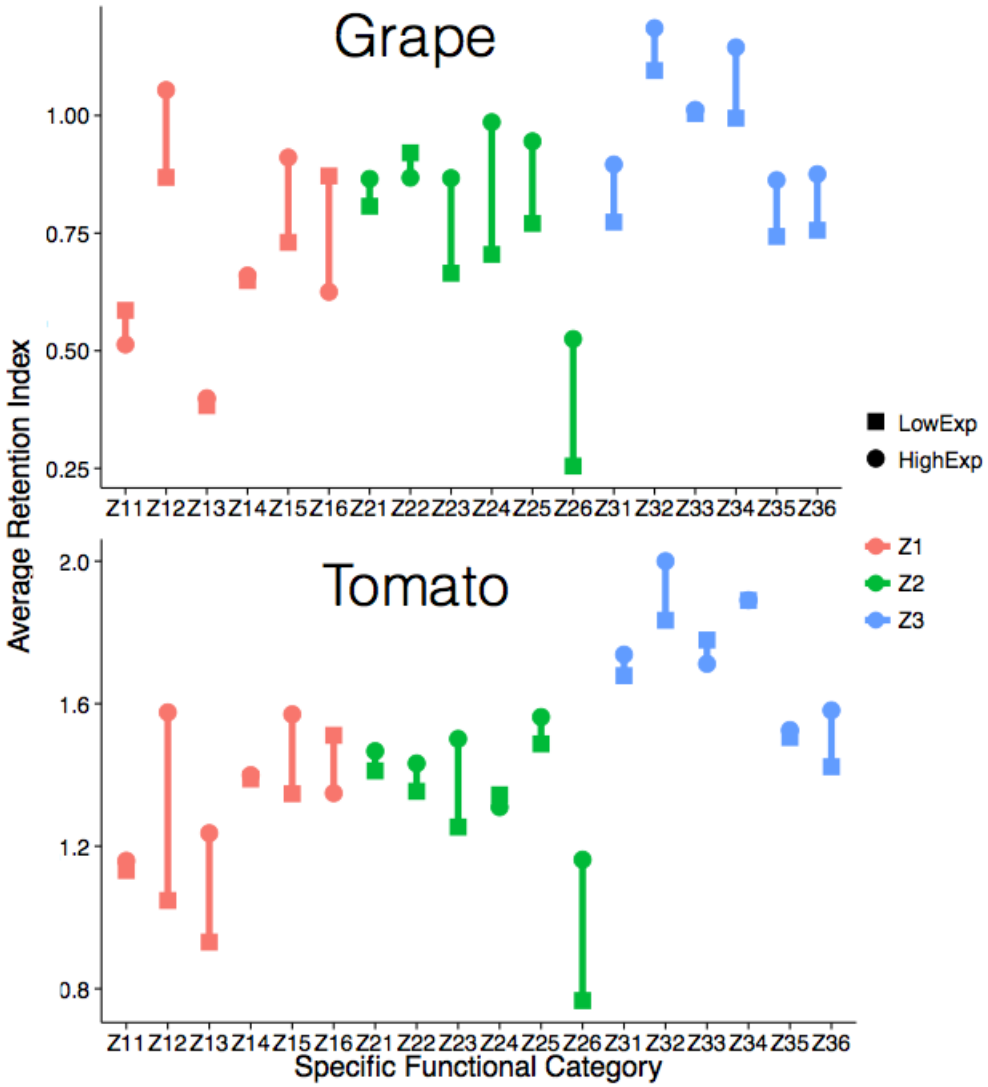


FIGURE 5.2: Summary of average retentions indices in grape and tomato. Each functional category has two data points: average retention index under low expression (LowExp) and average retention index under high expression index (HighExp).

functional category. This result is consistent with our prediction that genes of Z3 are more fractionation-resistant than gene of Z2 and Z1.

This is further reinforced in Figure 5.3. This supports our prediction that genes of Z3 are more fractionation-resistant than Z1 and Z2. In grape, the adjusted p-value for the statistical test of that difference between Z3 and Z2 is above the usual 0.05 limit but this is likely due to the trade-off for having balanced data.

Figure 5.3 also shows that in grape, the difference between fractionation-resistant Z3 and fractionation prone Z1 and Z2 are smaller than the difference in tomato. A reason for this observation is that that gene families that are singletons in all three species of the rosid data set constitute a far higher proportion than in the asterid data set, so even the fractionation-resistant functional category contain many singleton gene families.

