# Mathematical Models for Evolution of Genome Structure

Suja Thomas

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering.

Chapel Hill
2010

Approved by:

Todd J. Vision, Advisor,
Committee Co-Chair

Shawn M. Gomez, Committee Co-Chair

Morgan C. Giddings, Committee Member

Oleg Favorov, Committee Member

Jeffrey Thorne, Committee Member

# Abstract

**Suja Thomas: Mathematical Models for Evolution of Genome Structure.**
**(Under the direction of Todd J. Vision.)**

The structure of a genome can be characterized by its gene content. Evolution of genome structure in closely related species can be studied by examining their synteny or conserved gene order and content. A variety of evolutionary rearrangements like polyploidy, inversions, transpositions, translocations, gene duplication and gene loss degrade synteny over time. In this dissertation, I approach the problem of understanding synteny in genomes and how far back its evolutionary history can be traced in multiple ways. First, I present a probabilistic model of the rearrangements gene loss and transposition (gain) and apply it to the problem of estimating the relative contribution of these rearrangements within a set of syntenic genome segments. This model can be used to predict gene content in syntenic regions of unsequenced genomes. Next, I use optimization methods to recover syntenic segments between genomes based on reconstructions of their parent ancestry. I examine how these reconstructions can be used as input to programs that identify syntenic regions in genomes to reveal more synteny than was previously detected. I use simulations that incorporate each of the evolutionary rearrangements described above to evaluate the models presented in this dissertation. Finally, I apply these models to genomic data from yeast and flowering plants, two eukaryotic systems that are known to have experienced polyploidy. This application is of particular relevance in flowering plants, in which a lot of economically and scientifically important polyploid species have incompletely sequenced genomes.

To my parents, Thomas Verghese and Mollykutty Thomas, who gave me life and nurtured all my dreams and ambitions.

# Acknowledgments

I am grateful to my advisor, Dr. Todd J. Vision, for guiding me through a fascinating attempt to unravel part of the wonderful workings of our universe. I also acknowledge current lab members Thomas Clarke and Dr. Lex Flagel for their advice and for proof-reading the chapters included here. I thank lab alumni Dr. Eric Ganko, Dr. Amy Bouck, Dr. Jixin Deng and Dr. Stephanie Hartmann for helpful discussions. I also thank Dr. Jijun Tang, Dr. Laszlo Szekely, Dr. Eva Czabarka and Yuhui Dong at the University of South Carolina, Columbia, for exciting collaborations on topics covered in this dissertation.

A dissertation is not simply a catalogue of research required to earn a degree. To essay an undertaking of this magnitude would be impossible without the support of family and friends. I am very grateful to my family for their guidance and tireless encouragement through each of my endeavours. My parents Thomas Verghese and Mollykutty Thomas surrounded me with reading and books early on in my childhood, two gifts that opened a window to the many wonders of science. My sisters Dr. Susan Thomas, Dr. Sally Thomas and Sunita Thomas are and have always been inspiring role models to emulate. Over the years it has taken me to complete this dissertation, I have built a family away from home here in Chapel Hill. This family consists of friends I made through graduate school and the community. Many thanks to Sharif Razzaque, Caroline Green, Rakhi Kilaru, Elena Taranova and Cris Ledon-Rettig for meticulously reading drafts of chapters of my dissertation and providing valuable feedback. Special thanks also to Timothy Meehan, Montek Singh, Biljana Djukic, Victoria Graham, Allister Bernard, Maria Razzoli, Lindsay Dubbs, Maria Christensson, Mary Lindsley, Monika Lichtinger and Laura Faulconer for many hours of help, advice, love, cooking and listening.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The arrival of full genome sequencing in the early 21st century is perhaps the most significant development in the field of genomics (3; 4). We can now obtain the DNA sequences of different organisms and compare them with each other. The first few eukaryotic genomes to be sequenced were those of the yeast *Saccharomyces cerevisiae*, worm *Caenorhabditis elegans*, fruitfly *Drosophila melanogaster* and plant *Arabidopsis thaliana* (5; 6; 7; 8). The first complete human genome was sequenced in 2001 (4).

Sequencing greatly enhanced building genetic maps of different organisms where previously, molecular techniques like restriction fragment length polymorphism (RFLP) were used in identifying and isolating genetic markers (9). Markers like these which were produced in plants like tomato (10), corn and wheat (11; 12) are very useful to breeders for agronomic cultivation purposes. As more markers were produced and more sequences obtained, it became clear that many genetic markers were conserved amongst species both in content and order, also known as *synteny*.

Genome sequences of closely related species do not share much similarity in entirety, but their regions that encode for genes do. Only about 5% of the entire human DNA sequence is currently implicated in coding for a total of 24,800 verified proteins. Amongst these, humans share 70-90% of their gene content with mice, and 95-98% with apes. For a set of species that are so organismically and morphologically different from each other,

this may seem like an extraordinary amount of genic material to have in common. To characterize this property and use it to predict protein-coding sequences, many sophisticated models of sequence evolution were developed. They are reviewed comprehensively in (13).

The structure of a genome can be characterized by its gene content. A comparison of genome structure amongst different organisms can be done through a comparison of their synteny. Evolutionary rearrangements degrade synteny over time. Modeling the effects of rearrangements on synteny can inform us about the changes produced in genome structure and age of preservation of synteny.

This dissertation contributes to the study of evolution of genome structure by examining it through synteny in genomes and our ability to detect how far back in time we can trace it. This study is approached in the following ways. First, I define a probabilistic model of the processes of gene loss and gene transposition or gain and apply it to the problem of estimating their relative contribution to gene order within a set of syntenic genomic regions. This model can be used to predict gene content in syntenic regions of unsequenced genomes. Second, I apply optimization methods to reconstruct the ancestral gene order of syntenic regions within genomes generated by simulations. Third, I evaluate using the reconstructions along with an existing method used to identify pairwise synteny to see if there is a gain in synteny detection over using pairwise and profile synteny detection methods on simulated as well as on plant genomic data.

In this chapter, I provide a brief biological background for understanding this work. I also review the existing statistical and mathematical models for understanding different aspects of genome structure evolution. I then discuss the importance of synteny. Finally, I will introduce the questions that are studied in this dissertation and the motivation for doing so.

## 1.1 The Biology of Genome Structure Evolution

Synteny is preserved in mammalian genomes (14) since the last inferred common ancestor. Synteny is also shown to be highly preserved in yeasts (15). As the rates of rearrangements that shuffle gene order in mammals are relatively low and their genes are highly collinear in order, diagnosing synteny is relatively easy. Synteny between the first sequenced flowering plant or angiosperm genomes Arabidopsis and rice was shown in (16) and amongst the Arabidopsis, Carica and Populus genomes in (17). There is also a high degree of synteny between the grass genomes of maize, sorghum, rice, sugarcane, foxtail millet, pearl millet, Triticeae and oats (18). However, in flowering plants or angiosperms higher rates of rearrangement intervene in this preservation of collinearity (19).

A variety of rearrangements create disruption in genomic synteny. In this dissertation, I consider the following rearrangements: Whole Genome Duplication (WGD) or polyploidy, dispersed single gene duplications, inversions, translocations, transpositions and gene loss. In this section, I will review what is known about the impact of these rearrangements on genome structure evolution.

Figure 1.1 is an illustration of how the different processes contribute to difference in gene order from a starting ancestral state.

In 1970, Ohno proposed that gene duplication played a major role in evolution (20) and suggested that the vertebrate genome is the result of one or more entire genome duplications. Polyploidy or WGD has occurred many times in the eukaryotic lineage (21; 22) with at least one inferred WGD event in the last common ancestor (23). The first ancient WGD was shown for eukaryotes in yeast (24). It was shown in plants (25; 26; 27), teleost fish (28) and in paramecium (29). It has been estimated that about 2-4 % of speciation events are associated with polyploidy in flowering plants (30). It has been implicated in giving rise to species-rich groups in both plants and animals (31; 21). There are many current-day polyploid species, particularly in plants with

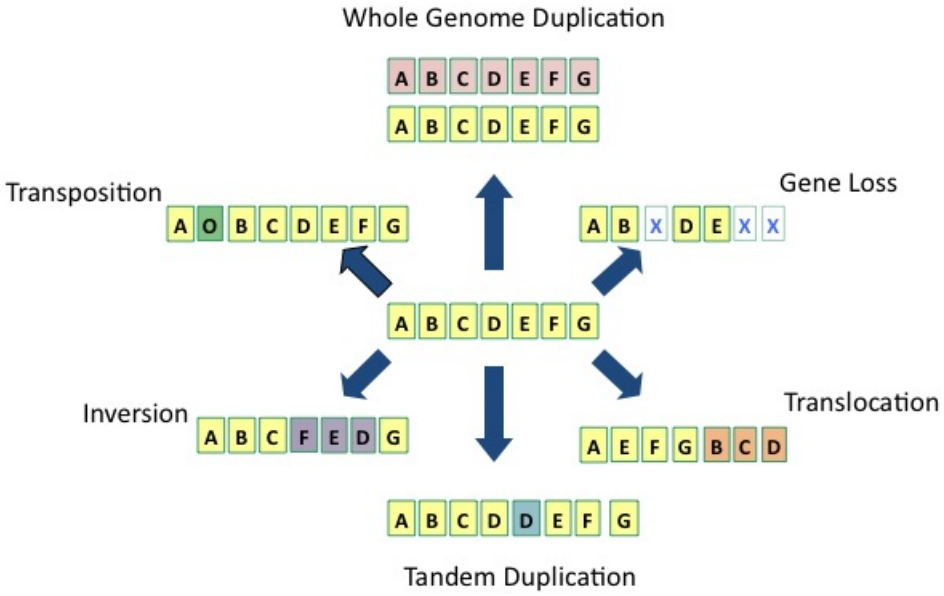


Figure 1.1: Illustration of evolutionary rearrangement processes for a genome segment with genes A, B, ..., G in the center of the figure. The rearrangements are denoted in colours different from the original genome

estimates ranging from 30 to 70 % (32). However, only a few ancient polyploidy events are thought to have survived (33).

Polyploidy has a big impact on gene order rearrangement because of the large number of genes that are duplicated during each event (34). It has been shown to precipitate massive gene loss , which is a major contributor to divergence among descendant polyploid genomes (35; 36). Asymmetric gene loss following polyploidy has been shown to obscure synteny between genomes, making its detection difficult (25).

Gene duplication on a smaller scale also plays an important role in gene content and order rearrangement (37). It can arise through tandem and segmental duplications (during DNA replication and recombination for example).Transposition of duplicated

genes by transposable elements by transduplication or stimulation of intrachromosomal recombination events (38; 39) can also interrupt gene order. 15-20% of the gene content of Arabidopsis and rice consists of tandemly arrayed gene clusters (40).

Large-scale duplication generated by polyploidy and smaller-scale gene duplications are not necessarily exclusive processes (37; 34). Transposition of genes by transposable elements has been shown to coincide with polyploidy (41). These two kinds of duplication have different effects on synteny. A single-gene duplication might not occur within the right regulatory context needed, or with all the required sequence required for its correct expression. This will affect its probabilities of retention in the organism (42), which in turn influences its contribution to synteny. With polyploidy however, the gene and its entire context are duplicated.

Inversions and translocations that occur at the scale of chromosomes have been known to occur in eukaryotes (43). Estimates from different organisms suggest that chromosomal scale inversions can occur at different rates in different organisms (44; 45; 46). Larger scale inversions have a more immediate impact on gene order between organisms. However, smaller-scale inversions that are only a few genes in length could be happening at a much more rapid rate. Added up over a period of time, this could produce considerable rearrangement over a larger size of the chromosomes that they occur in (34). Such inversions have been thought to rearrange gene order in organisms like yeast (47), but were considered to be rarer in plants (36). However, the cereal maize has to shown to have experience a high rate of inversions in comparison with rice (46).

The impact of each of these rearrangement on gene order vary on many levels. They occur with different rates within different lineages and in different organisms. A process like polyploidy operating on a genome scale automatically has a larger impact on gene order than a local inversion, for example. Rearrangements also do not occur indiscriminately in genomes. It has been shown that functionally related genes are preferentially retained over those that are not (48). In this dissertation, I account for these chro-

mosomal rearrangements and examine how informative they are in modeling synteny evolution.

## 1.2 Mathematical Models of Genome Structure Evolution

In this section, I review the history of mathematical models of genome structure evolution, starting with models of sequence evolution and proceeding to models of gene content and order evolution.

Prior to the new era of sequencing, there were a variety of models used in comparative genomics. In 1936, Dobzhansky and Sturtevant first proposed to use the amount of disorder between the gene order in two different genomes as an indicator of their evolutionary distance (49; 50). As rearrangements were considered to be relatively rare, the distance that minimized the disorder or was the most *parsimonious* was considered realistic. In fact, Sturtevant and Novitski stated that for numbers of loci greater than 9, this problem was intractable (51; 49). The first studies that examined chromosomes using techniques such as chromosome banding or in-situ hybridization focused on closely-related species, where the number of rearrangements were small (49). Established combinatorial techniques were used to address this parsimony criterion.

Initially, differences in genome structure were studied by a variety of models that examined it through differences in nucleotide sequences. Jukes and Cantor (52) described a model to describe changes from one nucleotide base to the other that was based on the assumption that substitutions are equally probable and that the frequencies of all the four bases in DNA are the same. This was followed by a variety of methods proposed to to estimate phylogenetic trees from sequence data using a probabilistic model of evolution and maximum likelihood (53), in contrast to the methods that traditionally used parsimony to do so till then. Further significant developments in modeling sequence

evolution followed subsequently (54; 55; 56; 57; 58; 59; 60).

Many of these models were evaluated on mitochondrial DNA of organisms, as mitochondrial DNA is small, readily available and much less complicated in structure than nuclear DNA. These models were used to characterize the variation in sequences and also to align sequences pair-wise. Most models developed for sequence evolution and alignment intrinsically assume that changes within sequences occur in a random fashion and that the changes are stochastic in nature. As a result, a lot of the models also used ideas and concepts from probability theory including Bayesian theory, Hidden Markov Models and maximum likelihood. These models are used in many applications like gene prediction, creating phylogenies and inferring rates of evolution in different organisms.

Apart from those methods that model sequence evolution, there are many methods that model genome structure evolution through gene content. A variety of probabilistic methods have been developed to create phylogenies from gene content. Steel and Huson (61) developed a method that models the evolution in the size of the genome with gene loss and horizontal transfer. Gu and Zhang (62) developed a model considering four genomes at a time, with gene loss and duplication in a maximum likelihood framework. Other methods construct phylogeny based on the difference in gene presence-absence content in genomes. Some of these methods model the evolution of presence/absence of genes on a phylogeny in a phyletic nature, while others use a continuous-time Markov process (63). Another set of methods model coevolution of gene content (64). There are also methods that use Bayesian theory with phylogenies to estimate ancestral sequence character states.

With sequencing, the importance of rearrangements to gene order was realized. Many methods have been proposed to measure the number of rearrangements. Palmer and Herbon (65) noticed that the mitochondrial genomes of cabbage and turnip were $\sim$ 99.9% identical in genic sequence, but very different in gene order. Watterson et al (66) stated the problem of representing the relative positions of genes in different genomes as

*permutations* of each other and solving the problem of transforming one into the other with a series of inversions. Sankoff proposed the study of using edit distances to measure gene order rearrangement as an alternative to studying genome divergence through differences in sequence evolution (67). An edit distance between two strings of characters is the number of operations required to transform one string into another. Following this, Pevzner and Waterman reviewed a series of open combinatorial problems to address gene order rearrangements as permutations (68). Sankoff first formulated the inversion distance problem and provided lower and upper bounds for it as well (69). Hannehalli and Pevzner (70) announced the solution to the problem of counting the minimum number of circular inversions (for a circular genome) in polynomial time in 1995. A lot of the initial studies with the inversion distance were performed on mitochondrial and bacterial genomes, which are small, have a higher number of conserved gene content and are circular in shape. Since then, a variety of solutions have been provided for the inversion distance problem, including extensions to linear chromosomes, signed and unsigned permutations (71).

Many other distances have been proposed as well. The breakpoint distance is an edit distance which was first proposed by Sankoff et al (72). This distance is the number of breakpoints or the number of adjacencies in one permutation that are not adjacencies in the other. The authors developed a heuristic to compute the breakpoint distance for genomes that have unequal gene content by calculating induced breakpoint distances (defined in detail in Chapter 3). The authors applied this method to compute a phylogeny for protist genomes (73). The transposition distance was first introduced by Bafna and Pevzner (74). This estimates the distance between two permutations as the number of times a block of contiguous elements is displaced in transforming one to the other. Several polynomial-time approximation algorithms and heuristic approaches have been described to compute it. For multichromosomal genomes, a reciprocal translocation distance was formulated by Kececioglu and Ravi (75), and Hannehalli (76) formulated it

with polynomial complexity. The Double-Cut-and-Join or DCJ distance was proposed by Yancopoulous et al (77) and is computed as the ways two breakpoints in gene order created by rearrangements can be connected back again. This distance measure models breakpoints that are created by inversions, transpositions, fissions, fusions and translocations.

Genome-halving is an algorithm proposed by El-Mabrouk and Sankoff that computes ancestral reconstructions for two genomes, one of which has undergone a WGD since its divergence from the other. This method has been used in reconstructing the pre-WGD ancestor of the yeasts *S. cerevisiae* and *C.glabrata* (78) and the pre-WGD ancestor of Populus (79). *DUPCAR* (80) is a method for reconstructing contiguous ancestral regions with duplications that was used to reconstruct the ancestral chromosome X of placental mammals and the ancestral genome of *Paramecium tetrauerila*.

There are also a variety of programs that have been implemented to reconstruct gene order. GRIMM (81) is a program that has been used to infer the number of rearrangements between human and mouse. MGR (82) is an extension of GRIMM for handling more than a pair of genomes and has been used to infer rearrangements in sequenced mammalian genomes. GRAPPA (83) is a suite of programs that computes several kinds of distances between genomes and computes phylogenetic trees from these distances. It has been used to reconstruct phylogenies for chloroplast genomes and recently, on bacterial genomes (84).

A common feature of all of these methods is that the distances are computed between genomes/segments of equal gene content. When these programs are applied to genomic data sets like in a comparison between human, cat and mouse genomes (81), reconstructions are provided for only the shared gene content between the genomes. This omission of gene content might not affect how the distance between two genomes are calculated with the methods used, but yields an incomplete reconstruction. Gene duplication is also not modeled by these programs. Angiosperm have large multi-gene

families (85). Genome-Halving considered only those sets of duplicates that have corresponding homologs in the reference non-WGD genome. These methods will not yield complete reconstructions for polyploid genomes, as a consequence.

Many algorithms have been developed to detect synteny within and among genomes have been developed for use in many systems. The program *i-ADHoRe* (86) detects synteny through pairwise comparisons and uses these syntenic regions as profiles to collect more regions of synteny. It was used to detect syntenic regions between the Arabidopsis and rice genomes. *FISH* (87) is a statistical method that calculates the probability of detecting syntenic clusters of given sizes in pairwise comparisons and was used to detect syntenic regions within the Arabidopsis genome. *CoGe* (88) provides an integrated Web-based system to find and align syntenic regions and was used to visualize synteny among the Arabidopsis, Popular, Carica and grape genomes. *CloseUp* (89) uses gene density parameters to identify pairwise synteny. These programs use distance between homologous genes on syntenic segments and density of such genes as parameters for searching for synteny. High fragmentation of synteny in a segment through gene loss makes synteny detection in a pairwise comparison difficult.

## 1.3 Importance of Synteny

Synteny amongst different species allows for extrapolating information from one genome to the other. Conserved order of shared genes in two genomes is a strong indicator of their functional and evolutionary relationship. Syntenic genes are markers for homologous regions within and between genomes. This is particularly useful in cases of synteny between genomes that have been well-studied like the model plant rice which is smaller relative to other grasses like barley, wheat and sugarcane that have intractable genomes (90). Grass genome comparisons revealed a high degree of collinearity in gene order and content which was easily visible when a set of conserved segments among them

were assembled into a comparative map (11). If a gene of interest is located within a syntenic region, its location can be narrowed down in the larger intractable genomes by exploiting its collinearity with the smaller, better-sequenced genome and extrapolating from the location of the gene in the smaller genome. The collinearity between the cereal genomes has been maintained since their descent from a common ancestor around 50 million years ago. A similar comparison facilitated the discovery that some important genes involved in domestication and other important traits like selection for large seeds and flowering time that appeared to be at the same loci across multiple grass genomes (12).

Rat, mouse, fruitfly and pufferfish gene models are used in characterizing gene models in the human genome because of their synteny with the human genome. This has been very useful in the field of human medicine and disease as illustrated by these examples of studies in haemophilia, diabetes and cancer (91; 92; 93; 94; 95). Many agronomically and scientifically important plants do not have complete gene maps as yet. There are a handful of plant genomes for which there are genetic maps available. Comparative maps utilizing synteny in plants with model genomes have been used in identifying candidate genes in a variety of plants (96). Arabidopsis thaliana was the first plant to be sequenced and comparison of its genome sequence to that of other plants enabled the study of many important quantitative trait loci, especially those involved with disease resistance (97), water-use efficiency(98) and heat resistance (99). Rice was the first cereal to be fully sequenced. This has spurred a lot of research in science and industry to make strains of rice that are genetically modified to increase production, resist parasites and grow in nutrient-poor regions. The poplar genome was the first tree genome to be sequenced and is a valuable model organisms for further studies on tree genomics.

Synteny comparison between organisms is also very valuable in studying the evolutionary causes for the difference in their genetic make-up and phenotype. The monkeyflower Mimulus is a model organism for the study of genetics and speciation and its

whole genome sequence will facilitate comparative genetics in asterid eudicots. Its relatively small genome of about 430 Mb facilitates the comparison of its genetic make-up with those of other plants to address questions in plant ecological adaptation. The mushroom *Copriopsis cinerea* is a model organism for multicellular development in Agaricomycotina fungi for which a genome sequence is now available (100). Synteny analyses between its genome and that of the another fungus *Laccaria bicolor* enabled the authors of the study to study the presence of key genomic features in Agaricomycotina genomes such as nitrogen metabolism, the cytoskeleton, metabolic regulation, etc.

The relative order of genes in plants at informative positions in the angiosperm phylogenetic tree is also very informative in understanding the evolutionary rearrangements in one plant relative to the other. Gene content and order of the plant genomes that have been sequenced to date have been used to infer WGD events are unique to their lineage as well as shared by groups of lineages. Traces of WGDs are detected when multiple regions in one genome are homologous to that of a region in another genome. For example, it is inferred that Arabidopsis experienced two-three recent WGD events (25). Through genome comparisons, only one of the events is inferred to have been shared by it and Populus, Vitis and Papaya (19). Maize is inferred to have undergone a WGD event since its divergence from Sorghum, whereas the most recent duplication in Sorghum is inferred to be shared with all other cereals (46). By comparing the disruption in collinearity in regions in one plant syntenic to another, rearrangements like inversions and translocations etc. in one plant can be inferred with respect to another (46; 11).

There are some features in plant genomes that make it hard to obtain genetic maps. A lot of flowering plant species are polyploid and fragmentation in patterns of synteny created by WGD events can confound correctly assigning a map location to a gene product. A high rate of repeat elements in plants makes it hard to correctly create

scaffolds in plants. 87% of the sequence in maize consists of repeat elements. There is a remarkable degree of conserved synteny in the plant kingdom, but detecting it is challenging due to the interference of one or many of the processes described above. Unscrambling the puzzle of detecting synteny in the face of these rearrangements is a very valuable tool. Leveraging what is currently known in model plants to those with incomplete sequences is of huge relevance to the studies of evolution in plants, agriculture and ecology.

## 1.4 Questions addressed in this dissertation

It is important to understand how synteny is maintained and to be able to diagnose how far back in time it has been preserved. Consequences of polyploidy can obscure synteny and it will be very useful to model synteny evolution in genomes, particularly in systems that have experienced WGD events for which we do not have complete genetic maps. Current probabilistic models of gene order and evolution do not model WGD events. One objective of this dissertation is to evaluate models of gene loss and gain in polyploid genomes and assess if it can enhance current gene prediction capabilities. I address this in Chapter 2.

The gene order and content in the ancestor of closely related species would be more similar to each of the species gene order and content than they are with each other. Most reconstruction algorithms optimize gene order for genomes that share equal gene content. How accurate these reconstructions are in the context of ancestral WGD events and asymmetric gene loss has not been characterized to date. I test an algorithm that assembles ancestors for genomic regions of unequal gene content in simulations that model this context in Chapter 3. In Chapter 4, I use the information gained about the reconstructions to assess the results of using reconstructions along with a synteny analysis on genomic data of rice and Arabidopsis. I will evaluate the advantages of using

this method over pair-wise synteny detection alone.

# Chapter 2

# A Probabilistic Model of Gene Loss and Gain after Whole Genome Duplication for Predicting Gene Content in Syntenic Segments

## 2.1 Abstract

Gene content among related genome segments diverges primarily through gene loss, particularly following Whole Genome Duplication (WGD) and through transposition. We currently lack tools to quantify the relative importance of these factors and we have limited power to predict what genes are present in related, but poorly characterized, genomic regions. By modeling the process of gene content divergence among homoeologous chromosome segments, I aimed to both predict the content of unsequenced genome regions and provide a statistical framework for studying the divergence process. I developed a probabilistic model of gene loss and gain among genome segments related by a known phylogeny and in the presence of occasional genome duplication events. I found it possible to resolve gene loss immediately following WGD from background gene loss, under what is considered a biologically realistic process. However, I found that it was

not always possible to resolve gene gain accurately. The accuracy in estimating the two gene loss rates and gene gain rate degrades with the amount of data that is missing and with lesser number of segments in the data. I found that predictions of unobserved genes are most enhanced with an increase in the number of genomic segments in the data, rather than the number of genes in the data set and completeness of the segments. I also tested the model on yeast genomic data and found that the predictive capabilities of the model worked as observed in the simulations even though the model was not expected to accurately account for the underlying biological processes. Moreover I found that the rate of loss following WGD is 4 times that of the background loss rate, and 11.5 times that of the gene gain rate.

## 2.2 Introduction

Closely related genomes share conserved gene content and order, or *synteny*. There is evidence for this collinearity of order of conserved genes in flowering plants (101; 16), animals (22; 102) and fungi (24). Synteny between genomes is used in comparative mapping to leverage information from genomes that are fully-mapped in identifying candidate genes in genomes that have intractable maps (11; 103). A variety of crops have been estimated to have only 5% of their constituent genes mapped to correct physical locations in their genomes (90; 104; 105). In particular, synteny has proved to be useful in predicting genes where a phenotypic effect could be linked to a part of an unsequenced genome, but the gene responsible for it could not be readily identified (96; 97; 98; 99). This region of the genome could correspond to a number of unsequenced genes; identifying the gene (or genes) responsible for the effect is as challenging as sequencing all of them. Leveraging synteny to narrow down the candidates, therefore, saves time and effort.

Synteny, however, is often obscured by a variety of evolutionary processes that cause

related genomes to diverge away from each other (19). This furthers the challenge in using it for comparative mapping. Signatures of these processes can be observed in genome sequences as both differences in their nucleotide sequence level and in their gene content. Gene loss is an example of such a process (106). Polyploidy, or WGD, has occurred multiple times in the history of eukaryotes (21; 22) and is implicated in immediately precipitating massive gene loss (25), a major contributor to divergence in descendant polyploid genomes (35; 36). Multiple rounds of WGD can confound the synteny that is descended from each WGD event. Asymmetric gene loss following polyploidy has been shown to obscure synteny between genomes, making its detection difficult (25). Chromosomal rearrangements like transposition, translocation, segmental inversion and tandem duplication of genes also cause decay in synteny (39; 107).

Models of how these rearrangement processes affect gene content evolution could aid us in estimating their relative contribution to discordance in synteny between related genomic segments. Accurate models of these processes could unscramble obscured synteny. This in turn can help in predicting the presence of genes in regions of completely mapped genomes that are syntenic to regions in incompletely mapped genomes; particularly those descended from WGD events. Such models would also make it possible for us to assess which process has a bigger impact on synteny evolution over the others. We approach the task of building a probabilistic model of two particular rearrangements in this article: gene loss and gene gain, or the transposition of a gene into a different part of the genome. To investigate resolving rearrangements in the face of WGD, two kinds of loss are considered: loss of those genes created immediately following WGD and a background rate of gene loss.

There are many probabilistic models of gene loss and gain that model gene content amongst related species (108; 109). To model synteny evolution amongst related genomes, we consider models in a phylogenetic context. A variety of models have been proposed for gene gain or loss in a phylogenetic context (62; 110; 64; 60; 63; 111). Some

of these methods model the evolution of presence/absence of genes on a phylogeny in a phyletic nature (110; 62), while others use a continuous-time Markov process (63). Some model gene duplication on a local basis (62) and others consider the expansion and contraction of sizes of gene families (111). These models do not incorporate gene content evolution under WGD events. The models proposed here differ from these in simultaneously accounting for the two different gene loss processes (due to WGD and background) and the gene gain process described above. I have addressed three questions in this chapter. First, whether these models can be used in predicting gene content in unsequenced genomes. Second, how the amount of data in terms of number of genes and genome segments affects our being able to do so. Third, whether and when the distinction between these rates is possible.

## 2.3 The Models

Gene loss and gain were modeled as stochastic processes (112) using a Hidden Markov Model (HMM) (59). The genes were assumed to evolve i.i.d with constant instantaneous rates of loss and gain. When loss due to WGD and otherwise is not resolved, a background rate of $\alpha_R$ was assumed. Else, I distinguished between the rate of gene loss precipitated by WGD $\alpha_D$ and the background rate of loss $\alpha_S$. Genes were "gained" onto a segment by being transposed there from elsewhere in the genome at a rate of $\beta$. Loss and gain of an individual gene was assumed independent of the other genes present on the segment. Figure 2.1 illustrates the processes being modeled and the input to the model.

The input syntenic segments may be derived from one or more genomes and were all assumed to be descended from the same ancestral segment. The completeness of each segment represented the extent of our knowledge about their gene content, or an estimate of the fraction of genes present on it, that have been sequenced and are known.

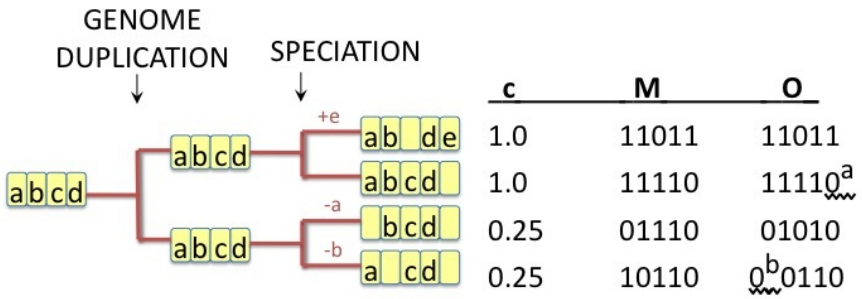


Figure 2.1: An example phylogeny of four segments, with a depiction of their evolutionary history. Their ancestor genome segment with 4 genes a,b,c,d underwent a WGD event, following which there was a speciation event. Gene e was gained onto the phylogeny after the speciation event. The top two segments are completely sequenced, and so have $c = 1$, but the other two segments are estimated to be 25% sequenced, and so have $c = 0.25$. The presence/absence matrix $M$ of the segments is not known to us, but what we can observe is the observation matrix $O$. As $c = 1$ for the top two segments, the rows corresponding to them in both matrices are equal; however, this is not the case for the lower two segments. **a** is unobserved because absent, **b** is unobserved as it hasn't been sequenced.

For example, a segment that is in a genome which has a high-quality genetic map like the plant Arabidopsis had $c = 1$. The phylogeny of the segments was derived from the phylogeny of the species they belong to.

The aim was to determine the probability that a gene not observed in an incompletely characterized segment was truly absent from it or not. Formally, given $O$, I wanted to obtain the estimate $\hat{M}$ of the presence-absence matrix $M$, where $M_{ij} = 1$ if gene $i$ is present on segment $j$, or 0 if it is absent, when $j$ is an incomplete segment. $M_{ij}$ is position probability matrix.

To obtain $\hat{M}$, I defined a hidden Markov model in which the hidden states are the

presence or absence of genes at each node in the phylogeny. The emissions of these states are whether $g_i$ is observed on the segment $G_j$ ($O_{ij} = 1$) or absent ($O_{ij} = 0.$). A gene present in $G_j$ is observed with probability $c_j$. The transitions along any branch of the phylogeny are governed by the rates [$\alpha_D$, $\alpha_S$, $\beta$].

Table 2.1: Input parameters, Models, Ranges Tested

| Input Parameter | Description | Model | Range examined |
|---|---|---|---|
| $\alpha_D$ | Rate of loss following WGD | $MLG$ | [0,1] |
| $\alpha_S$ | Background rate of loss | $MLG$,$LG$,$LO$ | [0,1] |
| $\beta$ | Rate of gene gain | $LG$, $MLG$ | [0,1] |
| $T$ | Topology | $MLG$, $LG$, $LO$ | Symmetric, bifurcating |
| $b$ | Branch lengths | " | Depth of tree = 1 unit |
| $N_S$ | Number of syntenic segments | " | 8,32 |
| $N_G$ | Number of genes | " | 50,500 |
| $c_j$ | Completeness of segment $j$ | " | $\frac{1}{4}$ and $\frac{3}{4}$ of segments incomplete $c_j = 0.05$ |
| $D : S$ | Ratio of Duplication:Speciation nodes | " | 1:1 |

I was interested in assessing whether resolving gene loss due to WGD as well as gene transposition made for better predictions in comparison to when I could resolve gene loss and gene transposition, or gene loss without transposition. To investigate the relative contribution of these different loss and gain processes, I defined three models in increasing order of rate complexity: $LO$ (Loss Only) with only loss, $LG$ (Loss Gain) with loss and gain and $MLG$ (Multiple Loss Gain) with background loss and loss due to WGD, as well as gain.

The transitions between states for the models and their emissions are summarized in Figure 2.2, and described in detail in section 2.4.

Bayesian estimates of the loss/gain rate parameters $\hat{\alpha_D}$, $\hat{\alpha_S}$ and $\hat{\beta}$ were used to compute $\hat{M}$ which contains the posterior probability of presence of each gene on each genome segment.

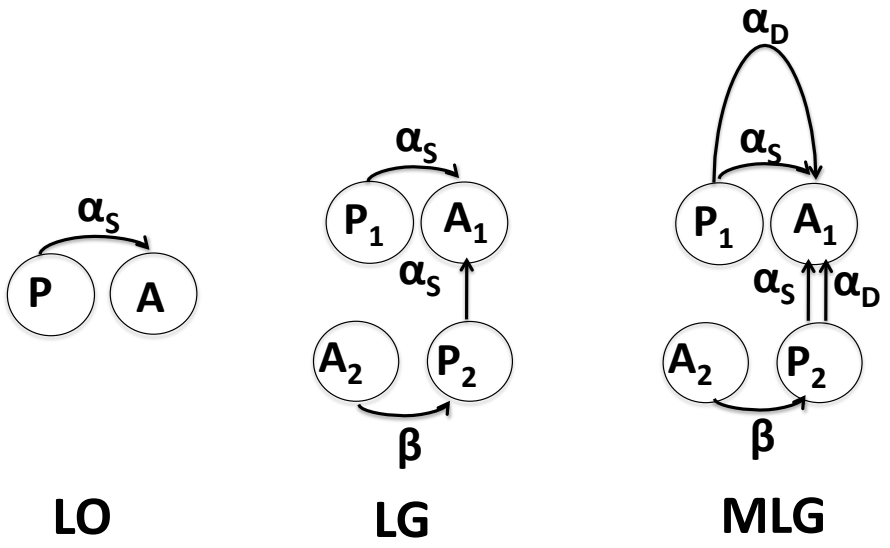


Figure 2.2: Building the HMMs for the 3 models: State Transitions for Loss Only (LO), Loss-Gain (LG) and Multiple Loss-Gain (MLG), and the Emission Probabilities for each set of states, determined by completeness of the segment $j$, $c_j$.

## 2.4 Methods

### 2.4.1 Input to the Models

The models required as input a phylogeny (topology $T$, branch lengths $b$) of the $N_S$ syntenic genome segments and a binary observation matrix $O$, that lists whether the $N_G$ genes are observed. or not on the segments (Figure 2.1). For example, if gene $g_i$ is observed present on segment $G_j$, then $O_{ij} = 1$, else 0. The internal nodes of the phylogeny are labelled as *Duplication (D)* or *Speciation (S)* nodes. The segments also

have associated with them a vector of *completeness* $c$, where $c_j \in [0,1]$, $j \in [1...NS]$. The completeness is the estimate of the the percentage of the genes that have been identified in the genome the segment belongs in. For example, a high-quality finished genome like Arabidopsis or *S. Cerevisiae* will have $c = 1$.

### 2.4.2 Simulation Studies

A symmetric, bifurcating tree topology was generated for $N_S$ segments and $N_G$ genes. The labeling on the nodes was assigned an equal number of Duplication, $D$ and Speciation, $S$ nodes, or with $D : S = 1$ with a Bernoulli process. Under the assumption that a gene is equally likely to have been present or absent at the ancestor, or prior probability of presence at the ancestor is 0.5, I modeled gain and loss as follows:
If a gene is present at the root node, for every internal node of the tree, gene loss is investigated with an exponential distribution to see whether its daughter nodes lost or retained genes, given the loss rate $\alpha_D$/ $\alpha_S$ specified by the node label. Once lost, the gene was not allowed to be gained again. If the gene was assigned absent at the root node and if the model allowed gain, a transition from absent to present was investigated on the branches. Once gained, the gene was then subject to the loss process described above and gain on subsequently sampled branches was disallowed.

Data was simulated for the values of parameters in the specified ranges for each $\alpha_D$, $\alpha_S$ and $\beta$. The rates of loss (and gain) are instantaneous rates, and the units are per gene per unit branch length. The depth of the trees are all unit branch length.

Summarized in Table 2.1 is a range of input parameters for which the model is tested.

### 2.4.3 Probabilistic Models of Gene Evolution

#### LO (Loss Only) model

Under this model, I only consider the effects of gene loss. A gene is subject to a loss at rate $\alpha_R$. An HMM is used to model the transition between the two states *presence* $P$ and *absence* $A$ of the gene on the internal nodes of the phylogenetic tree. Transition between the two states is governed by $\alpha_R$. At the leaves of the tree, hidden state $P$ emits 1 with probability $c_j$ and 0 with probability 1-$c_j$, corresponding to its state on segment $G_j$.

#### LG (Loss Gain) model

Under this model, I allowed gene gain due to transposition, in addition to gene loss due to speciation. A gene could be lost with rate $\alpha_R$, and can also be gained uniquely on the phylogeny (if absent previously) with rate $\beta$. I modeled 4 hidden states of the gene at the internal nodes - *presence* at the root node $P1$, *absence* under the root node $A1$. Similarly for the gene absent at the root node and gained subsequently, $P2$ and $A2$. Under the assumption that presence at the root was equally likely, hidden states $P1$ and $P2$ emit 1 with probability $0.5c_j$ and 0 with probability $0.5(1 - c_j)$, and hidden states $A1$ and $A2$ emit 0 with probability 0.5 each.