The ANOVA table (Table 5.2) answers the main objective of the paper: which of Gene Balance Hypothesis and Gene Dosage impact duplicate gene retention more? We answer this by calculating whether functional categories or expression levels have the bigger effect size in the two-way ANOVA. In the table, the effect size, measured in partial eta squared, supports the suspicion in the Chapter 4 [90] that functional categories carry more weight in determining retention indices more than expression levels. The table also shows that while functional categories strongly affect average retention indices, the effect of expression levels have on average retention indices are no longer significant. This non-significance may be a consequence of using average retention indices to balance the count

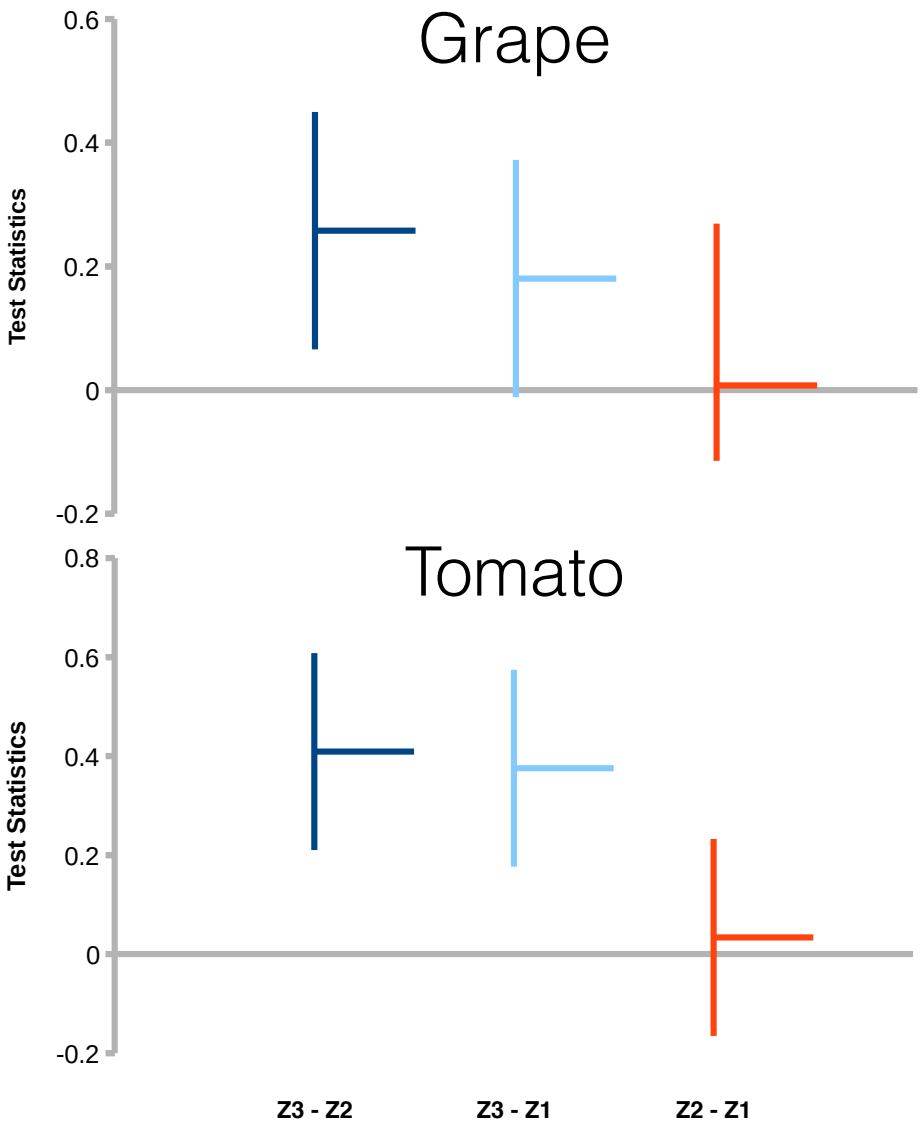


FIGURE 5.3: Tukey's honest significant difference test. The horizontal bar indicate the test statistics of the estimated difference between labelled categories. The vertical lines indicate the 95% confidence interval. In both grape and tomato, category Z2 and Z1, in red, are not significantly different from each other. In grape, category Z3 and Z2, in light blue is not significant (adjusted p-value is 0.06887)

TABLE 5.2: ANOVA table on balanced grape and tomato data.