#### MLG (Multiple Loss Gain) model

Under this model, I differentiated between gene loss due to WGD and due to speciation, in addition to gene gain. On the branches following a $D$ node, $\alpha_D$ was modeled and on those following an $S$ node, $\alpha_S$ was modeled. Gain $\beta$ was modeled as before. As for the $LG$ model, I modeled 4 hidden states of presence/absence of the gene at the internal nodes, $P1$, $A1$, $P2$, $A2$. Loss was modeled due to either $\alpha_D$ or $\alpha_S$, depending on the label of the node being $D$ or $S$, respectively.

**Probabilities of transition between states**

I determined the probability that $g_i$ is present on $G_j$ for every instance a gene was 'unobserved', i.e $O_{ij} = 0$ as follows. I used Felsenstein's peeling algorithm (53) to compute the 'forward' and 'backward' probabilities of observing the gene content.

The posterior probability that node $n$ is in state $K$ given the observed data can be written as

$$p(\pi_n = K|T, O_i, \alpha_D, \alpha_S, \beta, c) = \frac{(D_K(n)U_K(n))}{P(O_i|T, \alpha_D, \alpha_S, \beta, c)} \tag{2.1}$$

where $D_K(n)$ and $U_K(n)$are the downward (forward) and upward (backward) probability, respectively, that node $n$ is in state $K$. Here, the observed data is specific to gene $g_i$, which is the $i$th column of $O_{ij}$, $O_i$.

The downward and upward probabilities for the leaves of the tree were initialized as follows: If $n$ is a leaf with $O_{ij} = 1$,

$$D_P(n) = 1 \tag{2.2}$$

and

$$D_A(n) = 0 \tag{2.3}$$

else if its observed value is 0

$$D_P(n) = c_j \tag{2.4}$$

and

$$D_A(n) = 1 - c_j \tag{2.5}$$

Since the transitions from branch to branch are modeled by a continuous Markov process

in time, the transition probability matrix $Pr(t)$ is the solution to

$$Pr^{'}(t) = \rho Pr(t) \tag{2.6}$$

where $t$ is the branch length and $\rho$ is the rate matrix.

Shown below are the possible state transitions for the $LO$ model:

$$\begin{pmatrix} PP & PA \\ AP & AA \end{pmatrix}$$

For the $LO$ model $\rho$ is

$$\begin{pmatrix} -\alpha_R & \alpha_R \\ 0 & 0 \end{pmatrix}$$

for which $P(t)$ is

$$\begin{pmatrix} e^{-\alpha_R t} & 1 - e^{-\alpha_R t} \\ 0 & 1 \end{pmatrix}$$

The $LG$ and $MLG$ model have the following different state transitions

$$\begin{pmatrix} P1P1 & P1A1 & P1P2 & P1A2 \\ A1P1 & A1A1 & A1P2 & A1A2 \\ P2P1 & P2A1 & P2P2 & P2A2 \\ A2P1 & A2A1 & A2P2 & A2A2 \end{pmatrix}$$

$\rho$ for the LG model is

$$\begin{pmatrix} -\alpha_R & \alpha_R & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \alpha_R & -\alpha_R & 0 \\ 0 & 0 & \beta & -\beta \end{pmatrix}$$

for which $Pr(t)$ is

$$\begin{pmatrix} e^{-\alpha_R t} & 1-e^{-\alpha_R t} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-e^{-\alpha_R t} & e^{-\alpha_R t} \\ 0 & h1 & h2 & e^{-\beta t} \end{pmatrix}$$

where

$$h1 = \frac{e^{-\alpha_R t}\beta - \alpha_R e^{-\beta t} + (\alpha_R - \beta)}{\alpha_R - \beta} \tag{2.7}$$

and

$$h2 = \beta \frac{\left(e^{-\beta t} - e^{-\alpha_R t}\right)}{\alpha_R - \beta} \tag{2.8}$$

$\rho$, $Pr(t)$ and equations 7 - 8 are the same as that for the $MLG$ model except that the loss rates are $\alpha_S$ or $\alpha_D$ depending on the branch label.

### 2.4.4 Computing the likelihood of observing $O$ given the rate parameters

The probability that the daughter node $n$ is in state $W$ given that its parent node $m$ is in state $K$ is calculated as

$$Pr(\pi_n = W/\pi_m = K|T, O_i, \alpha_D, \alpha_S, \beta, C) = \frac{(D_W(n)U_K(m))}{P(O_i|T, \alpha_D, \alpha_S, \beta, C)} \tag{2.9}$$

The log-likelihood to be maximized over all nodes $n$, over all genes $i$, over all possible

internal states $K$

$$L_{O_i,T,\alpha_D,\alpha_S,\beta,C} = \sum_i \sum_n \sum_k log(Pr(\pi_n = W/\pi_m = K|T, O_i, \alpha_D, \alpha_S, \beta, c)) \qquad (2.10)$$

Note: This log-likelihood is derived using the $MLG$ model. For the $LO$ and $LG$ models, the likelihood depends on $\alpha_R$ and ($\alpha_R$,$\beta$), respectively.

## 2.5 Parameter Estimation

### 2.5.1 Markov Chain Monte Carlo analysis of $\alpha_D$, $\alpha_S$ and $\beta$

$\hat{\alpha_R}$, $\hat{\alpha_D}$, $\hat{\alpha_S}$ and $\hat{\beta}$ are used to estimate the Presence/Absence matrix of the models which is the posterior probability of presence of each gene on each genome segment.

I used Bayesian inference to estimate $\hat{\alpha_D}$, $\hat{\alpha_S}$ and $\hat{\beta}$ and used the log-likelihood distribution as the target distribution for inference. I used a prior distribution on the rates to obtain an initial value for $\alpha_D$, $\alpha_S$, and $\beta$. This was done by doing a coarse grid search on the initial likelihood surface with initial proposed rates and adding a random perturbation to the maximum obtained to preclude starting off with a value that biases the algorithm to either stay near the maximum or stray too far from it. A standard normal distribution is used as the proposal distribution for the chain, centered around the current values of the rate parameters, and the Metropolis-Hastings ratio to compute the acceptance probability of the proposed move (as reviewed in (113)).

The proposals for each of the rate parameters were done in sequence, with the rates that are not under proposal at their current values.

A pilot sample was run to determine burn-in, number of iterations for the MCMC chain, and the values in the chain that are to be sampled to estimate the posterior cumulative distribution function of the q-quantile to within +/- The chain was run till convergence was determined. To determine convergence, the average standard deviation

of the split frequencies was measured for convergence to a steady value after burn-in. The posterior distribution was examined to make sure that the within-variance of the values of the chain and in-between variance of the runs were within expected values, with the Gelman-Rubin criterion (as reviewed in (113)). The rate estimates were then computed as the mean of the sampled posterior distributions.

### 2.5.2 Tests of Sensitivity and Specificity in Predicting Unobserved Genes.

With the simulated data, I estimated the accuracy with which the presence of unobserved genes are predicted.

Sets of presence-absence matrices $M$ are generated for specified $\alpha_D$,$\alpha_S$,$\beta$, for given $N_S$ and $N_G$. A specified percentage of segments (25 or 75 %) are randomly assigned to be incomplete, and assigned $c_j = 0.05$. Associated columns in the presence/absence matrices $M$ are masked to be 'unobserved' (i.e. 1 was changed to 0 in corresponding segment column for 95% of the values) and sent as input to the models, as a new input observation matrix $O$. With the rates estimated from these new matrices, the posterior probability of presence of the unobserved genes in $O$ are computed. Based on their corresponding values in $O$ and for specified probability cut-offs of presence of Pr(presence of unobserved gene) = [0.01,0.99], I calculated the true positive TP, false negative FN, true negative TN and false positive FP rates of presence.

The sensitivity is defined as

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.11}$$

and specificity is defined as

$$Specificity = \frac{TN}{TN + FP} \tag{2.12}$$

.

### 2.5.3 The Yeast Data Set

Genome structure evolution is more biologically complex than what is modeled in the simulations. I tested the $LO$, $LG$ and $MLG$ models on genomic data simulated to be incomplete, to measure the accuracy of predictions of presence of unobserved genes.

The gene content of 11 yeast species and the synteny observed between them as used in (1) was used to test the models. This data set was chosen in particular because five of these species showed evidence of a WGD event. The species used were *S. cerevisiae, V. polyspora, N. castelli, S. bayanus, C. glabrata, K. waltii, L. thermotolerans, L. kluyveri, E. gossypii, K. lactis* , and *Z. rouxii*.

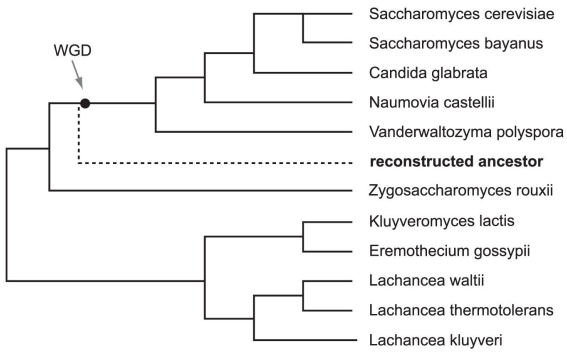


Figure 2.3: Phylogeny of the 11 yeast species showing the WGD event and position of the inferred ancestor. (Adapted from (1))

The authors reconstructed the pre-WGD ancestor that was dated to exist just before the WGD event, a 100 million years ago. This reconstruction contained the 8 inferred ancestral chromosomes, along with their gene content and order. From the 5 post-WGD yeast species, 2 inferred regions or 'tracks' each map to an ancestral chromosome reconstruction. An ordered list of the ancestral chromosomal gene content reconstruction, along with a list of the orthologs that correspond to these reconstructions in each of the 11 species (2 each for the 5 post-WGD species) was used here to test the model.

Presence/absence matrices were constructed as follows:

Each ancestral chromosome reconstruction and its associated orthologs were considered to form a 'multiplicon' of syntenic segments, with 16 (6 + 5x2) segments in each multiplicon. The segments from contemporary genomes contain presence/absence information for each of the $N_k$ reconstructed ancestral genes on the $k$ chromosomes (k $\in$ 1... 8) and also include 'singleton' genes from the genomic segments defined by the ortholog genes to the ancestral segment.

For each of the contemporary segments, if an ortholog to a gene in ancestral chromosome is present, 1 was entered into its position in the observation matrix, otherwise 0. If there were genes in between two recorded orthologs to ancestral genes that are not present in the ancestral chromosome, the observation matrix was re-sized to accommodate entries for them if they were within the threshold of what constitutes a neighbourhood of genes that display segmental homology.

To determine whether the run of genes in between two orthologs to ancestral genes belong in a syntenic segment, it must be determined how many singleton genes can be found in between two orthologs in a syntenic segment.

I used a parameter defined in (87) to do so. Here, a simple null model for homologies amongst genes in the absence of synteny was used to define what constitutes a neighbourhood of genes that significantly displays segmental homology. If $h_j$ is $W_j/(W_T W_j)$, $W_j$ is the number of orthologs in genome $G_j$ $j \in 1...11$ to the ancestral reconstruction, $W_T$ is the total number of ancestral genes (here, 4703) and $W_j$ is the total number of genes listed for genome $G_j$,

$\kappa$ defined as:

$$\kappa \leq 0.5 + [\frac{log(1-T)}{log(1-h)}] + 0.25 \tag{2.13}$$

is the threshold number of singleton genes that are determined significant in between two orthologous genes.

T is the probability cut-off at which a run of singletons was decided to be significant

or not.

All 16 constructed genomic segments (2 each from the 5 post-WGD species, 1 each from the 6 pre-WGD species) and the list of genes that constitute them defined the observation matrix. If 2 segments shared the same singleton gene, only one instance of that gene was recorded for the observation matrix, i.e the gene did not form 2 separate 'gain-like' columns.

### Phylogeny: Topology, Branch Lengths

The phylogeny of the 16 segments for each multiplicon was inferred from that used in (1), using their placement of the WGD event. The branch-lengths were obtained from (114).

### Completeness

I estimated the completeness of the 11 yeast genomes in this data set. *S.Cerevisiae* is the most completely sequenced annotated fungal genome to date. The definition of completeness here is an estimate of how many of the protein-coding genes in the contemporary genomes have been identified up to date. Even though the sequencing of more and more genomes sheds light on the presence of more genes than previously identified in *S. Cerevisiae* (115) I considered its genome completely or a 100% sequenced in this regard.

To obtain a relative estimate of how completely sequenced the other genomes are in relation to that of *S. Cerevisiae*, I considered those genes in the ancestral reconstruction for which *S. Cerevisiae* has orthologs. I then calculated the percentage of genes for which the other contemporary genomes have orthologs to this gene set and designated this percentage to be my estimate of how completely each yeast genome has been sequenced.

Table 2.2: Completeness estimates for the yeast data

| Organism | Estimate |
|---|---|
| *S. cerevisiae* | 1.0 |
| *V. polyspora* | 0.96 |
| *N. castelli* | 0.97 |
| *S. bayanus* | 0.91 |
| *C. glabrata* | 0.97 |
| *K. waltii* | 0.94 |
| *L. thermotolerans* | 0.96 |
| *L. kluyveri* | 0.96 |
| *E. gossypii* | 0.94 |
| *K. lactis* | 0.96 |
| *Z. rouxii* | 0.97 |

# Results

### 2.5.4 Simulation Tests

In order to determine how gene predictions and rate estimates were affected by the amount of data in the observation matrix, the following were varied in the simulations:

1. Number of genes $N_G$
2. Number of segments $N_S$
3. The fraction of incomplete segments

When incomplete, segments were incomplete at $c_j = 0.05$.

We can expect to find data sets with a minimum of 8 syntenic segments in a multiplicon from the angiosperms (eg. Arabidopsis with 2 suspected WGD events (116; 25; 26), rice with 1 suspected WGD event (16), etc), and in yeast (24); hence the lower value of 8 for $N_S$. An $N_G$ of 50 is similarly an estimated lower bound for the number of genes we can expect to see in such multiplicons (117). To test the limits of the model and its predictive abilities, I assigned a very low level of completeness to a genomic

segment when incomplete at 0.05, or that only 5% of the total genes in the genome have been assigned to physical map locations. This is the estimate to which a variety of agronomically important crop plants have been sequenced. To assess prediction for mostly complete to very incomplete data, I considered data sets with 25% - 75% of the genomic segments incomplete. In the models, the deepest branch in the phylogeny is of unit branch length.

I assessed the accuracy with which $[\hat{\alpha_D},\hat{\alpha_S},\hat{\beta}]$ and $\hat{M}$ could be estimated using simulated data. The data was simulated using the $MLG$ model as described in section 2.4.

Figure 2.4 below shows $[\hat{\alpha_D},\hat{\alpha_S},\hat{\beta}]$ estimated for $\alpha_D$ ranging from 0.1 to 1, $\alpha_S = 0.25$,$\beta = 0.25$, $NG = 50$, $NS = 32$ and an equal number of $D$ and $S$ nodes for complete data.

With 100% complete data, $[\alpha_D,\alpha_S]$ were accurately estimated in the range of $[0.1, 1]$ for the values of $N_S$ and $N_G$ specified; hence I performed the simulations with the parameters in this range. Low rates of gain were recovered accurately. I found that $\beta$ was consistently underestimated for the values simulated in the range $[0, 0.2]$, even though I tested it over a range of $[0.1, 1]$ (not shown in the figure 2.4). This was not unexpected, as the model does not disallow gain occurring more than once per lineage.

I then compared the models on different parameter regimes by observing trends in the rate estimates and analyzing the Receiver-Operator Characteristic (ROC) curves for the gene content predictions. The base parameter values for these simulations were $N_S = 8$, $N_G = 50$, $c_j = 0.05$ for 25% of the segments.

**Varying the number of genes, $N_G$**

To assess the variation in prediction as a function of $N_G$, I tested the model on sets of simulated data for which $N_G$ was either low at 50 or ten times higher at 500. In Figure 2.5, I have shown the fit of the $LO$, $LG$ and $MLG$ models for $N_S$=8 and 32. I found

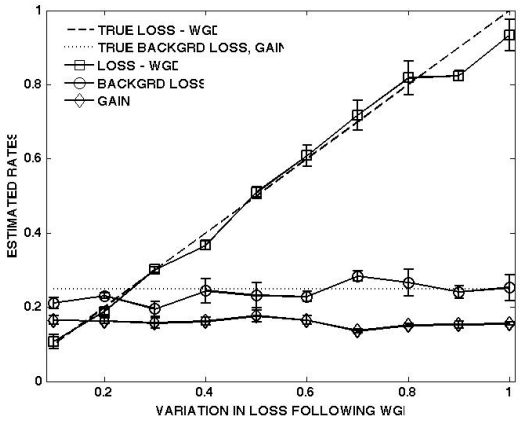


Figure 2.4: $\hat{\alpha_D}$,$\hat{\alpha_S}$,$\hat{\beta}$ plotted for when $\alpha_D$ is varied from [0.1,1], $\alpha_S = 0.1$, $\beta = 0.2$. $N_G = 50$, $N_S = 32$, all the segments are complete, i.e $c_j = 1$ for all of them, in a and b.

that there was no noticeable difference in the gene predictions using 50 or 500 genes for 8 segments.

Table 2.3: Rate Estimates in varying $N_G$

True and Estimated Rates

| Model | NG | $\alpha_R$ | $\alpha_D$ | $\alpha_S$ | $\beta$ |
|---|---|---|---|---|---|
| | **True** | | 0.4 | 0.1 | 0.2 |
| LO | 50 | 0.41 | - | - | - |
| LO | 500 | 0.42 | - | - | - |
| LG | 50 | 0.21 | - | - | 0.56 |
| LG | 500 | 0.17 | - | - | 0.47 |
| MLG | 50 | - | 0.27 | 0.12 | 0.56 |
| MLG | 500 | - | 0.25 | 0.07 | 0.48 |

At 50 genes, a sensitivity of 0.8 and higher is attained only for values of specificity of 0.6 and lower. With 500 genes however a high sensitivity (0.9 and higher) is attained at specificity values of 0.8 and lower. For values of specificity 0.9 and higher, the $MLG$ model had a 50% higher sensitivity than the $LG$ or $LO$ model. For all models, however, the rate estimates were inaccurate. $\beta$ was over-estimated to be more than twice as much of its real value for both the models. $\alpha_R$ was estimated to be twice as much as $\alpha_S$ with

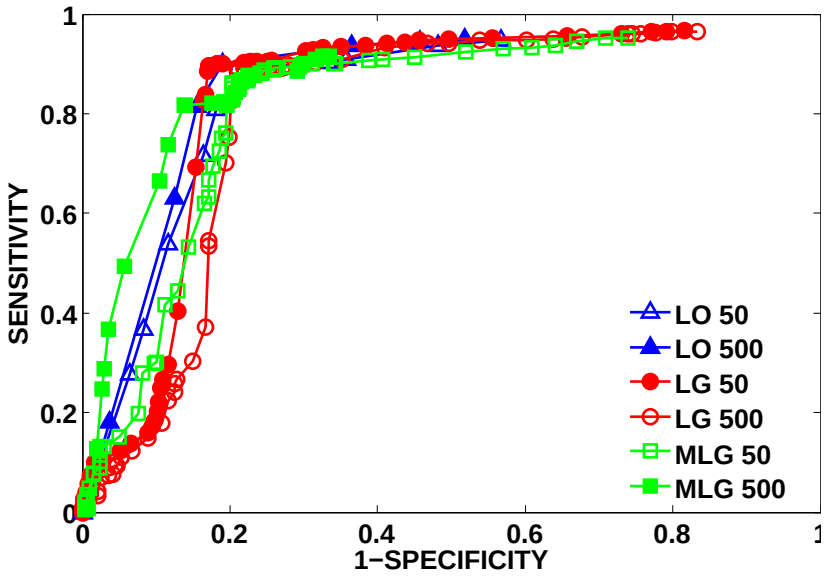


Figure 2.5: 8 segments, 25% incomplete data for the $MLG$ (squares), $LG$ (circles) and $LO$ (triangles) models, with 50 (open) and 500 genes (filled).

the LG model and four times as much with the $LO$ model. For the $MLG$ model, $\alpha_D$ was under-estimated by as much as 50% of the true simulated value of 0.4, while $\alpha_S$ estimated to within 20% of its true value of 0.1.

### Varying the number of segments, $N_S$

To examine the effect of number of segments on gene predictions, I considered simulations with 8 and 4 times as much segments, with $N_G$ fixed at 50, 25% of the segments

incomplete. In Figure 2.6, I have shown the fit of the $LO$, $LG$ and $MLG$ models for $N_S$=8 and 32. I found that there was a sharp increase in accuracy of gene predictions with more segments in the data set.

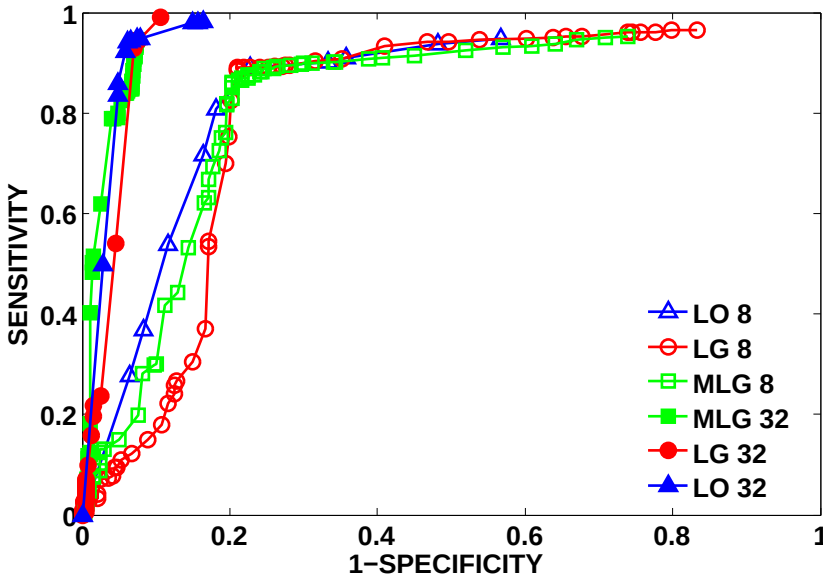


Figure 2.6: ROCs for $MLG$ (squares), $LG$ (circles), and $LO$ (triangles) for 8 (open) vs 32 (filled) segments.

With 32 segments, a high level of sensitivity (0.9 and higher) was achieved for a specificity as high as 0.9. All predictions were made at very high values of specificity ranging from [0.8,1]. Again, there seemed to be no clear difference in using one model over the other, for either 8 or 32 segments. The $MLG$ model had the highest sensitivity

Table 2.4: Rate Estimates in varying $N_S$

| True and Estimated Rates | | | | | |
|---|---|---|---|---|---|
| Model | NS | $\alpha_R$ | $\alpha_D$ | $\alpha_S$ | $\beta$ |
| | **True** | | 0.4 | 0.1 | 0.2 |
| LO | 8 | 0.41 | - | - | - |
| LO | 32 | 0.47 | - | - | - |
| LG | 8 | 0.21 | - | - | 0.59 |
| LG | 32 | 0.24 | - | - | 0.17 |
| MLG | 8 | - | 0.27 | 0.12 | 0.56 |
| MLG | 32 | - | 0.38 | 0.07 | 0.17 |

among the three models for very high specificity [0.9,1]. In this case much more accurate estimates of the true rate parameters were obtained with the $MLG$ model. Interestingly $\beta$ was more accurately estimated when $NS = 32$ with the $LG$ and $MLG$ models.

**Varying completeness of the segments, $c_j$**

I considered data sets where most (25%) of the segments were complete and where most (75%) were incomplete.

In Figure 2.7, I have shown the fit of the $LG$ and $MLG$ models for $c_j = 0.05$ for either 25% or 75% of the segments. This range represents the range of completeness found in genomic data, like that of many flowering plants. $N_S = 32$. I found that though not as stark as for variation in $N_S$, there is an increase in accuracy of gene predictions the more complete segments there are in the data set.

When 25% of the segments are incomplete, the highest sensitivity obtained was 0.9 for a specificity of 0.8. For high values of specificity $\in$ [0.8,1] the predictions made when 25% of the data is incomplete was 50% higher that those made when 75% of the segments are incomplete. As for the previous two cases, there was no clear trend among the models for each case, though it was interesting to note that the $LG$ model had a steeper ROC curve than the other two models. Also interestingly enough, there was no

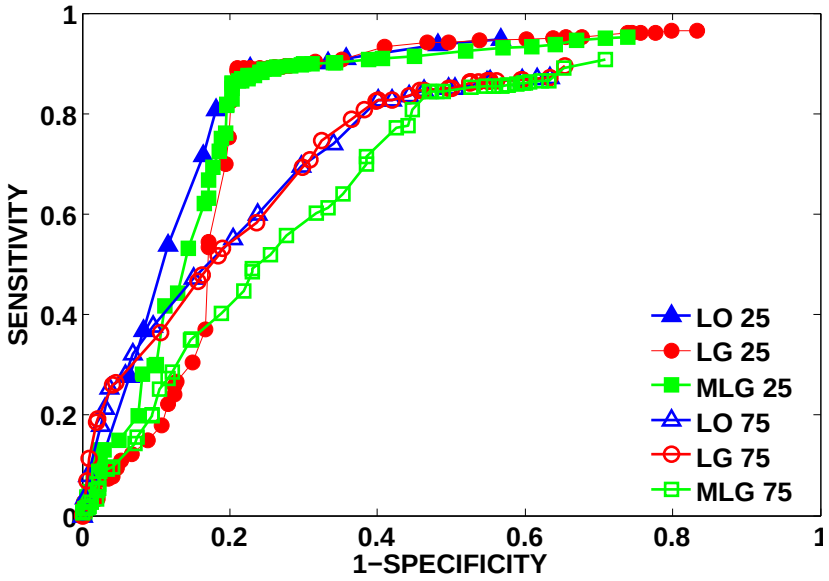


Figure 2.7: ROCs for $MLG$ (squares), $LG$ (circles), and $LO$ (triangles) for segments that are 75% (open) vs 25% (filled) incomplete

appreciable difference in the estimates of $\alpha_R$, $\alpha_D$ and $\alpha_S$ for the three models, while there was a 50% difference in the $\beta$ estimates.

### 2.5.5 Genomic Data

#### Tests on yeast data

I tested the models on yeast data to assess how the predictions of unobserved genes varied when genomic data was simulated to be incomplete.

Table 2.5: Rate Estimates for varying $c_j$

| Model | % Incomplete | $\alpha_R$ | $\alpha_D$ | $\alpha_S$ | $\beta$ |
|---|---|---|---|---|---|
| | **True** | | 0.4 | 0.1 | 0.2 |
| LO | 25 | 0.47 | - | - | - |
| LO | 75 | 0.46 | - | - | - |
| LG | 25 | 0.24 | - | - | 0.17 |
| LG | 75 | 0.22 | - | - | 0.28 |
| MLG | 25 | - | 0.38 | 0.07 | 0.17 |
| MLG | 75 | - | 0.39 | 0.065 | 0.27 |

True and Estimated Rates

The 8 reconstructed ancestral chromosomes described in (1) led to the design of 8 multiplicons (as described in 2.4) with gene lists ranging from 579 - 1198 in number.

Table 2.6: Sizes of the Multiplicons

| # | Ancestral | Inferred |
|---|---|---|
| 1 | 536 | 814 |
| 2 | 670 | 906 |
| 3 | 581 | 881 |
| 4 | 389 | 579 |
| 5 | 719 | 984 |
| 6 | 381 | 577 |
| 7 | 548 | 738 |
| 8 | 879 | 1198 |

The $MLG$ and $LG$ model were applied to each of the multiplicons and the rates inferred are summarized in Table 2.7. The value of $\alpha_D$ inferred is $\sim 4$ times the value of $\alpha_S$, and $\sim 11.4$ times the value of $\beta$ with the $MLG$ model. The rates of $\alpha_S$ and $\beta$ estimated from the $LG$ model are very similar to that estimated by the $MLG$ model.

The authors of (1) inferred 124 gene gains in S. Cerevisae, which corresponds to a rate of 0.0378 per gene per unit time (124/0.5859/5601) for comparison to $\beta$ inferred for the $MLG$ and $LG$ models. They also inferred the ancestral gene set to have 4703 genes in total. The most complete genome *S. Cerevisiae* has 5158 orthologs to this ancestral

Table 2.7: Estimated rates from yeast data

| Model | $\alpha_R$ | $\alpha_D$ | $\alpha_S$ | $\beta$ |
|---|---|---|---|---|
| LO | $0.803 \pm 0.0375$ | - | - | - |
| LG | $0.53 \pm 0.029$ | - | - | $0.172 \pm 0.0042$ |
| MLG | - | $1.93 \pm 0.689$ | $0.49 \pm 0.042$ | $0.168 \pm 0.0049$ |

gene set, with which we inferred that 4248 genes (4703 x 2 - 5158) were lost in total. Hence, a measure of $\alpha_S$ for the $LG$ model is $\sim 0.771$ which is 1.5 times that estimated by the $MLG$ and $LG$ models.

**Incompleteness and gene prediction**

Incomplete observation matrices were generated from the complete observation matrices constructed from the 8 multiplicons. I applied the $LO$, $LG$ and $MLG$ models to estimate rates and predict the presence of unobserved genes.

Figure 2.8 summarizes the trends observed in the rates estimated. For the $MLG$ model, while $\alpha_D$ and $\beta$ were both over-estimated at roughly greater than 5 and 4 times as more of the segments in the data set were incomplete, the opposite trend was observed for $\alpha_S$. With the $LG$ model estimates, both $\alpha_S$ and $\beta$ are over-estimated by $\sim 1.5$ and 3 times greater with highly incomplete data.

Shown in Figure 2.9 is an example ROC curve of the predictions obtained. These curves are predictions of simulated incompleteness of the multiplicon defined by the ancestral chromosome 6 reconstruction.

The probability cut-off used to generate this figure was in the range of [0.1,0.9]. With 25% of the segments incomplete at 5% incompleteness, a sensitivity of 0.8 and higher was obtained for values of specificity of 0.8 and lower, for both the $LG$ and $MLG$ models, similar to what was observed for the simulations involving $N_S = 32$ segments and $N_G = 50$ genes with 25% of the segments incomplete. These values of sensitivity and specificity began to degrade for a probability cut-off of $\sim 0.4$. When 75% of the segments

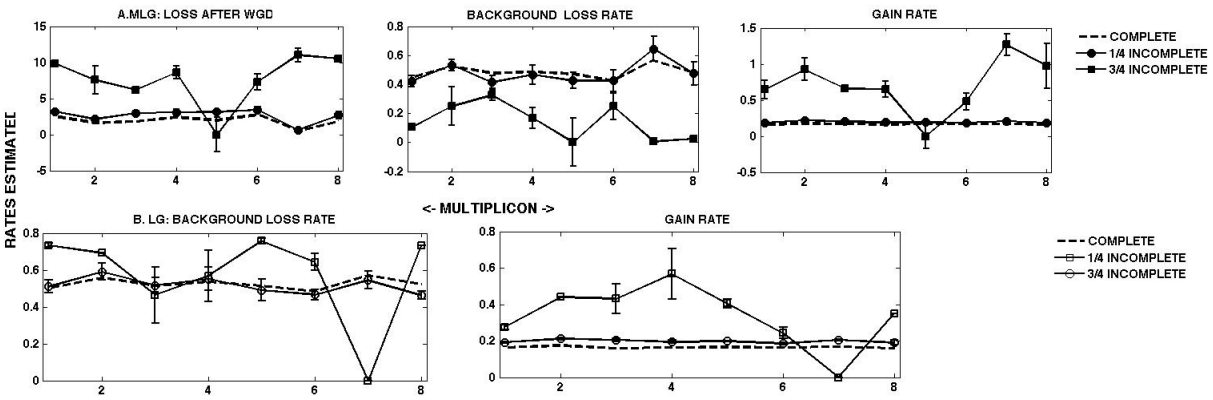


Figure 2.8: Estimated rates for all multiplicons: a. $\alpha_D$ b. $\alpha_S$ and c. $\beta$ with all segments complete(blue),1/4 incomplete at 5% (circles) and 1/4 incomplete at 5% (squares). The three panels of the first row order are the $MLG$ estimates of $\alpha_D$, $\alpha_S$ and $\beta$, respectively. The second row shows the $LG$ estimates of $\alpha_S$ and $\beta$.

were incomplete, this sensitivity dropped to 0.6 for the same range of specificity. The $LG$ model made slightly stronger predictions than the $MLG$ model in this instance.

### Simulations based on Yeast data multiplicon sizes

While there can be no direct measure of $\alpha_D$ from the yeast data set, an approximate measure of $\alpha_S$ and $\beta$ was inferred from the loss and gain events in the lineage leading to *S. Cerevisiae*. Using the same phylogeny obtained for the $NS = 16$ yeast genome segments and node labels, simulations were performed using the values of $N_G$ corresponding to the 8 multiplicon sizes to compare the estimates on simulations the size of the yeast data. The simulations were performed with the $LG$ model, with $\alpha_S^Y = 0.77$, and $\beta^Y = 0.038$ (the superscript 'Y' to denote that these parameters were inferred from the yeast data set). I estimated the rates with the $MLG$ and $LG$ models.

The variance in $\hat{\alpha_D}$ was much higher than that of the other two parameters. $\alpha_S$ was under-estimated and $\beta$ was over-estimated by 4-5 times as much over $\beta^Y$ used in the simulations. The values of $\hat{\alpha_D}$ observed in the individual iterations ranged from

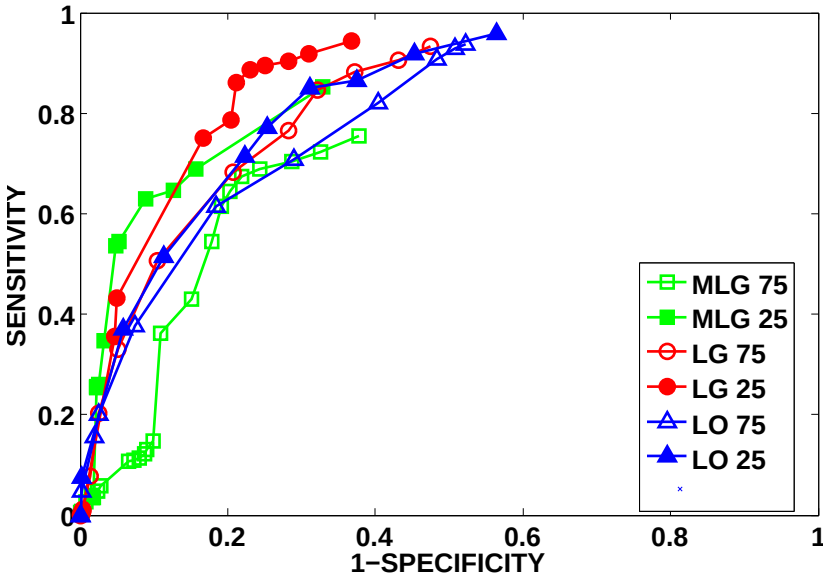


Figure 2.9: ROC curves as estimated for multiplicon 6, with curves for the 1/4 (25%) incomplete data in filled, 3/4(75%) in open shapes. Pr(An unobserved gene is present) = [0.1,0.9].

Table 2.8: Estimated rates from simulated data with yeast multiplicon sizes

| Model | $\alpha_R$ | $\alpha_D$ | $\alpha_S$ | $\beta$ |
|---|---|---|---|---|
| LG | 0.67 $\pm$0.039 | - | - | 0.19 $\pm$ 0.004 |
| MLG | - | 1.93 $\pm$ 0.689 | 0.49 $\pm$ 0.042 | 0.17 $\pm$ 0.005 |

[0.47,1.19].

## 2.6 Discussion

I have demonstrated that with this simple probabilistic model of gene loss and gain it is possible to predict gene content in syntenic regions of incompletely sequenced genomes with reasonable accuracy. I find that a non-trivial amount of data is needed. Of the factors examined, including the number of genes, segments and completeness of the data set, the largest increase in accuracy of prediction came with an increase in the number of segments. Differentiating between $\alpha_S$ and $\alpha_D$ did not make a profound difference in predictions when sensitivity was high, though the $MLG$ model consistently had higher values of sensitivity than the $LG$ model for high values of specificity.

From the simulations, I have demonstrated that parameters for background loss and loss following WGD and transposition can be differentiated, even for the lower limits of $N_G = 50$ and $N_S = 8$ tested. $\alpha_D$ was set to be 4 times as high as $\alpha_S$. For the $MLG$ model, I found that $\alpha_D$ is consistently under-estimated, particularly in the presence of highly incomplete data. The limit of accurately estimating $\alpha_D$ and $\alpha_S$ is in the number of segments in the data set, as the estimates improved from $N_S = 8$ to 32 with a corresponding decrease in error of estimate from 37.5% to 2.5% of the simulated value. I also found that $\beta$ is best estimated for values in the range [0,0.2]. As there is no distinction between loss due to speciation and duplication, for the $LO$ model, I expected $\alpha_R$ to be estimated at a value higher than $\alpha_D$ and $\alpha_S$ and at a value intermediate them

with the $LG$ model. I found that with the $LO$ model, $\alpha_R$ was estimated at values 2..5 - 17% higher than $\alpha_D$, while it was estimated at a value between $\alpha_D$ and $\alpha_S$ for the $LG$ model. I observed that $\alpha_D$, $\alpha_S$ and $\beta$ are best resolved at the larger limit of the data tested in the simulations ($N_S = 32$, $N_G = 500$, 25% segments complete).

With our current understanding of how various rearrangement processes impact genome structure evolution (34), I also wanted to assess the accuracy of the predictions made on biological data with WGD events. The size of the data set of the 11 yeast species studied in (1) was comparable to the ranges of $N_S$ and $N_G$ that I tested in my simulations. The numbers of genes in the multiplicons inferred from the reconstructions were much larger: 579 - 1198 as compared to the 50-500 that I tested. The 16 segments in the yeast multiplicons were at a value intermediate in the range of 8 and 32 that we tested.

The predictions made by both $LG$ and $MLG$ models were as good - and even better - than expected from the simulation tests. Unlike in the simulations, where there were as many duplication nodes as speciation nodes, there was only one duplication node and 14 speciation nodes in the yeast data set, as there was only one WGD event. As observed in the simulations, differentiating between $\alpha_D$ and $\alpha_S$ did not influence the predictions. It helped to account for elevated loss under WGD, in all of the multiplicons, except for one of them. While $\hat{\alpha_D}$ from 7 out of the 8 multiplicons were estimated within one standard deviation around the mean, $\hat{\alpha_D}$ from multiplicon 7 was estimated at two standard deviations away from the mean and was also very close to $\hat{\alpha_S} = 0.57$, from which an elevation in loss immediately following WGD cannot be inferred in regions syntenic to ancestral chromosome 7. The unusual pattern of retention in this multiplicon could arise from strongly linked functional properties of the genes descended from this chromosome (48; 118).