Grape Anova Table (Type II tests)	Partial eta^2	Sum Sq	Df	F value	Pr(>F)	
GOf	0.9071	36.193	3	97.64771	<1e-15	***
ExpQ	0.06236	0.121	1	1.9953	0.1681	
GOf:ExpQ	0.02797	0.017	2	0.4317	0.6534	
Residuals		1.0918	30			
Tomato Anova Table (Type II tests)						
GOf	0.96865	36.193	3	308.9591	<2e-16	***
ExpQ	0.0937	0.121	1	3.1016	0.08841	.
GOf:ExpQ	0.01395	0.017	2	0.2121	0.81005	
Residuals		1.171	30			

GOf is the High level functional category. ExpQ is the expression category.

data.

5.5 Discussion

Expression has been suggested to be the biggest factor in determining duplicate retention after whole genome duplication events [27]. However, our results suggest that functional category is the more dominant factor of the two.

The simplest way to reconcile these contradictory results of Gene Dosage from Gout *et al.* [27] and Gene Balance Hypothesis from Birchler *et al.* [7] is to consider that some functional categories contributes to duplicate retention drastically more than others such

that the effect of expression on duplicate retention may be greater for genes of some functional categories.

Another explanation is that the inherent correlation between gene function and expression levels may introduce enough bias against crediting expression levels the proper amount of impact on fractionation. Our results in Table 5.2 suggests that there is no interaction between functional categories and expression levels.

Going forward, we want to further explore the role of expression on fractionation. One direction is to explore the different types of expression. Some genes are only expressed in certain stages of life, such as the development of flowers, or genes that have organ specific expression pattern, or genes that are always on but fluctuates depending on the situation. This is analogous to different functional categories having different fractionation tendencies.

Another direction would be to expand the analysis to other genes that are currently not part of the analysis. One particular analysis for future work is the relationship between retained duplicates and the nearby genes. Retained duplicates is reported to have an effect on the distribution of genes with copy number variation in humans [85]. We can explore if similar effects are also present in plants.

In summary, we have evidence to suggest that functional categories play a more important than gene expression levels in duplicate gene retention after whole genome duplication.

There are many challenges and possibilities that can build upon this work to better explain the mechanisms and the effects of the fractionation process.

Author's contributions'

EC and DS did the research and wrote the manuscript.

Chapter 6

Conclusion and future direction

6.1 Concluding statements

In this thesis, we analyzed the effect of functional categories and gene expression levels on fractionation. We first presented a workflow (Chapter 2) for identifying gene families related by the common ancestry of present-day genes in polyploid last common ancestor. We then applied it to multiple genome data sets to analyze the effect of functional categories in Chapter 3. We studied the comparison between functional categories and expression levels on duplication gene retention in Chapter 4 and Chapter 5.

We draw our conclusion from functional analysis on three genomic data sets and from expression analysis on representative species of two of the three genomic data sets. Each

data set represents a different scenario regarding whole genome duplication, as first described in Chapter 3. In summary, we find both functional categories and expression levels to be important in the fractionation process, with functional categories having greater impact on duplicate gene retention than expression levels.

6.1.1 Effect of functional category

We find that functional category of the genes does have an effect on duplicate gene retention post-whole genome duplication events. We have "metabolic process" as the most fractionation prone functional category and "catalytic activity" being a close second. This makes sense because enzymatic genes and metabolic genes have a large degree of overlap with many metabolic intermediates requiring enzymes to catalyze the reaction. Metabolic genes also mostly interact with metabolites such as fat and sugar, and not with other gene products. The most fractionation resistant functional categories are "biological regulation" and "response to stimulus". Again, both of these functional categories share a large overlap of genes whose products interact with other gene products.

As a whole these results are as predicted by the Gene Balance Hypothesis as presented in the latest review by Conant *et al.* [20] since the most fractionation resistant functional category contains genes typically involved in complex regulatory systems. The most fractionation prone functional category contains gene typically not involved in multi-gene-networks where the stoichiometry ratios of the gene product is extremely important.