To examine whether the rate estimates were in a reasonable range, I simulated data sets for the sizes of the multiplicons inferred from the reconstructions under the $LG$

model (as I could not readily infer the number of genes lost immediately following WGD). $\hat{\alpha_S}$ from these simulations was close in value to that used for the simulations ($\alpha_S^Y$). $\hat{\beta}$ was not, but it was close to the estimates obtained from the yeast multiplicons themselves. The model is not sensitive to values of $\beta$ that are an order of magnitude smaller than the loss rates. The authors in (1) parsimoniously reconstructed the ancestor of the 11 yeast species considered here. What I estimate here with my models are instantaneous rates of loss and gain per gene per unit branch length (in these simulations $\sim$ 170 million years). With the $MLG$ model, I estimate that $\alpha_D$ was $\sim$ 5 times higher than $\alpha_S$ following WGD. The authors of (1) estimated a total of 4248 gene losses and 127 gene gain events in the lineage leading to *S. Cerevisiae* using their ancestral reconstruction. Using the rates of $\alpha_R$ and $\beta$ for the $LG$ model, I estimate a smaller $\sim$ 2700 loss and much higher $\sim$ 948 gain events. With the $MLG$ model, I estimate 2733 loss events and 925 gain events. For the $LO$ model, I estimate 4425 loss events, which is much closer in magnitude to that estimated by (1) than for the other models. The $LO$ model that does not account for gene gain at all has the closest estimates of loss events to that inferred by (1). As the Bayesian rate estimates account for all possible events of loss and gain along the different branches of the phylogeny, they are expected to be different from the parsimonious estimates described in (1). This suggests that the gene content and order observed among the 11 yeast species considered here could have been generated by a lower instantaneous rate of loss and much higher rate of gene transposition, not detected by parsimony. The authors only included genes in their ancestral reconstruction for which they were able to resolve its location at the time of the WGD event in the yeast lineage. This excluded gene content in subtelomeric regions. The excluded gene content could contribute to the difference in the events estimated. Some of the genes in the ancestral gene set considered here might have been absent at the ancestor and transposed into a set of the lineages at some internal node that descended from the ancestor and therefore incorrectly inferred to have been present at

the ancestor. I also noted that the sizes of the multiplicons generated by the $MLG$ model with the ancestral set of 4703 genes on 8 chromosomes inferred by (1) were $\sim$ 1.5 times larger than those observed in the data set. This difference in size can arise both through different in rate estimates as well as a difference in prior probability of presence of the genes in the ancestor, which I consider here to be 0.5. This suggests that a different ancestral gene content size and different prior probabilities of presence also factor into the difference seen in the numbers of estimated gain and loss events. In terms of the relative frequencies of the rate estimates, for the $MLG$ model, $\alpha_D$ was $\sim$ 10 times of $\beta$ which is higher than the magnitude of 2 times that we simulated (0.4/0.2) and $\sim$ 5 times that of $\alpha_D$, comparable to what I simulated (4 times - 0.4/0.1). For the $LG$ model, $\alpha_R$ was $\sim$ 4 times higher that of $\beta$. This suggests that irreversible gene transposition could be occurring at much higher frequencies than normally suspected (39).

The accuracy of these predictions are contingent on assumptions of an error-free phylogeny of the participant genome segments and estimates of how far they have been sequenced, which are clearly not realistic (53). The rates are assumed to be homogenous along the branches and there is evidence for this not being the case (115). Therefore my estimates represent the average effect of different episodes of loss following WGD. Rates of loss are also known to be different for different classes of gene families (48). The current framework of the model allows for extensions to fit and test variation in the rates used, to investigate if better differentiation between rates is possible and/or make more accurate gene content predictions. One way of doing so would be implementing different functions for the rates (57) and specifying branch- and gene-specific distributions of rates. These extensions to the models may provide insight into the evolution of syntenic gene content after WGD.

# Chapter 3

# Reconstruction of Ancestral Gene Content and Order of Syntenic Genomic Segments

## 3.1 Abstract

Gene order and content between closely related species diverge through chromosomal rearrangements like gene loss, gene duplication, inversions, transpositions and translocations. Extensive research has been done on computing the distance between genome segments to their ancestor when they share equal gene content under rearrangements that preserve gene content like inversions, translocations and transpositions. Such distances are not expected to adequately model the divergence between species that experience episodes of polyploidy and the increased gene loss that follows after. *eAssembler* is a heuristic algorithm that reconstructs ancestral gene order and content for genomic segments of unequal gene content. In this chapter, I evaluate the accuracy with which *eAssembler* reconstructs ancestral gene order and content for genomic segments simulated under a model of evolutionary rearrangements that include polyploidy, gene loss, dispersed gene duplication, inversions, translocations and transpositions. I use the breakpoint, inversion and DCJ distances within *eAssembler* to measure the merits of one distance over the other for a variety of rearrangement regimes. I also propose values for the input parameters to *eAssembler* to guide statistically significant clusters

of segments for reconstructions. I find that the accuracy of reconstruction is affected by the distance measure used in *eAssembler* and the evolutionary regime simulated.

## 3.2 Introduction

Decoding the evolution from ancestral genomes to current-day genomes presents many challenges. Fossils have been used in uncovering the hidden steps in evolution. Fossils are not available for all taxa and different methods have been developed to infer ancestral character states on ancestral nodes in the eukaryotic phylogeny. Some of these methods, particularly the earlier ones, used maximum parsimony to infer ancestral character states, with both heuristic and probabilistic models (119; 120). Maximum likelihood methods have also been developed (121) to reconstruct ancestral states.

In the last two decades, the comparison of sequences of whole genomes enabled the inference of their most recent common ancestor or MRCA in ways that were not available before. A lot of the first comparisons were performed with mitochondrial, plastid and prokaryotic genomes which are an order of 5 - 10 times smaller than nuclear genomes and have simpler structures. Genomes were represented by their constituent markers with a beads-on-a-string model. These markers were usually genes. More comparisons are done with the gene content of the genomes, rather than their entire nucleotide sequences (122).

Some of the challenges in using the nuclear genomes of species for comparison are the difficulties in sequencing them and from unequal copies of genes across species for comparison. We have complete genome sequences for only 38 eukaryotic genomes today and there is a dearth of phylogenetically informative species genomes that remain to be sequenced fully. Many of these are commercially and scientifically important plant genomes like wheat, barley, sugarcane, etc. (9).

Genomes of many species have partially conserved *synteny*, or conserved gene con-

tent and gene order. Synteny is sometimes highly preserved and easily visible between close relatives (1). It can be very hard to detect in species that have experienced a lot of disruptive evolutionary rearrangements and subsequent divergence, as is the case in distantly related flowering plants (21). Synteny that is conserved in spite of rearrangements can be detected by a variety of methods, experimental and computational. Synteny is leveraged in many applications in comparative mapping and a very useful example is determining the genetic maps of intractable genomes using those of fully sequenced model organism genomes (21). In flowering plants, multiple instances of polyploidy coupled with massive gene loss produce unequal gene content across and within the species genomes. At least three suspected rounds of polyploidy in the Arabidopsis lineage has resulted in many duplicated chromosomal segments (123; 26). Small scale duplications can also produce synteny. Distinguishing between long preserved regions of ancestral synteny and duplications on a smaller scale is made difficult by fragmentation of gene order. These ancestral patterns of synteny are often not readily identifiable by *ad hoc* methods.

If we were able to observe the common ancestor of two contemporary genomes, each of these genomes would clearly be more similar to their ancestor than they are to each other. In fact, the gene order and content of the extant genomes can potentially be more easily resolved in comparison with the MRCA ancestral gene content and order. Particularly in organisms that have undergone multiple rounds of polyploidy, synteny between regions that correspond to older duplications are harder to detect than between regions derived from younger duplications. If we assume that the genes observed in a set of regions did exist in the ancestor that pre-dated the duplication events, then an accurate reconstruction of the ancestor could help connect the hidden synteny between the regions derived from different duplication ages. Blanc et. al used an ancestral reconstruction in guiding their search for synteny in the Arabidopsis genome, and discovered 68 more pairs of syntenic segments in the genome in this fashion (124). In a more re-

cent example, Gordon et. al manually assembled the MRCA of 11 yeast species, five of which had undergone a WGD. With this reconstruction, they both validated existing evolutionary hypotheses and tested some new ones. In this case, however, they had the advantage of having the complete whole genome sequences of the species they used in their study. An accurate reconstruction of the ancestor of related genomes could uncover much more of the synteny than is currently observed.

There are several methods that have been developed for the reconstruction of ancestral gene order and content for genomes that have equal gene content. Equal gene content automatically precludes gene loss and duplications - two major disrupters of synteny. Due to this preclusion alone, there is a known inaccuracy in these reconstructions. Reconstructions are used most often to estimate rearrangement rates in the evolutionary path leading to contemporary genomes, as well as to estimate the correct phylogeny of related species. When the gene content being compared is equal, the gene orders are considered as permutations of each other and of the ancestral genome. The first method that was developed for reconstructions was the reversal or inversion distance method by Hahhenhalli and Pevzner (70). The distance between two contemporary genomes was computed as the number of reversals it would take to transform one genome to the other. To obtain an ancestral reconstruction, this method was extended to compute the ancestor for which the distance between the ancestral reconstruction and the extant genomes was minimized (71; 83).

Another heuristic for computing the distance between genomes is to count the number of rearrangements between the two, or the number of *breakpoints*. The breakpoint distance was first proposed by Sankoff and colleagues (72) and applied to the task of reconstructing the mammalian ancestral genome. The inversion distance models rearrangements as inversions, while the breakpoint distance does not discriminate between them. A third distance measure that has been widely used is the double cut and join distance method (77) proposed by Yancopoulos et al. This method invokes the various

rearrangements that produce two cuts and a subsequent join in gene order, such as transpositions, fusions, fissions, inversions and translocations. It was been widely used and extended (125; 126).

Many extensions of these distance methods have been proposed to deal with unequal gene content between genomes. This would mean the inclusion of insertion, deletion (loss) and duplication of individual genes, or whole stretches of genes. Two approaches to deal with unequal strings have been proposed: the block-edit model, and match-and-prune model (49). The match-and-prune model transforms strings into permutations and then minimizes the distance between them or maximizes the similarity between them. The block-edit model is based on counting the number of operations needed to transform one string into another. The DCJ with deletions, insertions, etc. would be an example of the match-and-prune model.

There are a few algorithms that have been developed for reconstructing ancestors for genomes of unequal content, like *eAssembler* (2), *DUPCAR* (80) and Genome-Halving (127). The program *eAssembler* reconstructs ancestors of syntenic genome segments (identified by a synteny-finding program like *FISH* (87)), by reconstructing the breakpoint median or ancestor (72) for segments that are clustered together if they share a minimum of $\Upsilon$ genes in common and are at a breakpoint distance of no more than $\tau$ genes away from the reconstruction. The program *DUPCAR* looks at gene family phylogeny reconciliations to infer a segment/species phylogeny for the input genomic segments/genomes, and reconstructs the ancestor at every node of the reconciled phylogeny. In the Genome-Halving context, under the assumption that one species in the input data is a polyploid descendant of the common ancestor and the other is a non-polyploid descendant, the non-polyploid genome is used as a 'guide' to reconstruct the ancestor of the two species. *DUPCAR* reconciles orthologous gene families into ancestral regions on a gene family basis and does not implicitly model the context of syntenic regions. As Genome-Halving does not incorporate gene content not present in both the

polyploid and non-polyploid daughter genomes used to reconstruct the pre-polyploid ancestor, the reconstruction is incomplete with respect to the synteny of the ancestral gene order.

Reconstructions of ancestors of contemporary genomes could be more accurate if the evolutionary history within their *entire* gene content is used. This could be especially true in identifying synteny among genomes that are descendants of multiple polyploidy events. *eAssembler* is a computationally fast method that provides reconstructions for clustered segments with unequal gene content. It takes advantage of the overlap among syntenic genomic segments to assemble ancestral segments, that contain more distinct genes than any single segment in the cluster. An illustration of how the algorithm assembles reconstructions is shown in in Figure 3.1.

*eAssembler* has some limitations. The breakpoint distance with which the reconstructions or *medians* are currently computed is known to be non-discriminatory to evolutionary rearrangements that create breakpoints in synteny. Moreover, the parameters of $\Upsilon$ and $\tau$ that are used for clustering segments are currently arbitrarily decided. Distance measures like the inversion and DCJ measures (70; 128) are known to be superior distance measures for reconstructions, in some specific evolutionary contexts (71; 125).

In this chapter, I test two proposed improvements to ancestral reconstruction in the *eAssembler* algorithm. The first improvement is to evaluate how the choice of distance measure in *eAssembler* affects the accuracy of reconstructions. The second improvement is to use genome-derived values for the clustering parameters $\tau$and $\Upsilon$, to identify segments that are significantly syntenic over those that might be seen by chance in the genome. I use simulated data to test whether the use of one or both of these alterations actually improve the accuracy of reconstructions, as unlike for biological genomic data, we have the luxury of knowing the exact simulated evolutionary history of simulated genomic segments.

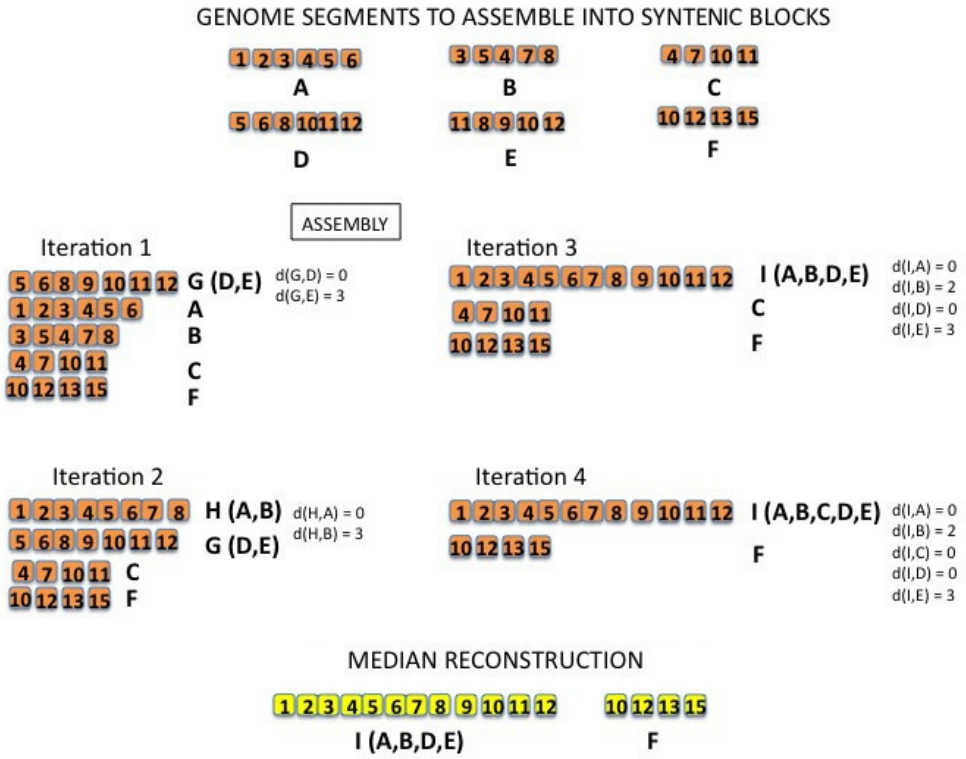


Figure 3.1: An illustration of the eAssembler algorithm, adapted from (2). Six segments are shown at top, each with four to six genes. Genes shared among segments are labeled with identical numbers. The bottom half of the figure shows four iterations of the agglomerative clustering process, with the corresponding medians in each step, and the breakpoint distance of each assembled segment to the median. In this example, the medians satisfy the assembly parameters of at least three shared genes and a maximum breakpoint distance of three from the median ($\tau = 3$, $\Upsilon = 3$).

Table 3.1: Parameters and Definitions

| Input Parameter | Description |
|---|---|
| $\tau$ | Minimum number of genes to be shared between clusters |
| $\Upsilon$ | Maximum allowed distance between cluster segments and median |
| B | Set of input segments to *eAssembler* |

## 3.3 Methods

### 3.3.1 Methods

The original eAssembler algorithm as described in (2) is implemented here in MATLAB, with some modifications. Briefly, the objective of the program is to reconstruct medians for all the segments assembled into clusters. For each cluster, the segments in the cluster share at least $\tau$ genes in common and are at a distance of no more than $\Upsilon$ from the reconstructed ancestor of the cluster. The input to eAssembler is a list of genomic segments or ordered list of genes that are identified to be syntenic by programs like FISH (87) and *i-ADHoRe* (86) and clustering parameters $\tau$ and $\Upsilon$.

A list of all the parameters used in this chapter is defined in table 3.1.

The Sankoff median (72) is used in *eAssembler*. In summary, a gene $g \in G = \bigcup g_i$, $i \in 1...n$, where $G$ is the union of all the genes in the $n$ segments in cluster $C$. At each iteration of the computation, the gene $\hat{g} \in \hat{G}$ that minimizes the cost function $\psi(M)$ is inserted, where $\hat{G}$ is $G \setminus M$, the set of all genes in G that are not in median $M$. These iterations are continued until $\hat{G}$ is empty or all the markers have been inserted into $M$. If more than one choice of $g \in \hat{G}$ satisfies the optimization criteria, one of the choices is randomly picked.

The cost function $\psi(M)$ that is minimized is

$$\psi(M) = \sum_{i=1}^{i=n} d(S_i, M) \tag{3.1}$$

Input : Segments $S_i$, $i \in 1...N$, $\tau$,$\Upsilon$
Output : $M$ clusters and reconstructions
Initialize: Clusters $C_i = S_i$ $i \in$ [1,N]
J = $Join(\{C\})$
**while** $J \neq \emptyset$ **do**
  $\{ C_a, C_b \}$ = max($Join\{C\}$)
  $M_{ab}$ = Median($C_a C_b$)
  **if** d(s,M) $\leq \Upsilon \; \forall$ s $\in C_a \cup C_b$ **then**
    clustered = 1
    Merge $C_a$ and $C_b$
  **end if**J = $Join(\{C\})$
**end while**
**if** clustered =1 **then**
  **for** $C_i \in 1...M$ **do**
    R = $C_i \setminus$ M
    **if** $R \neq \emptyset$ **then**
      Insert all g $\in R$ in $M$
    **end if**
  **end for**
**end if**

where $S_i$ are the $n$ segments in cluster $C$, and d is the distance between the median reconstruction $M$ and $S_i$.

If there are multiple candidate clusters that can be joined during the cluster joining process, one of the candidates is randomly chosen for joining. During the cluster joining process, a reduced median is computed. The reduced median is reconstructed with only those genes that are shared by at least a pair of segments in the cluster, i.e. genes $g \in \cap_{i=1}^{i=n} s_i \; \forall \; s_i \in C$. This reduction is made possible since breakpoints can only be inferred in shared content from one sequence relative to the other. It also reduces the time to compute the median as was demonstrated in the original implementation of *eAssembler* (2), as the number of genes that are to be inserted is reduced. The remaining genes are inserted into the median after the cluster joining process.

One important feature of *eAssembler* is that the medians are larger than the segments they are built from and can be used to cluster additional segments that did not have sufficient overlap of genes to be joined by themselves.

When there are instances of duplicated genes within a segment, the gene (and its position in the segment) selected for insertion into the median is chosen randomly from the set of duplicated genes in the segment.