Another important result from these datasets is that the duplicate gene retention pattern is consistent across all three datasets. Genomes of the rosids and the asterids datasets have comparable genome sizes of around 25000 genes or more [12, 14, 29, 31, 35, 43] and the additional rounds of whole genome duplication events results in increased scrambling of the genome; yielding lower number of gene families compared to the rosids data set. The Poaceae data set of monocot species also show very similar retention patterns despite being evolutionarily distant from the rosids and the asterids dataset (Chapter 3). Even if we relax the criteria for gene families (Chapter 3) we still observe the same duplicate gene retention pattern with respect to the functional categories. This suggest that the effect of functional category on gene retention post-whole genome duplication events are general, at least within the angiosperms.

6.1.2 Effect of expression

Earlier work done on *Paramecium* by Gout *et al.* [27], the inspiration for Chapter 4 and 5, suggests that expression is the most important factor in duplicate gene retention; although they do note the role of gene function in expression levels. By controlling bias from functional categories we do see clear trends in increased expression with increased duplicate gene retention across. Also in agreement is that this observation, to be true for vast majority of the functional categories (Chapter 4) where there are enough genes to perform a reasonable statistical test. However upon further analysis (Chapter 5) we were

able to show that functional categories have a bigger effects on duplicate gene retention post whole genome duplication events.

6.2 Future extensions on the study of whole genome duplications

From the results of this thesis there are a few different directions to take to further the study of whole genome duplications. Possible new directions include but are not limited to: further method development in aggregating sequencing data, novel experimental designs on synthetic polyploids, broader generalization of duplicate gene retention in more diverse lineages, and the effects these duplicate genes have on current genomes.

6.2.1 RNA-seq normalization

In this thesis, the expression analysis is only done on the rosids and the asterids dataset because of the availability of high quality RNA-seq data that cover multiple organs. There currently are no good way to compare different RNA-seq data across different platforms and different depth of sequencing.

A recent comparative study on the various RNA-seq platforms and strategies by Li *et al.* [91] has been encouraging on the potential of meaningful expression analysis across

RNA-seq data obtained from different platforms. I remain hopeful that the advancement in RNA-seq algorithms and the continuing maturation of RNA-seq technologies will open the door to more precise and more elaborate representation of gene expression data.

6.2.2 Triangle of U

Another natural direction is to analyze the genomes of genus *Brassica* that make up the "triangle of U" [92]. These genomes are of agricultural interest because various important crop species are actually hybridizations of *Brassica rapa*, *Brassica nigra*, and *Brassica oleracea* [92, 93]. The *Brassica* species datasets will pose an important extension to this work by offering a look at duplicate gene retention pattern from a comparatively very recent (relative to the rosids, asterids, and the monocots datasets) whole genome duplication events.

6.2.3 Whole genome duplication in Salmonidae

As first mentioned in Chapter 1, flowering plants are not the only lineages that experienced whole genome duplication events. The family of Salmonidae, which contains economically important species such as the atlantic salmon, the pacific salmon, the rainbow trout, the freshwater whitefishes, and more [94], are all derive from a common ancestor that experienced whole genome duplication [95]. From this tetraploid ancestor of 4n =

96 chromosomes, massive genome reorganization is hypothesized to have taken place to produce current diploid species that have a diverse chromosome number (from 2n = 54 to more than 80) [94].

In particular, the rainbow trout genome paper by Berthelot *et al.* [96] showed functional enrichment after whole genome duplication similar to functional categories reported in this thesis. Both the rainbow trout paper and Chapter 3 of this thesis reported the enrichment of transcription factor genes. Furthermore, in Chapter 3, one of the reported fractionation resistant GO terms is "developmental process". This is a high level GO term which encompasses both the embryonic development and the synaptic development terms that Berthelot *et al.* reported to be enriched in the rainbow trout paper. The rainbow trout paper also reported the enrichment of GO term "receptor activities", which is also fractionation resistant in our analysis but is too specific a category to be included in Chapter 3.

The expression analysis portion of the rainbow trout paper focused only on genes that are retained as duplicates [96]. They found functional enrichment based on the expression pattern and expression level of the duplicates. For example, metabolic and house-keeping genes are enriched in duplicates where they are highly correlated with each other on which organs they are expressed but have divergent expression levels. Developmental process genes are mostly enriched in duplicates that have similar expression levels and the same

expression pattern. The exceptions are eye development and vision perception genes which are enriched when the duplicates does not have correlated expression patterns.

The rainbow trout expression analysis complements the functional and expression analysis presented in this thesis. It is worth examining if a similar result can be replicated on plant datasets, or if these results will change when we include singletons.

Berthelot *et al.* [96] also reported high collinearity between the paralogous regions from the newest whole genome duplication in fish as evidence against the current hypothesis of rapid gene fractionation and genome rearrangements following whole genome duplication. This claim is echoed by Uno *et al.* [97], where they reported that rapid gene loss and genome rearrangements are not essential for the diploidization of the frog model organism's allotetraploid ancestor.

These results are in contrast with experiments on the resynthesized allopolyploid plant *Brassica napus* by Xiong *et al.* [98]. They showed, by the tenth generation of the resynthesized allopolyploid, involvement of aneuploidity, a great number of chromosome rearrangements, and dosage constraints maintained the chromosome number close to the initial resynthesized polyploid. More experiments and a standardized metric for measuring genome rearrangements are needed to clarify the necessity and the extent of genome rearrangements following whole genome duplications.

6.2.4 Gene number variation deserts in vertebrates

Finally, yet another direction to take is to analyze the effects duplicate gene retention have on current genomes. Makino *et al.* [85] reports that neighbouring regions of dosage sensitive genes are copy number variation deserts. They argue that this is an efficient way of identifying diseases related to copy number variation.

Before analyzing diseases related to copy number variations in plants, a survey to confirm whether this pattern of gene number variation deserts in plants is needed. Also, it is important to identify if retained duplicate genes are surrounded by other high expressing genes or are part of a chromosome region of elevated expression. This will help to develop models to explain why high expression seems to be universally important in duplicate gene retention.

Bibliography

[1] Ohno S, et al.: *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. 1970.

[2] Force A, Lynch M, Pickett FB, Amores A, Yan Yl, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**(4):1531–1545.

[3] Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast**. *Nature* 2003, **424**:194–197.

[4] Birchler JA, Riddle NC, Auger DL, Veitia RA: **Dosage balance in gene regulation: biological implications**. *Trends in Genetics* 2005, **21**(4):219–226.

[5] Edger PP, Pires JC: **Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes**. *Chromosome Research* 2009, **17**(5):699–717.

[6] Freeling M: **Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition**. *Annual review of plant biology* 2009, **60**:433–453.

[7] Birchler JA, Veitia RA: **Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines**. *Proceedings of the National Academy of Sciences* 2012, **109**(37):14746–14753.

[8] Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Wall PK, Soltis PS, et al.: **Polyploidy and angiosperm diversification**. *American journal of botany* 2009, **96**:336–348.

[9] Hegarty M, Coate J, Sherman-Broyles S, Abbott R, Hiscock S, Doyle J: **Lessons from natural and artificial polyploids in higher plants**. *Cytogenet Genome Res* 2013, **140**(2-4):204–225.

[10] Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M: **Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs.** *PLoS Biol* 2010, **8**:e1000409. [Doi:10.1371/journal.pbio.1000409].

[11] Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X: **Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa**. *PLoS One* 2012, **7**(5):e36442.

[12] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyére C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**:463–467.

[13] Soltis DE, Albert VA, Leebens-Mack J, Palmer JD, Wing RA, dePamphilis CW, Ma H, Carlson JE, Altman N, Kim S, et al.: **The Amborella genome: an evolutionary reference for plant biology**. *Genome Biol* 2008, **9**(402):10–1186.

[14] Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJ, Simpson J, Fernández-Cortés

A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, de Jesús Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L: **Architecture and evolution of a minute plant genome**. *Nature* 2013, **498**:94–98.

[15] Byrnes JK, Morris GP, Li WH: **Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion**. *Molecular biology and evolution* 2006, **23**(6):1136–1143.

[16] Wang B, Zheng C, Sankoff D: **Fractionation statistics**. *BMC bioinformatics* 2011, **12**(Suppl 9):S5.

[17] Sankoff D, Zheng C, Wang B: **A model for biased fractionation after whole genome duplication**. *BMC genomics* 2012, **13**(Suppl 1):S8.

[18] Rubio-Somoza I, Cuperus JT, Weigel D, Carrington JC: **Regulation and functional specialization of small RNA–target nodes during plant development**. *Current opinion in plant biology* 2009, **12**(5):622–627.