The original implementation of *eAssembler* used only the breakpoint distance. The two alternative distance measures that are incorporated into *eAssembler* in this chapter are the Double-Cut-and-Join or $DCJ$ distance, and the inversion distance. The breakpoint distance between two segments of genes is the count of the number of adjacent genes within one segment that are not adjacent in the other, or the number of breakpoints between the two. The inversion distance is the minimum number of inversions of sub-segments of different length within one genome segment required to transform it into the other. The DCJ distance accounts for the number of ways two cuts of breaks in the sequence of genes in one segment can be joined by translocations, fissions, fusions, transpositions and segment interchange to transform into the other segment. For the DCJ distance, translocation and segment interchange have a weight of 2 units, whereas the other operations have unit weight. I have implemented the breakpoint distance algorithm in MATLAB for this chapter Both the DCJ and inversion distance measures have been adapted from the program GRAPPA (83). Both the inversion and DCJ distances have been adapted for use in *eAssembler* by the Tang lab in the University of South Carolina, from the *GRAPPA* program suite (83).

Rather than set $\tau$ and $\Upsilon$ arbitrarily, the properties of the dataset can be used to decide parameter thresholds to reduce false positives for a given null model of the distribution of homologies among genes in the input genome data. Based on the work of (129), optimal values of $\tau$ can be computed for two segments of a given length, for the given genomes and their associated gene families.

Equations 3.2, 3.3 and 3.4 define the probability of a seeing a gene cluster in a genomic region of size $r$ genes under the hypothesis of random gene order using the number of shared gene families $m$ as a statistic. An assumption is made that all gene

families in the genome are of the same size $\phi$ and further that the average length of the genomes is $n$. Using these assumptions, the authors showed that using $\phi$=2 for genomes of size $\leq$ 25000 genes fit power-law based gene cluster probabilities (129).

$\phi$ is set to be 2. I then calculate the probability of seeing $m$ homologous matches in a pair of genomic segments of size $r$ as $q(m)$. Here, $r$ is the average size of genomic segment from the input data.

$$q(m) = \sum_{k=m}^{r} \left[ \binom{n}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2(l|k) \right] \tag{3.2}$$

$$p_1(k) = \left( \binom{n}{r} \right)^{-1} (-1)^k \sum_{u=\frac{r}{\phi}}^{k} \left[ (-1)^u \binom{k}{u} \binom{u.\phi}{r} \right] \tag{3.3}$$

$$p_2(l|k) = \left( \binom{n-r}{r} \right)^{-1} \sum_{z} (-1)^l \sum_{u=\frac{r-z}{\phi}}^{l} \left[ (-1)^u \binom{l}{u} \binom{u.\phi}{r-z} \right] \binom{n-k.\phi}{z} \tag{3.4}$$

Hence, for a given significance threshold $\alpha$ and segment size $r$, $m$ can be selected such that $q(m) \leq \alpha$. This value of $m$ will then be suggested as the optimal value for $\tau$ for the assembly process.

For the breakpoint and inversion distances, it is possible to calculate the expected distance between two random permutations. From (130), I estimate that under the hypothesis of random gene order, for genomes of length $n$, the expectation of a breakpoint distance $d$ can be derived from

$$n - d = O\left( \frac{log(n)}{2} \right) \tag{3.5}$$

and from (131), I estimate that the expectation of the reversal distance $d$ between two random permutation of length $n$ (same hypothesis)

$$n + 1 - \frac{1}{2}log(n) - \frac{3}{2} + O\left(\frac{1}{n}\right) \leq E[d] \leq n + 1 - \frac{n+1}{2n}log(n+1) + O\left(\frac{1}{n}\right) \tag{3.6}$$

Therefore, once an optimal value of $\tau$ is calculated, by setting $n = \tau$, the expected distance that would be observed under a random distribution of matches amongst the genomes (and within them) and their segments. This defines a lower threshold for what value of $\Upsilon$ should be used in the assembly process. For all the experiments described in this chapter, I used the expected inversion distance as a proxy for the expected $DCJ$ distance.

In addition to the median computation method described above, an alternative method for computing the median is used here. The method optimizes the many ways of determining the sequence of inversions to transform one permutation into another (71). I modified the framework of *eAssembler* for this comparison as follows. At initial clustering steps, if a pair of segments are grouped into a cluster, the cluster is replaced by the computed median of shared gene content. At subsequent clustering steps, each cluster therefore consists of a single median segment that represents all median reconstructions upto that step. The coverage of genes in this median is therefore only as high as the gene content shared by all the segments that have participated in the clustering steps leading to the final median reconstruction. Hereafter, I will refer to this alternative median as the optimal inversion median.

### 3.3.2 Genome Data Simulator

To test the reconstructions of *eAssembler* , I designed a forward genome evolution simulator in MATLAB for data that models rearrangements scenarios similar to those inferred from biological data.

Simulations are initiated with a unichromosomal genome containing $A_G$ genes. A segment phylogeny is then simulated under given speciation and polyploidy rates. The depth of the phylogeny is set to be of unit length, which corresponds to approximately 150 million years, or the root of the angiosperm phylogeny. Inversions, translocations, dispersed gene duplications and gene losses are then simulated as stochastic processes

Table 3.2: Parameter Rates used for testing the three distance measures in *eAssembler*

| Parameter | Description | Dimension | Range | Default |
|---|---|---|---|---|
| $A_G$ | Ancestral genome size | Number of genes | 50,500 | 50 |
| $\lambda_s$ | Speciation | per unit time | [0.5,1.5] | 1.2 |
| $\lambda_p$ | Polyploidy | per unit time | [0.5,1] | 0.5 |
| $\lambda_i$ | Inversion | per unit time | [0.5,2] | 0.5 |
| $\lambda_t$ | Translocation | per unit time | [0.5,2] | 0.5 |
| $\lambda_d$ | Dispersed Duplication | per gene per unit time | [0.5,2] | 0.5 |
| $\lambda_l$ | Gene Loss | per gene per unit time | [0.5,2] | 0.5 |

occurring along the branches of the phylogenetic tree. A dispersed gene duplication is the duplication of a single gene and transposition of its duplicate elsewhere in the genome. Whereas translocations and inversions are modeled as processes that apply to the entire genome per unit time, the dispersed duplication and gene loss parameters are modeled as processes per unit gene per unit time.

These processes are formally defined as follows.

For G, a list of N genes $g_1, g_2, ...g_N$:

An **inversion** between the $i$ and $j$th genes where 1≤ i ≤ j ≤ results in $g_1..g_j g_{j-1}..g_{i+1} g_i..g_N$.

A **loss** of the $i$th gene where 1≤ i ≤ n results in $g_1..g_{i-1} gi + 1..g_N$.

A **translocation** of length $l$ starting at the $ith$ gene to a location starting at the $k$th gene where 1 ≤ i ≤ N and 1≤ k ≤ N results in $g_1..g_{k-1} g_k g_i g_{i+1}..g_L g_{k+1}..g_N$.

**Dispersed duplication** of gene $i$ to a location next to the $k$th gene results in $g_1..g_i...g_k g_i g_{k+1}...g_N$.

The dispersed gene duplications serve to create a 'cloud' of homology within the data that can lead to false positive homologies relative to homologies generated by polyploidy events. I am particularly interested in regimes of the data that have properties similar to that observed in angiosperms. To test reconstructions for this chapter, I took a reduced set of those parameter regimes (which are detailed in chapter 4). They are summarized in Table 3.2, along with their default values in simulations.

For both inversions and translocations, a single chromosome is selected for rearrangement. A inversion with a length that has a lower bound of $\frac{1}{5}$ to an upper bound of $\frac{1}{2}$ the total length of the chromosome is generated.

In this framework, each of the processes are assumed to operate independent of each other. The events are simulated as Poisson processes.

A phylogeny was simulated with a Yule process using speciation rate $\lambda_S$ (132). If the number of nodes generated was $v$, the number of WGD nodes to be assigned was calculated as $\lambda_p$v. Candidates for the WGD labels were selected uniformly randomly from the set of $v$ available nodes. The root node had $A_G$ genes at time $T$=0. The simulation was initiated at the root node and continued along each branch of the phylogeny. If the parent node of a branch had a WGD label, the genome was duplicated before any of the other processes are simulated. A Gillespie process (133) was used to stochastically simulate rearrangements. The simulation ends at time $T = 1$.

A typical simulation result for the default parameter regimes used here resulted in a phylogeny of 4 genomes, with at least two of the genomes having experienced WGD events. At least one genome per simulation underwent 2 rounds of polyploidy on average. As a result, the resultant genomes had anywhere from 1 - 4 chromosomes, resulting in a total number of 7 - 9 chromosomal segments in the data set.

For this chapter, my objective was to simulate genomic segments that are all derived from one ancestral segment of size $A_G$ with the processes listed above and measure the quality of reconstructions obtained under the regimes tested.

### 3.3.3 Measures of Performance

The quality of the reconstructions was assessed by three metrics: **Coverage**, **Normalized Induced distances** and **Quality of Reconstruction**. Coverage measured what proportion of the genes seen in contemporary genomes were present in the reconstruction and was calculated as the ratio of the distinct genes of all reconstructed segments

to the number of genes in the original genome. For a segment, the Normalized Induced Distance (e.g. breakpoint distance) is the distance between the reconstruction and the true gene order, and is defined as the ratio of its distance to its length in genes. Hence, a lower NI distance indicates a more accurate construction, as does a higher value of coverage. I compute the NI distance using all three distance measures; NB with breakpoint, ND with DCJ and NI with inversion.

Quality of Reconstruction **QR** is defined as the ratio of coverage to NI distance. The higher the value of this ratio, the higher the quality of the reconstruction.

The coverage and the three normalized induced distance measures for *eAssembler* were compared for data sets of evolutionary regimes where one process is at a higher rate in comparison with the others (e.g.. frequency of inversion higher than that of gene loss, transposition and translocation, etc.) to examine the potential advantage of using one distance measure over the other. This was used to evaluate the relative performance of one distance measure over the other for a particular evolutionary regime.

Coverage and normalized induced distances were measured from each reconstruction to the starting ancestral segments in the simulations. For the tables in the results, QR for the breakpoint reconstruction was calculated as coverage/NB, for the inversion as coverage/NI and for the DCJ as coverage/ND. For the optimized inversion median, QR was calculated as coverage/NI.

## 3.4 Results

To test the reconstructions obtained from my modified version of *eAssembler* with each of the breakpoint, inversion and DCJ distances, I simulated data sets under a set of different parameter regimes. The chromosomal segments of the genomes at the end of the simulation were sent as input to *eAssembler*.

I simulated a regime *EQ* where all the parameters are at their default frequencies. I

also simulated a regime $HI$ where the frequency of inversion is set to its high value, i.e. $\lambda_i = 2$ with the other parameters at their default frequencies and a regime $HL$ where the frequency of loss is set to its high value, i.e. $\lambda_l = 2$, with the other parameters at their default frequencies. I used these three regimes to infer the relative contribution of inversions and gene loss over other rearrangements to accuracy in the reconstructions.

### 3.4.1 Clustering parameters $\tau$ and $\Upsilon$

I wanted to measure whether the proposals for clustering parameters that I have described in equations 3.2 - 3.4 yield reconstructions with optimal coverage and Normalized Induced Distance. To test this, I used simulations to measure coverage and ND obtained for a range of values of $\tau$ and $\Upsilon$ for the different regimes. The data sets were generated with a starting $A_G = 50$. The lengths of the chromosomal segments obtained varied from a minimum of 11 genes to a maximum of 30 genes and had an average length of 16 genes. The proposed value for $\tau$ and $\Upsilon$ is calculated for the average size of genomic segment. I therefore looked at the range of $\tau$ and $\Upsilon$ values calculated for ranges of segment length between the minimum and maximum value.

The values of $\tau$ derived were $m = \tau = 3.5$ and 15 for which the corresponding $\Upsilon$ values were 2, 4 and 13 for the breakpoint and inversion distances at a level of significance $\alpha$ of 0.0001. $\tau = 5$ and $\Upsilon = 4$ correspond to the average length of segment of 16 genes. I added in an additional intermediate value of $\tau = 10$ for which $\Upsilon$ =9.

The coverage and ND obtained with the breakpoint and inversion reconstructions for all combinations of $\tau$ and $\Upsilon$ used on data sets from the $EQ$ and $HI$ regime are shown in Figure 3.2. Panels A, B, E and F in the left of the figure correspond to the $EQ$ regime and panels C, D, G and H correspond to the $HI$ regime. Panels A, E, C and G were generated using the breakpoint reconstruction and panels B, D, F and H were generated using the inversion reconstruction. The first row of panels A, B, C and D show the coverage over the $\tau$-$\Upsilon$ grid while the second row below with E, F, G and H

show the corresponding ND. Five sets of simulation replicates were used to derive the results in this figure.

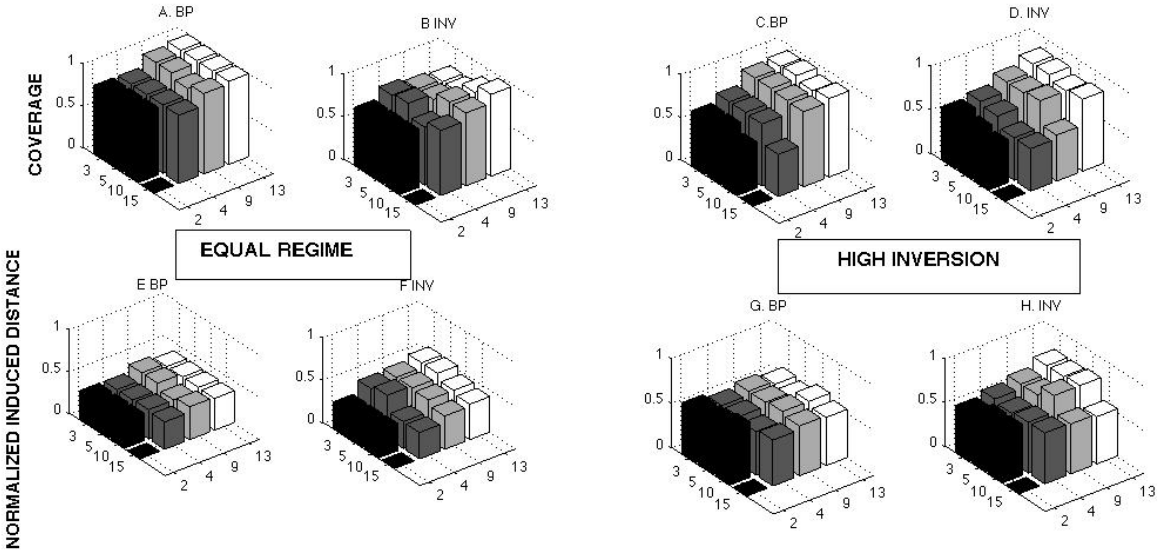


Figure 3.2: Values of coverage and ND for $\Upsilon$ = 2,4,9,13 and $\tau$=3,5,10,15 for two evolutionary regimes. Figures A,B,E and F correspond to the equal frequency regime, while C,D,G and H correspond to a high inversion regime.

The most optimal values are those that yield the highest coverage for the lowest ND. The proposed value of $\tau = 4$ has the lowest ND at $\Upsilon = 4$ for all the panels except F corresponding to the inversion reconstruction in the $EQ$ regime. The coverage obtained for this set of $\tau$ and $\Upsilon$ was within a 25% neighbourhood of the highest coverage observed. The trends for coverage and ND were similar for the breakpoint and inversion reconstructions for the $EQ$ regime. For the $HI$ regime, the breakpoint reconstruction yielded the lower ND.

As the proposed values for $\tau$ and $\Upsilon$ yield reconstructions with high coverage and low ND, I used this method in the rest of the experiments described in this chapter.

### 3.4.2 Measuring Reconstruction Quality

I tested the difference in accuracy of the ancestral reconstruction under different the evolutionary regimes $EQ$, $HI$ and $HL$ with different distance measures in the median computation of *eAssembler*. ($\tau$, $\Upsilon$) were calculated as described in the Methods 2.4 to be (5,4) for the breakpoint reconstruction and (5,5) for the inversion and DCJ reconstructions. Each of the measures are summarized from the results of 10 simulation sets.

Table 3.3 summarizes the different performance measures for the $EQ$, $HI$ and $HL$ regimes with $A_G = 50$ for all of them except for $HL$, where $A_G = 500$.

Table 3.3: Comparison of the different distance measures in *eAssembler* for **EQ**, **HI** and **HL** regimes

| **Measure** | **Reconstruction** | | |
|---|---|---|---|
| | Breakpoint | Inversion | DCJ |
| | **EQ** | | |
| **Coverage** | 0.64 ± 0.024 | 0.59 ± 0.037 | 0.65 ± 0.039 |
| **NB** | 0.50 ± 0.061 | 0.46 ± 0.068 | 0.50 ± 0.074 |
| **NI** | 0.58 ± 0.063 | 0.55 ± 0.042 | 0.59 ± 0.037 |
| **ND** | 0.55 ± 0.065 | 0.59 ± 0.042 | 0.55 ± 0.039 |
| **QR** | 1.28 | 1.07 | 1.18 |
| | **HI** | | |
| **Coverage** | 0.65 ± 0.025 | 0.62 ± 0.016 | 0.601 ± 0.02 |
| **NB** | 0.51 ± 0.046 | 0.49 ± 0.047 | 0.45 ± 0.044 |
| **NI** | 0.68 ± 0.045 | 0.59 ± 0.027 | 0.59 ± 0.018 |
| **ND** | 0.65 ± 0.025 | 0.56 ± 0.028 | 0.56 ± 0.017 |
| **QR** | 1.27 | 1.05 | 1.07 |
| | **HL** | | |
| **Coverage** | 0.17 ± 0.039 | 0.15 ± 0.033 | 0.17 ± 0.041 |
| **NB** | 0.46 ± 0.069 | 0.48 ± 0.07 | 0.42 ± 0.082 |
| **NI** | 0.61 ± 0.078 | 0.56 ± 0.051 | 0.51 ± 0.065 |
| **ND** | 0.58 ± 0.087 | 0.52 ± 0.062 | 0.47 ± 0.073 |
| **QR** | 0.29 | 0.27 | 0.36 |

When all rearrangements are simulated in equal frequency with each other in the $EQ$ regime, there is no expectation for which distance method is the better one to use for reconstructions. Here I found that the DCJ and breakpoint method have higher coverage than the inversion method. The inversion method has lower normalized induced distances than the other two methods except in the case of ND, where the DCJ has the lowest value. The breakpoint reconstruction had the highest $QR$ for this regime.

Rows 5 - 10 in Table 3.3 summarizes the performance measures for the high inversion $HI$ regime, with $A_G = 50$. The DCJ reconstruction produced the lowest normalized induced distances NB, NI and ND, though the mean NI and ND values the same for both the DCJ and inversion reconstructions. The breakpoint reconstruction yields the highest coverage and measure for $QR$.

Rows 11 - 15 in Table 3.3 summarize the performance measures for the $HL$ regime with $A_G = 500$. The mean coverage for all methods for this regime is $\sim 4$ times lower in magnitude than for the $EQ$ or $HI$ regimes. The DCJ and breakpoint reconstruction had the same mean coverages which are higher than the mean coverage for the inversion reconstruction. The DCJ reconstruction has the lowest values for normalized induced distances NI, NB and ND in all cases and has a quality of reconstruction about 25% times higher than that of the inversion and breakpoint reconstructions.

In order to test the difference in changing the way the median computation itself is performed, I compared the Sankoff median computation method (with the breakpoint, inversion and DCJ distances) with the optimized inversion median for the EQ, HI and HL regimes.