[19] Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Camara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouel A, Lepere G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Betermier M, Weissenbach J, Scarpelli C, Schachter V, Sperling L, Meyer E, Cohen J, Wincker P: **Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia**. *Nature* 2006, **444**(7116):171–178.

[20] Conant GC, Birchler JA, Pires JC: **Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time**. *Current opinion in plant biology* 2014, **19**:91–98.

[21] Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae**. *Nature* 2004, **428**(6983):617–624.

[22] Gout JF, Duret L, Kahn D: **Differential retention of metabolic genes following whole-genome duplication**. *Molecular biology and evolution* 2009, **26**(5):1067–1072.

[23] Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution**. *The Plant Cell* 2004, **16**:1679–1691.

[24] Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP: **Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus**. *Genome Research* 2012, **22**:95–105.

[25] Bekaert M, Edger PP, Pires JC, Conant GC: **Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints**. *The Plant Cell Online* 2011, **23**(5):1719–1728.

[26] Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR: **Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa***. *The Plant Cell Online* 2012, **24**(6):2415–2426.

[27] Gout JF, Kahn D, Duret L, Paramecium Post-Genomics Consortium: **The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution**. *PLoS Genet* 2010, **6**(5):e1000944.

[28] Wu X, Qi X: **Genes encoding hub and bottleneck enzymes of the *Arabidopsis* metabolic network preferentially retain homeologs through whole genome duplication**. *BMC Evolutionary Biology* 2010, **10**:145.

[29] Jung S, Cestaro A, Troggio M, Main D, Zheng P, Cho I, Folta KM, Sosinski B, A A, Celton JM, Arús P, Shulaev V, Verde I, Morgante M, Rokhsar DS, Velasco R, Sargent DJ: **Whole genome comparisons of *Fragaria*, *Prunus* and *Malus***

reveal different modes of evolution between Rosaceous subfamilies. *BMC Genomics* 2012, **13**:129.

[30] Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**:635–641.

[31] US Department of Energy JGI: ***Mimulus* version 1** 2010, [http://www.phytozome.net/mimulus].

[32] Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, ur Rahman M, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS: **The *Sorghum bicolor* genome and the diversification of grasses**. *Nature* 2009, **457**:551–556.

[33] Kawahara Y, de la Bastide M, Hamilton J, Kanamori H, McCombie WR, Ouyang S, Schwartz D, Tanaka T, Wu J, Zhou S, Childs K, Davidson R, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T: **Improvement of the *Oryza sativa Nipponbare* reference genome using next generation sequence and optical map data**. *Rice* 2013, **6**:4.

[34] Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, Jenkins J, Barry K, Lindquist E, Hellsten U, Deshpande S, Wang X, Wu X, Mitros T, Triplett J, Yang X, Ye CY, Mauro-Herrera M, Wang L, Li P, Sharma M, Sharma R, Ronald PC, Panaud O, Kellogg EA, Brutnell TP, Doust AN, Tuskan GA, Rokhsar D, Devos KM: **Reference genome sequence of the model plant *Setaria***. *Nature Biotechnology* 2012, **30**:555–561.

[35] The International *Brachypodium* Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon***. *Nature* 2010, **463**:763–768.

[36] Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences**. *The Plant Journal* 2008, **53**:661–673.

[37] Zheng C, Swenson K, Lyons E, Sankoff D: **OMG! Orthologs in multiple genomes - competing graph-theoretical formulations**. In *Algorithms in Bioinformatics*. Edited by Przytycka T, Sagot MF, WABI 2011, 11th Workshop on Algorithms in Bioinformatics 2011:364–375.

[38] Zheng C, Chen E, Albert VA, Lyons E, Sankoff D: **Ancient eudicot hexaploidy meets ancestral eurosid gene order**. *BMC Genomics* 2013, **14**(Suppl 7):S3.

[39] The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29. [Data Version 2012-04-20].

[40] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674–3676.

[41] Vitulo N, Forcato C, Carpinelli E, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C, Lucchin M, Valle G: **A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype**. *BMC Plant Biology* 2014, **14**:99.

[42] Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps**. *Genome research* 2008, **18**(12):1944–1954.

[43] Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Risterucci A, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quétier F, Panaud O, Wincker

P, Bocs S, Lanaud C: **The genome of *Theobroma cacao***. *Nature Genetics* 2011, **43**:101–108.