Shown below in 3.4 are the results for the comparison of the two median computation methods on the HI regime, with $A_G = 500$.

With the optimized inversion median, the coverage obtained was lower than that for the other reconstructions. This is not unexpected, as the inversion median is only computed for segments that have the same gene content. However, the normalized

Table 3.4: Comparing different median computations

| | **Breakpoint** | **Inversion** | **DCJ** | **Optimal Inversion** |
|---|---|---|---|---|
| | High Inversion | | | |
| **Coverage** | 0.65 ± 0.025 | 0.62 ± 0.016 | 0.601 ± 0.02 | 0.56 ± 0.044 |
| **NB** | 0.51 ± 0.046 | 0.49 ± 0.047 | 0.45 ± 0.044 | 0.42 ± 0.073 |
| **NI** | 0.68 ± 0.045 | 0.59 ± 0.027 | 0.59 ± 0.018 | 0.52 ± 0.059 |
| **ND** | 0.65 ± 0.025 | 0.56 ± 0.028 | 0.56 ± 0.017 | 0.48 ± 0.06 |
| **QR** | 1.27 | 1.05 | 1.07 | 1.07 |

induced distances were the lowest for this method, suggesting that the accuracy of the reconstruction is higher than that of the other methods.

## 3.5 Discussion

Through these simulation studies I have shown that for the evolutionary regimes that include genome structure rearrangements like WGD, gene loss, inversion and transposition, using different distance methods in *eAssembler* produces differences in the quality of reconstruction. The DCJ reconstruction had the lowest normalized induced distances from the ancestor for both the high loss $HL$ and high inversion $HI$ regimes. For the $EQ$ regime, the inversion reconstruction had the lower normalized induced DCJ and inversion distances. The breakpoint reconstructions had the highest mean coverages for all the regimes although the DCJ reconstruction mean coverage was $\sim 1\%$ higher for the $EQ$ regime and the same for the $HL$ regime. For regimes that have high gene loss, the DCJ distance method yields the most accurate reconstructions in *eAssembler*. For the other two regimes, different distance measures produced the more desirable measures of reconstruction.

I found that the proposed value for clustering parameters $\tau$ and $\Upsilon$ adapted from (129; 131; 130) yield higher quality reconstructions with high coverage for lower NI distances. The choice for $\tau$ adapted from (129) identifies the number of gene homologies that are shared between two genomic regions of the same size $r$ that are spatially significant.

This choice is better than an arbitrary choice for $\tau$.

Different reconstruction methods are optimal for different rearrangement regimes. For regimes that have predominantly inversion rearrangements, the breakpoint method provides the best reconstructions. This is relevant in systems like the cereals, where maize is inferred to have undergone many rearrangements since its divergence from rice, for example (46; 134). Polyploidy events are known to precipitate massive gene loss (35; 36; 25), particularly in angiosperm regimes. The DCJ method provided the best reconstruction among the three methods and can be used for reconstructing ancestral gene order and content in such lineages; for example, the angiosperm ancestor prior to the divergence between the monocots and eudicots. These lineages are known to have undergone several rounds of polyploidy (21), gene loss (25), inversions (134; 135; 45), translocations (45) and dispersed duplications (112). In a yeast pre-WGD ancestral reconstruction study (1), 73 inversions and 66 reciprocal translocations, 4248 gene loss and 124 gene gain events were inferred in the lineage leading to S. Cerevisiae from the inferred pre-WGD ancestor (1). Apart from the gain parameters, the inferred yeast genome rearrangement parameters are comparable to those used in the HL regime. Therefore, a DCJ distance method can be used to provide a reconstruction of the pre-WGD ancestor of the 11 yeast genomes used in the (1) study.

The yeast genomes have rates of rearrangements comparable to those of the angiosperm lineages The breakpoint distance has been previously shown to provide reconstructions that are inferior to those computed with the inversion or even the DCJ distance (71; 125). The data sets the authors tested the inversion and DCJ methods on in (71; 125) modeled inversions and transpositions but did not account for gene loss or duplication. Most other genome rearrangement simulations (80; 136; 71; 125) have a framework where either an evolutionary tree is present, with a fixed number of events to be simulated on each edge of the tree, or a regime where the objective is to maximize the number of rearrangements in order to use unique breakpoints, and/or to achieve a

reduced genome size (here also an evolutionary tree is present). The distance measures described in this chapter have been tested on simulated data that incorporates all of these rearrangements, as well as polyploidy and gene loss. The conclusion that the breakpoint distance is inferior in performance to the DCJ or inversion distance cannot be drawn by looking at the coverage of reconstructions for the three different methods in the $HI$ and $EQ$ regimes. The DCJ distance accommodates many of the rearrangements that were used for the simulations in this chapter and is considered more realistic. The fact that it was not always better in both coverage and normalized induced distances of the reconstructions than the breakpoint distance method is a little unexpected. This might not be as surprising since each of the rearrangements are given equal weight in the DCJ operations, which can be considered analogous to the fact that the breakpoint distance does not discriminate between which rearrangements cause the breakpoints in synteny that it accounts for. In the case of the $HI$ evolutionary regime with high gene loss rate, the breakpoint distance reconstructions are less accurate than those of the other two distances. This is significant as a high rate of gene loss has been inferred to have occurred in angiosperm genomic data.

Recent studies (137) have used measures like genomic distance (analogous to normalized induced distances), breakpoint re-use rate and dispersion of sets of alternate solutions (dispersion in degeneracy) to evaluate the genome-halving technique for ancestral reconstruction. In the introduction to this chapter, I mention the issue of using even those genes that do not have homologs in any other genomic region. In this context, a method like *eAssembler* can be expected to have a much higher coverage of the genes present in extant syntenic genomic segments in the ancestor. The authors in (137) argue that the non-inclusion of such 'singleton' genes do not deteriorate their reconstructions. However, the simulations from which the authors measured their reconstructions modeled inversions and translocations and did not include rearrangements like gene loss, which degrade gene content. From the simulations in this chapter, high

coverage does not necessarily always result in high normalized induced distances, implying that the inclusion of singleton genes does not necessarily deteriorate the quality of reconstructions.

There are many opportunities for future work. The current median computation method can definitely be improved in a variety of ways. The heuristic method of inserting the gene that minimizes the distance function at each iterative step in computing the median is a local optimization function, and need not necessarily be the global optimal solution. Moreover, there are a lot of degeneracies in the choice of gene to insert at every iteration. There might be a more principled way of addressing the degeneracies over the current method of randomly picking a gene from the degenerate set. Rather than picking just one, a few could be picked at every clustering step to fork different median computation processes. These could be interrogated for a few more iterations till one choice optimizes the distance function better than the rest. If there is no optimal choice over the other within a few iterations, one of the processes could be chosen to continue. This kind of choice is relevant to the computational cost of the algorithm; the number of decisions to be made increases with the number of genes to be included in the reconstruction.

A different median computation method like the inversion method tested in this chapter can also be more useful in the context of reconstruction. However, the inversion method would have to be modified to accommodate segments of unequal gene content, which is an ongoing research problem. There are many proposals for DCJ-based methods that can accommodate segments of unequal content (128), but none that have been published for use as yet.

Though I have used biologically derived rearrangement rates in my simulations, they still cannot account for the all of the rearrangements that constitute genomic data. The authors in (137) suggest that differences between evolutionary rates amongst different genomes, for example, could affect the reconstructions. *eAssembler* could be modified

in a future version to account for clustering parameters for genomic segments specific to the properties of the genomes they are derived from. A guide segment phylogeny could also contribute to the clustering process, in which the clustering parameters would be derived for the node of the phylogeny clustering is performed at. For instance, the deeper the node is in the phylogeny, the more stringent the clustering parameters are as the segments are expected to be less diverged from each other at these nodes.

A more effective proposal for deriving $\tau$ could be to derive it for every pair of genomic segments/medians at every clustering stage. That way it would be derived from the genomic properties of the segments, instead of using a single universal value averaging the properties of all segments input to *eAssembler*. This might significantly increase the computation time of the algorithm. However, it can potentially help in reducing the degeneracies in the choice of the medians at every clustering step, particularly as the parameters can be expected to be more stringent at internal clustering steps.

# Chapter 4

# Segmental Homology Identification using Ancestral Reconstruction of Gene Content and Order with Synteny detection programs

## 4.1 Abstract

Identification of syntenic regions between closely related genomes is important in studying their genome structure evolution. There are a variety of methods that identify synteny through pairwise genome comparisons and detect profiles of synteny amongst genomes through multiple pairwise comparisons. Genomes that have undergone polyploidy and subsequent rearrangements like gene loss experience extensive degradation in their synteny. Pairwise comparisons might not able to detect synteny in these genomes. These genomes are expected to share more synteny with their ancestor than they do with each other. In this chapter, I compare the differences in accuracy of synteny measured by using the reconstruction of syntenic genome segments that are detected pairwise computed by the program *eAssembler* and the multi-segmental synteny detected by the synteny-detection program *i-ADHoRe*. I evaluate these programs with simulated data

sets that model inferred angiosperm rates of polyploidy, gene loss, inversions, translocations and transpositions. I also apply this method to reconstructing the ancestor of the angiosperm Arabidopsis and using it in a synteny analysis with the angiosperm rice genome.

## 4.2 Introduction

Closely related species share similar gene order and content, or *synteny*, in their genomes. With comparative mapping (138; 139), we can identify genomic regions that are homologous within and between genomes. Syntenic segments are descended from a single common ancestor and the present day order of genes suggests the order that existed in the ancestral genes. Recognizing genomic regions that are syntenic amongst species has been very useful in uncovering candidate genes in incompletely characterized genomes (138). Synteny is rarely conserved perfectly between species, especially when they are highly divergent. A variety of evolutionary processes contribute to disruption in synteny, namely inversion, transposition, translocation (reciprocal and otherwise), gene loss and gene duplication (individual, segmental and whole genome). *Ad hoc* methods for identifying syntenic regions (18) in the face of these rearrangements, particularly Polyploidy or Whole Genome Duplication (WGD) and the massive gene loss that usually follows are challenging problem. Computational methods have been designed to enhance our ability to identify syntenic segments. Some methods (13; 140) align DNA sequences to detect homologous regions. Homology detection at the nucleotide level becomes difficult with large sequence divergence. Other methods compare genetic or physical maps of genomes and genome segments and provide the advantage of being able to detect homology even between very divergent genome sequences. *FISH*, *i-ADHoRe* and *CloseUp* are examples of such methods (86; 87; 89).

*FISH* (87) and *i-ADHoRe* (86) both use Gene Homology Matrices or GHMs (141)) in their synteny analysis. A GHM is an information matrix where the rows and columns correspond to the positions of genes in their genomic sequences. A cell in the matrix contains a non-zero value if the genes corresponding to the row and column positions are homologous to each other or not. *i-ADHoRe* clusters points of homology in the GHM by minimizing a distance function that returns lower distances for points that cluster diagonally. The program detects all possible pairwise segments of synteny that are identified to be statistically significant with an input maximum distance between two points in a cluster and a distance-defined threshold. It uses these pairwise segments as profiles with which to collect additional syntenic segments. *FISH* (87) utilizes a GHM with a different distance function and null distribution to identify statistically significant clusters. The orientation of the points in the clusters do not influence the scoring function for both these methods. *FISH* detects syntenic segment pairs and does not build multi-segmental homology profiles like *i-ADHoRe*. *CloseUp* (89) identifies synteny based on parameters of proximity between genes homologous to others and their density within clusters that are identified. Clusters identified are evaluated for significance with Monte Carlo tests.

All of these algorithms start searches for synteny through pairwise comparisons. In searching for synteny pairwise, the most preserved synteny is likely the kind that is easiest detected. Particularly in a genome that has experienced several rounds of WGD, high gene loss and other rearrangements can lead to the fractionation of segment synteny. It is therefore very useful to compare multiple related genomes simultaneously for synteny. This is especially relevant when at least one of the genomes is considered to closely reflect a pre-WGD ancestral state in comparison to other post-WGD genomes, as in the case of the Vitis genome in comparison with other sequenced angiosperms like Arabidopsis (142). The Arabidopsis genome is inferred to have undergone at least two rounds of WGDs (26; 25). The advantage that comes out of multiple genome

comparisons with Arabidopsis, Carica, Populus and Vitis is illustrated in (143). In (86), the authors used *i-ADHoRe* to detect syntenic blocks in an Arabidopsis-rice genome comparison. They detected 23.8% of the genome in duplicated blocks than the 20.9 % they had found previously. They found that they were able to detect a 38.3% percent of the Arabidopsis genome in higher levels of synteny that in pervious analyses (16) as well. They did not, however, detect more syntenic regions in rice with this approach than they had in a previous study (117) where they inferred syntenic region pairs within a genome by using their separate synteny to a region from another genome.

A collection of syntenic segments, or a *multiplicon*, is illustrated in Figure 4.1. A multiplicon defines a unit of segmental homology. It is a collection of two or more segments (or contiguous intervals) within an ordered set of genes, or features (not spanning concatenation junctions). Those features that are homologous to a feature on at least one other segment within a multiplicon are the *anchors* of that multiplicon. Anchors can connect two or more segments at a time in a multiplicon. Segments within multiplicons begin and end with anchors at either end. Their intervals are defined by the position of the anchors with the lowest and highest index within that segment. *i-ADHoRe* defines levels for multiplicons as the number of genomic segments that they contain. For example, a multiplicon with 2 genomic segments is a level 2 multiplicon.

From Figure 4.1, without being able to see the evolutionary history of the genomes A, B and C, it would be hard to infer their inherited synteny, given their very limited shared gene content and order. Identifying pairs of syntenic segments with more than two anchors also would recover the multiplicon. However the synteny is evident among the four segments collectively. The synteny between them is also very clear when compared with the ancestor of these genomes. Therefore, synteny might be more clearly identified by looking at more than just a pair of genomic segments where anchors are sparse after large scale gene loss due to polyploidy. We can see that the resulting segments have very few genes in common - in fact, the common intersection of the segment genes is 0.

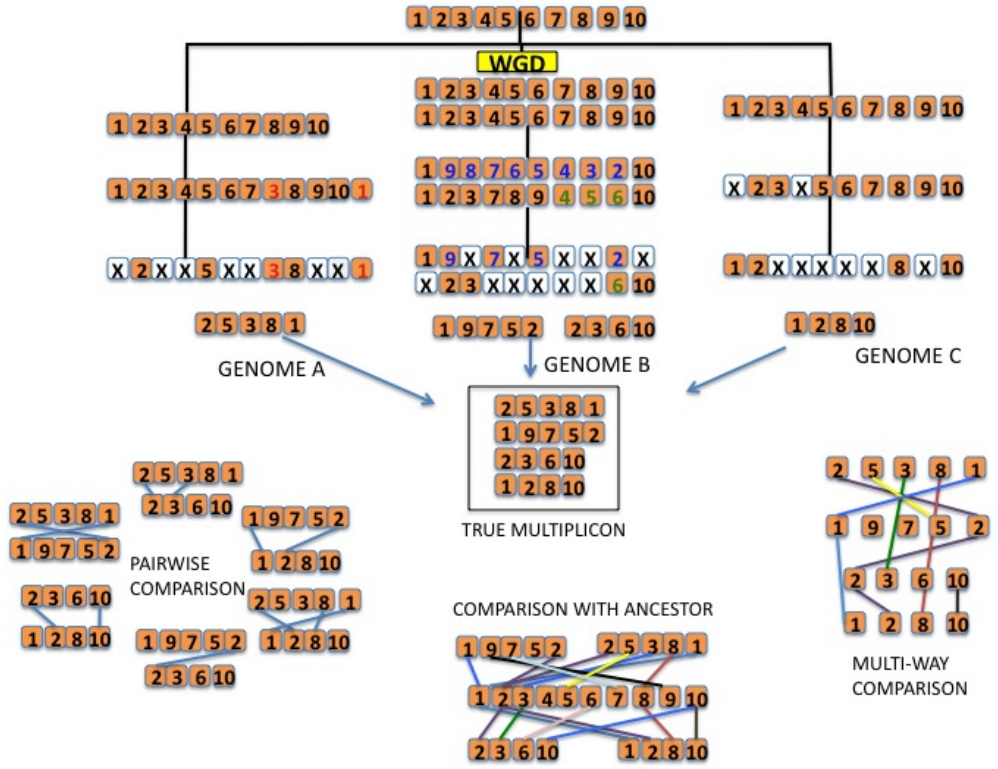


Figure 4.1: An Illustration of the difference between pairwise and multiway synteny detection. An ancestral genome segment with 10 genes is inherited by three extant species, one of which has undergone polyploidization. In addition, the segments have independently undergone single-gene duplications, transpositions, inversions, and many individual gene losses. The 'true' multiplicon contains two segments from genome B and one each from genomes A and C. All possible pairwise segmental homologs are shown in the lower left. Only one pair shares three anchors (indicated by blue lines), and, if that were the significance threshold in a pairwise comparison, only that one pair would be detected. However, in the lower right it can be seen that each segment has at least three anchors within the multiplicon as a whole. In the middle lower half of the Figure, each of the segments has high synteny with the ancestor. Pairwise synteny comparisons would not detect this multiplicon as a whole.

In all the programs mentioned above, a particular definition of synteny is assumed; either in terms of the genes shared in common between genomic segments, or the density between homologous genes on a segment. It is currently a challenge to be able to derive appropriate parameters for these two criteria, without knowledge of the rearrangement rates associated with the evolutionary process, or the length of the state of the ancestral and intermediary genomes. Different values for these parameters can yield different estimates of synteny amongst segments - ranging from inaccurate to over-estimates of synteny, to accurate, but overly stringent estimates of synteny.

In chapter 3, I evaluated *eAssembler* as a program to reconstruct gene order and content of syntenic genome segments and pointed out that the reconstructed ancestor is much more similar to its descendant genomes than they are to each other. In this chapter, I examine whether the reconstructions of syntenic genomic segments can identify more synteny that is sparse due to polyploidy and gene loss than through pairwise comparisons. I use *eAssembler* to reconstruct the ancestor of pairwise syntenic segments immediately prior to the WGD events that the segments are derived from. As the reconstruction has the union of all the genes in the segments, it can be used to cluster together segments fragmented by loss that do not share many genes in common. I also compared the depth of synteny identified using this approach with the multi-level synteny identified by *i-ADHoRe* to see if there was any increase in synteny detection.

There is no way to determine the ancestral gene order and content in the absence of a fossil DNA record. Therefore, I test the reconstruction and synteny identified with simulated genomic data as described in Chapter 3. The simulator has the advantage of being able to track the multiplicon through time. I measure the accuracy with which the synteny-detection programs identify the multiplicon before and after the use of the reconstruction programs. I use different evolutionary regimes to explore the effects of different rearrangement parameters on synteny detection.

I have also tested this approach on the genomic data sets of the plants Arabidopsis

Table 4.1: Parameter rates used in the simulations

| Parameter | Description | Dimensions | Range | Default |
|---|---|---|---|---|
| $A_G$ | Ancestral genome size | Number of genes | 50,500 | 50 |
| $\lambda_s$ | Speciation | number of events per unit time | [0.5,1.5] | 1.2 |
| $\lambda_p$ | Polyploidy | number of events per unit time | [0.5,1.5] | 0.5 |
| $\lambda_i$ | Inversion | calculated in a few ways | [30 ,750] | 120 |
| $\lambda_t$ | Translocation | number of events per unit time | [0.5,2] | 0.5 |
| $\lambda_d$ | Dispersed Duplication | duplication per gene per unit time | [0.5,2] | 0.5 |
| $\lambda_l$ | Gene Loss | loss per gene per unit time | [0.5,2] | 0.5 |

and rice. When *i-ADHoRe* was used previously on the combined data sets of Arabidopsis and rice an increase in synteny between and within the two genomes as well as in its levels was discovered. I compare the two approaches to see if using a reconstruction of a genome prior to its polyploidy events can uncover more synteny with a genome not inferred to have experienced those events and gain evidence for more ploidies in the other.