[44] Salse J: **In silico archeogenomics unveils modern plant genome organisation, regulation and evolution**. *Current opinion in plant biology* 2012, **15**(2):122–130.

[45] Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al.: **A genome triplication associated with early diversification of the core eudicots**. *Genome Biol* 2012, **13**:R3.

[46] Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J: **Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution**. *Genome research* 2010, **20**(11):1545–1557.

[47] Zheng C, Sankoff D: **Gene order in Rosid phylogeny, inferred from pairwise syntenies among extant genomes**. *BMC bioinformatics* 2012, **13**(Suppl 10):S9.

[48] Jung S, Cestaro A, Troggio M, Main D, Zheng P, Cho I, Folta KM, Sosinski B, Abbott A, Celton JM, et al.: **Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies**. *BMC genomics* 2012, **13**:129.

[49] Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al.: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PloS one* 2007, **2**(12):e1326.

[50] Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al.: **The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)**. *Nature* 2008, **452**(7190):991–996.

[51] Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al.: **Draft genome sequence of the oilseed species Ricinus communis**. *Nature biotechnology* 2010, **28**(9):951–956.

[52] Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al.: **The genome of woodland strawberry (Fragaria vesca)**. *Nature genetics* 2011, **43**(2):109–116.

[53] Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al.: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *science* 2006, **313**(5793):1596–1604.

[54] Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M: **Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids**. *Plant Physiology* 2008, **148**:1772–1781.

[55] Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M: **Screening synteny blocks in pairwise genome comparisons through integer programming**. *BMC bioinformatics* 2011, **12**:102.

[56] Galil Z: **Efficient algorithms for finding maximum matching in graphs**. *ACM Computing Surveys (CSUR)* 1986, **18**:23–38.

[57] El-Mabrouk N, Sankoff D: **The reconstruction of doubled genomes**. *SIAM Journal on Computing* 2003, **32**(3):754–792.

[58] Warren R, Sankoff D: **Genome aliquoting revisited**. *Journal of Computational Biology* 2011, **18**(9):1065–1075.

[59] Nadeau JH, Sankoff D: **Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution**. *Genetics* 1997, **147**(3):1259–1266.

[60] Chain FJ, Dushoff J, Evans BJ: **The odds of duplicate gene persistence after polyploidization**. *BMC genomics* 2011, **12**:599.

[61] Zheng C: **Pathgroups, a dynamic data structure for genome reconstruction problems**. *Bioinformatics* 2010, **26**(13):1587–1594.

[62] Xu AW: **On exploring genome rearrangement phylogenetic patterns**. In *Comparative Genomics*, Springer 2010:121–136.

[63] Tannier E, Zheng C, Sankoff D: **Multichromosomal median and halving problems under different genomic distances**. *BMC bioinformatics* 2009, **10**:120.

[64] Boyd S, Haghighi M: **A fast method for large-scale multichromosomal breakpoint median problems**. *Journal of Bioinformatics and Computational Biology* 2012, **10**(01).

[65] Edmonds J: **Paths, trees, and flowers**. *Canadian Journal of mathematics* 1965, **17**(3):449–467.

[66] Comai L: **Genetic and epigenetic interactions in allopolyploid plants**. *Plant Molecular Biology* 2000, **43**:387–399.

[67] Crow KD, Wagner GP: **What is the role of genome duplication in the evolution of complexity and diversity?** *Molecular Biology and Evolution* 2006, **23**(5):887–892.

[68] Birchler JA, Veitia RA: **The Gene Balance Hypothesis: implications for gene regulation, quantitative traits and evolution**. *New Phytologist* 2010, **186**:54–62. [Doi:10.1111/j.1469-8137.2009.03087.x].

[69] Zheng C, Sankoff D: **Practical aliquoting of flowering plant genomes**. *BMC Bioinformatics* 2013, **14**(S-15):S8.

[70] Duarte J, Wall PK, Edger P, Landherr L, Ma H, Pires JC, Leebens-Mack J, dePamphilis C: **Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels**. *BMC Evolutionary Biology* 2010, **10**:61.

[71] Abrouk M, Murat F, Pont C, Messing J, Jackson S, Faraut T, Tannier E, Plomion C, Cooke R, Feuillet C, et al.: **Palaeogenomics of plants: synteny-based modelling of extinct ancestors**. *Trends in Plant Science* 2010, **15**(9):479–487.

[72] Schnable JC, Freeling M, Lyons E: **Genome-wide analysis of syntenic gene deletion in the grasses**. *Genome Biology and Evolution* 2012, **4**:265–277.

[73] Sweetlove LJ, Fernie AR: **The Spatial Organization of Metabolism Within the Plant Cell**. *Annual Review of Plant Biology* 2013, **64**:723–746.

[74] Lynch M: **Evolutionary diversification of the multimeric states of proteins**. *Proceedings of the National Academy of Sciences USA* 2013. [Published ahead of print July 8, 2013].

[75] Nussinov R, Ma B, Tsai CJ: **A broad view of scaffolding suggests that scaffolding proteins can actively control regulation and signaling of multienzyme complexes through allostery**. *Biochim Biophys Acta* 2013, **1834**:820–829.

[76] Schnable JC, Wang X, Pires JC, Freeling M: **Escape from preferential retention following repeated whole genome duplications in plants**. *Frontiers in Plant Science* 2012, **3**:94.

[77] Cannon S, Mitra A, Baumgarten A, Young N, May G: **The roles of segmental and tandem gene duplication in the evolution of large gene families in** *Arabidopsis thaliana*. *BMC Plant Biology* 2004, **4**:10.

[78] Brett D, Pospisil H, Valcárcel J, Reich J, Bork P: **Alternative splicing and genome complexity.** *Nature Genetics* 2002, **30**:29–30.

[79] Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al.: **Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Research* 2003, **31**(19):5654–5666.

[80] Sémon M, Wolfe KH: **Consequences of genome duplication**. *Current Opinion in Genetics & Development* 2007, **17**(6):505–512.

[81] Schnable JC, Wang X, Pires JC, Freeling M: **Escape from preferential retention following repeated whole genome duplication in plants**. *Frontiers in Plant Science* 2012, **3**(94).

[82] Chen ECH, Najar CBA, Zheng C, Brandts A, Lyons E, Tang H, Carretero-Paulet L, Albert VA, Sankoff D: **The dynamics of functional classes of plant genes in rediploidized ancient polyploids**. *BMC Bioinformatics* 2013, **14**(S-15):S19.

[83] Lyons E, Pedersen B, Kane J, Freeling M: **The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids**. *Tropical Plant Biology* 2008, **1**(3):181–190.

[84] Schnable JC, Pedersen BS, Subramaniam S, Freeling M: **Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses**. *Frontiers in Plant Science* 2011, **2**(2). [Doi:10.3389/fpls.2011.00002].

[85] Makino T, McLysaght A, Kawata M: **Genome-wide deserts for copy number variation in vertebrates**. *Nature Communications* 2013, **4**.

[86] Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: **Two evolutionarily distinct classes of paleopolyploidy**. *Molecular biology and evolution* 2014, **31**(2):448–454.

[87] Thomas BC, Pedersen B, Freeling M: **Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes**. *Genome Research* 2006, **16**(7):934–946.

[88] Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu SH: **Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species**. *The Plant Cell Online* 2014, :tpc–114.

[89] Hardison RC: **A brief history of hemoglobins: plant, animal, protist, and bacteria.** *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(12):5675.

[90] Chen EC, Sankoff D: **Gene expression and fractionation resistance**. *BMC genomics* 2014, **15**(Suppl 6):S19.

[91] Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, et al.: **Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study**. *Nature biotechnology* 2014, **32**(9):915–925.

[92] Nagaharu U: **Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization**. *Jap J Bot* 1935, **7**:389–452.

[93] Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, et al.: **The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes**. *Nature communications* 2014, **5**.

[94] Davidson WS: **Understanding salmonid biology from the Atlantic salmon genome 1**. *Genome* 2013, **56**(10):548–550.

[95] Allendorf FW, Thorgaard GH: **Tetraploidy and the evolution of salmonid fishes**. In *Evolutionary genetics of fishes*, Springer 1984:1–53.

[96] Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al.: **The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates**. *Nature communications* 2014, **5**.

[97] Uno Y, Nishida C, Takagi C, Ueno N, Matsuda Y: **Homoeologous chromosomes of Xenopus laevis are highly conserved after whole-genome duplication**. *Heredity* 2013, **111**(5):430–436.

[98] Xiong Z, Gaeta RT, Pires JC: **Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus**. *Proceedings of the National Academy of Sciences* 2011, **108**(19):7908–7913.