## 4.3 Methods

### 4.3.1 Multiplicon generation with Simulated Data

A *multiplicon*, as illustrated in (Figure 4.1), is a collection of two or more segments (as defined previously in the Methods section in Chapter 2). The ideal multiplicon is that in which the segments are descended from a common ancestral segment.

The simulator described in the previous chapter is run with a set of parameters (defined in chapter 3) summarized in the table below.

The default rates of speciation and polyploidy used yielded at least genome that had experienced two rounds of WGDs, with a total of 7-9 chromosomal segments in the input data set. As the simulations are generated from a unichromosomal ancestral genome, there is only one resultant true multiplicon. The dispersed duplications are locally

Table 4.2: *i-ADHoRe* input parameters

| Parameter | Description | Default values |
|---|---|---|
| Gap size | Maximum distance between two anchor genes in a cluster | 40 |
| Cluster gap size | Maximum distance between individual elements in a cluster | 50 |
| Q-value | Minimum $r^2$ value a cluster must have | 0.9 |
| Minimum number of anchors | Minimum anchor genes that a segment in a multiplicon should contain | 4 |
| Probability cut-off | Maximum probability that a cluster is generated by chance | 0.001 |

generated random homologies that do not derive from the starting ancestor chromosome.

These segments are sent to the program *i-ADHoRe* along with a file that details all the homologous pairs of genes in comparisons of the genomes pairwise. Unless otherwise specified, *i-ADHoRe* is run with its default input parameters summarized in the table 4.2 below along with what they stand for.

The pairwise profile with which the multiplicons were detected were collected and sent as input genomic segments to *eAssembler* with clustering parameters $\tau$ and $\Upsilon$ calculated as described in Chapter 3 based on the average length of input segment $r$, number of gene families $n_f$ in the simulated genomes and their lengths $n$. The resulting reconstruction is added to the original syntenic segments and sent back again to *i-ADHoRe* for synteny analysis. The accuracy with which the multiplicon is detected both before and after using the reconstruction is measured in the following ways:

### Counting by Anchors

Let $A_M$ denote the total number of anchors (matches) summed over all the intervals in the true multiplicon. Dispersed duplications are matches which do not actually derive from true synteny with the ancestral genome. I distinguish between anchors that are

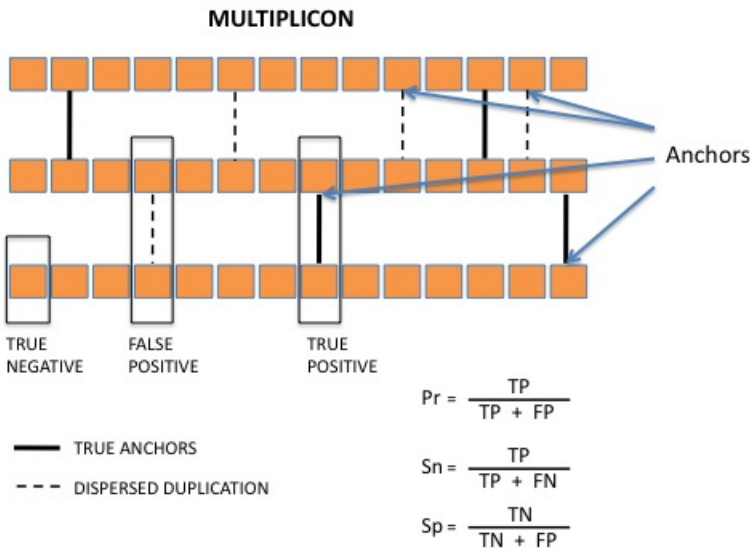


Figure 4.2: An example of counting anchors in 3 syntenic segments, with anchor genes that derive from synteny (solid black lines) and from dispersed duplications (dashed lines).

derived from true segmental homology as opposed to the 'noise' dispersed duplications. In every replicate, $A_P$ is the total number of anchors reported. From this, $A_{TP}$ is the number of anchors reported that are actually anchors in the true multiplicon, $A_{FP}$ is the number of anchors that are falsely reported as anchors (duplicative transposition matches), $A_{TN}$ is the number of genes not reported as anchors (i.e singletons), and $A_{FN}$ is the number of anchors not reported as anchors in the program results, but are anchors in $A_M$. Hence, $A_P = A_{TP} + A_{FP}$. These counts are counted from each segment pair, and are then summed over all segments reported by it i-ADHoRe.

Figure 4.2 summarized the categories of counts.

Using these counts, we can measure

$$Precision = \frac{A_{TP}}{A_{TP} + A_{FP}} \tag{4.1}$$

$$Specificity = \frac{A_{TP}}{A_{TP} + A_{FP}} \tag{4.2}$$

$$Sensitivity = \frac{A_{TN}}{A_{FP} + A_{TN}} \tag{4.3}$$

### Counting by Intervals

The intervals are counted in a similar fashion.

A true positive $I_{TP}$ is any subset of segment reported that corresponds to a segment in the true multiplicon. A true negative $I_{TN}$ is any subset of segment not reported as a syntenic segment that is not in the true multiplicon either. Similarly, a false positive $I_{FP}$ is any subset of segment that is reported in a syntenic segment but is not present in the true multiplicon, and a false negative $I_{FN}$ is any subset of the true multiplicon that is not reported as a syntenic segment.

These counts are summed over all the segments reported, and as described above, precision, sensitivity and specificity can be measured.

Figure 4.3 summarized the categories of counts.

### Counting by Multiplicon Levels

The levels of multiplicons for segments identified within the genome are reported for each experiment. If more levels of synteny are detected than with a pairwise comparison *i-ADHoRe* is expected to report more multiplicons with levels higher than 2. For the Arabidopsis-rice data, the percentage of the two genomes reported in multiplicons of different levels is reported. This also enables comparison with the analysis done in (86) for the Arabidopsis-rice synteny comparisons, with caveats about the sets that are compared which are described below.

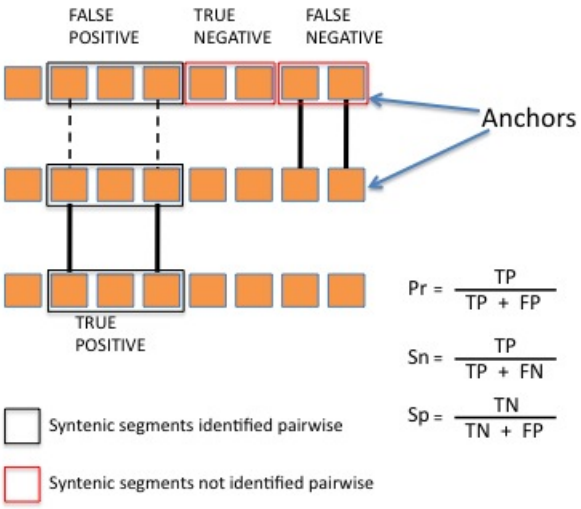


Figure 4.3: An example of counting intervals in 3 syntenic segments, with anchor genes that derive from synteny (solid black lines) and from dispersed duplications (dashed lines).

The quality of the reconstructions is measured again by Coverage and the three normalized induced distances, NB, NI and ND, detailed in Chapter 3. Coverage is measured as the proportion of genes in the reconstruction that are present in the ancestor. Normalized induced distances are measured as the distances between the ancestor and reconstructions obtained divided by the length of the reconstruction. The normalized induced distances calculated with the breakpoint, inversion and DCJ distances are NB, NI and ND respectively.

### 4.3.2 Angiosperm Data Analysis

The lineage leading to Arabidopsis, a eudicot, is known to have undergone multiple rounds of ancient WGD since the divergence of the monocots from the eudicots (25; 26). It has also undergone many re-arrangements, especially massive gene loss (106; 39; 25; 135). Rice, a monocot, is inferred to have undergone one and maybe two lineage-specific WGD events since the monocot-dicot divergence (117) and other re-arrangements as well in comparison with other cereals and angiosperms (134; 46).

The plant data was downloaded from Phytozome, a resource that facilitates comparative genomic studies amongst green plants http://www.phytozome.net. Genes both within and across genomes are clustered into families with a unique cluster family identifier based on the similarity metric between their associated peptides and evolutionary hierarchy. The data set that the authors used in (86) in their *i-ADHoRe* analysis of rice and Arabidopsis are different in the following ways. First, there have been changes in the annotation of both genomes since this study was done in 2004. Second, Phytozome has smaller gene families than the data set used by the authors and so the number of large-family homologies are reduced in the rice and Arabidopsis data set used in this chapter.

The positions of the genes on the chromosomes for each plant were downloaded from Phytozome. The gene order and sequence for Arabidopsis was the TAIR release 9 data

set, and for rice, the MSU Release 6.0. Homologies between rice and Arabidopsis genes were inferred from their Phytozome cluster ids. The list of all the homologies within and between Arabidopsis and rice was sent as input to *i-ADHoRe* along with the ordered list of the genes on their chromosomes.

Genome structure rearrangement rates were reviewed in the literature to inform parameters for the simulator are summarized in Table 4.3. As stated in chapter 3, the unit time in the simulations is intended to be roughly equivalent to 150 mya.

The rearrangements for most of the parameters were inferred from studies that used genetic linkage maps and in some cases, physical maps of the organisms studied. Comparative bayesian analyses with mapping data of the diploid *A. lyrata* with *A. thaliana* and *Capsella* were used to parsimoniously infer 2 inversions specific to the *A. thaliana lineage* (135). A physical map of *Z. mays* that covered 93.5 % of the genome and was integrated to 86.1% of its genetic map was compared with the *O. sativa* (rice) genome to infer 39 inversions in *Z.mays* since its divergence from *O.sativa* 50 mya ago (46). A genetic map for *Ae. tauschii* was used in a chromosomal orthology an paralogy analysis with *O. sativa* and *S. bicolor* to resolve rearrangements of which 27 inversions and 13 translocations were assigned to *Ae. tauschii*, 3 inversions and 5 translocations were assigned to *S. bicolor* and 1 inversion and 1 translocation assigned to *O. sativa* (134). Genetic linkage maps of *H. annus*, *H. argophyllus* and *H. petiolaris* were used to infer 2-4 inversions and 5 translocations relative to *H. annus* (45). The inversions in the chromosomes of *D. Melanogaster* and their length distributions have been studied and reported in (44).

For default rates in my simulations in this chapter, I used rates that were inferred with the highest quality map data over the time period that closest matched the time scale in my simulations. For inversions and translocations rates, this corresponds to the rates inferred from rice, sorghum and maize from (46). Maere et al. (112) developed an evolutionary model that simulations whole-genome and small-scale duplication and

Table 4.3: Genome Rearrangement parameters

| Rearrangement | Rates | Simulation Rate | Organism | Reference |
|---|---|---|---|---|
| Inversion | 1/50 mya | 3 | *O. sativa* | Luo et. al. (134) |
| | 3/50 mya | 9 | *S. bicolor* | Luo et. al.(134) |
| | 2/5 mya | 60 | *A. thaliana* | Yogeeswaran et. al. (135) |
| | 27/50 mya | 81 | *A. tauschii* | Luo et. al. (134) |
| | 39/50 mya | 117 | *Z. mays* | Wei et al. (46) |
| | $\frac{2-4}{0.75-1.67}$ mya | 295 - 817.5 | *H. annus* | Heesacker et al. (45) |
| | 10/1mya | 1500 | *D. melanogaster* | Richards et al. (44) |
| Dispersed Duplication | (0.03/0.1Ks) | 1.5 | *A. thaliana* | Maere et al.. (112) |
| Speciation | 1.2 | 1.2 | - | Nee (144) |
| Polyploidy | 2-4% of speciation events | 0.02 | angiopserms | Otto & Whitton (30) |
| | 7% in ferns | 0.04 | ferns | Otto & Whitton (30) |
| Translocations | 8/50 mya | 24 | *Z. mays* | (46) |
| | $\frac{5}{1.67}$ mya | 448.5 | *H. annus* | (45) |
| Loss | [0.5-1] | 0.7 | *A. thaliana* | Maere et al. (112; 115) |

Table 4.4: Arabidopsis and rice genomes

| **Property** | **Description** | |
|---|---|---|
| | Arabidopsis | rice |
| Number of chromosomes | 5 | 12 |
| Genome size | 115 Mb | 430 Mb |
| Number of genes (total) | 27098 | 40557 |
| Number of gene families | 14109 | 26005 |

loss dynamics of genes, which they fit to the Arabidopsis genome. As the rate of gene loss and single-gene duplication inferred from Arabidopsis data capture the dynamics of WGD, gene loss and single-duplication in a system like which I model in my simulations, I used their estimated rate loss for the whole genome which corresponds to 0.7 per gene per unit time and duplication rate which corresponds to 1.5 per gene per unit time as default parameters in my simulations. I used the other rates in Table 4.3 to estimate upper and lower bounds for the rates.

The genomic properties of Arabidopsis and rice are summarized in Table 4.4.

*i-ADHoRe* was run on the genome data sets of Arabidopsis and rice separately, as well as the combined Arabidopsis-rice data set with its default parameter settings. The parts of the genome that were not assembled into syntenic blocks were collected for each genome and for the combination separately and designated as 'orphan' intervals.

The syntenic segment pairs that were used as profiles for each multiplicon identified

from each run were sent as input to *eAssembler* with the breakpoint, inversion and DCJ distances, with $\tau$ determined to be 13 genes for the Arabidopsis data set ($\Upsilon = 12$) and 14 for the rice data set ($\Upsilon = 13$). *i-ADHoRe* was run with the same input parameters that was used in (86), which was a probability cut-off = 0.0001, a gap-size of 30 and a q-value of 0.9.

The reconstructions obtained from each run from one genome were collected together with its 'orphan' intervals, and re-submitted to *i-ADHoRe*, along with the other (current-day) genome. The idea behind this is that the reconstruction with the orphan intervals represents an approximate gene content and order of the ancestor of the genome relative to the ancestor at the time of divergence from the other genome. The results of the second *i-ADHoRe* runs are then analyzed to see whether any of the 'orphan' intervals of the current-day genome have now been assembled into syntenic blocks. This would imply that additional synteny has been identified. The percentage of genes in the genome now present in syntenic blocks of different levels was examined for before and after these two iterations of *i-ADHoRe* on the current-day genome data set. The amount of synteny detected before and after was compared.

The analysis presented in this chapter was done with building reconstructions of the Arabidopsis genome with *eAssembler* and the rice genome.

## 4.4 Results

I estimated if and by how much a reconstruction of ancestral gene order and content can unravel more synteny in polyploid systems, when used with synteny detection programs.

I set up simulations with a unichromosome genome containing $A_G = 50$ and 500 genes. A phylogeny was simulated with speciation rate $\lambda_s = 1.2$ and polyploidy rates $\lambda_p = 0.5$ as described in Table 4.1. The syntenic blocks identified in the simulated chromosomal segments were sent to *eAssembler* for reconstruction. Synteny analysis

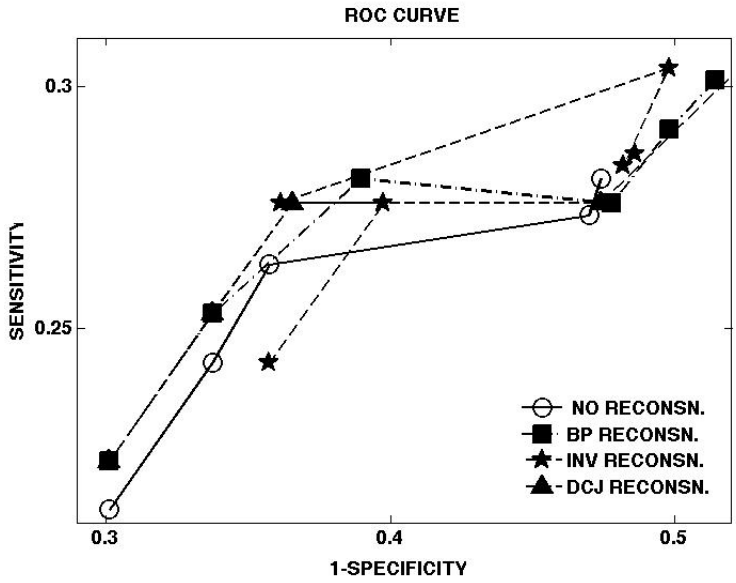


Figure 4.4: Variation in interval sensitivity vs specificity detected by *i-ADHoRe* alone (open circles), and *eAssembler*-aided *i-ADHore* with breakpoint (filled diamonds), DCJ (filled squares) and inversion (filled triangle) distances.

was performed again on the combined data set of the original chromosomal segments and the reconstructions obtained.

To determine how the input parameters to *i-ADHoRe* affect the accuracy of synteny identified in the input genomic segments, I measured the sensitivity and specificity in synteny detection both before and after running *i-ADHoRe* with the different *eAssembler* reconstructions. Hereafter, I refer to this procedure as 'before' and 'after' reconstructions. Shown here in Figure 4.4 is the variation in sensitivity and specificity for syntenic intervals reported for an input probability cut-off from [0.00001, 0.1]. There is a greater range in specificity values over sensitivity values. The turning point in the ROC curves correspond to the probability cut-off of 0.001, with which *i-ADHoRe* was run for all the experiments described here.

I measured synteny detection under four different parameter regimes, where I used either default or high values for the parameters described in 2.4. The first regime $HI$ had a high inversion rate $\lambda_i = 750$ with the rest at default, the second regime $HL$ had a

Table 4.5: Accuracy in synteny analysis with and without reconstructions under the $HI$ regime

| | **Before** | **After** | | |
|---|---|---|---|---|
| | | BP | INV | DCJ |
| **Intervals** | | | | |
| Sens | 0.203 | 0.221 | 0.206 | 0.206 |
| Spec | 0.861 | 0.841 | 0.852 | 0.858 |
| Prec | 0.689 | 0.687 | 0.682 | 0.691 |
| **Anchors** | | | | |
| Sens | 0.378 | 0.367 | 0.258 | 0.388 |
| Spec | 0.899 | 0.499 | 0.667 | 0.501 |
| Prec | 0.181 | 0.155 | 0.251 | 0.256 |
| **Performance Measures** | | | | |
| Coverage | | 0.186 | 0.172 | 0.185 |
| NB | | 0.597 | 0.567 | 0.593 |
| NI | | 0.762 | 0.805 | 0.808 |
| ND | | 0.633 | 0.663 | 0.675 |
| **Multiplicon Level** | | **Counts** | | |
| 2 | 7.8 | 2 | 3.1 | 3.1 |
| 3 | - | 4.1 | 4.3 | 4.4 |

high gene loss rate $\lambda_l = 2$ with the rest at default and the third regime $HD$ had a high rate of dispersed duplications $\lambda_d = 2$ with the rest set at default. In the fourth regime $AR$ I used the default parameters that I adapted for angiosperm rearrangement rates $[\lambda_i, \lambda_d, \lambda_l, \lambda_t] = [120, 1.5, 0.7, 1.5]$.

I report if any increase in the accuracy of synteny detection is observed when reconstructions are used in synteny analysis (**After**) in comparison to when they are not (**Before**).

Table 4.5 shows the results for the $HI$ regime with high inversion rate. $A_G = 50$ in these simulations. Values reported in the table are summarized for 10 simulations.

In terms of the intervals reported, there was no noticeable gain in accuracy with using the reconstructions. The breakpoint reconstruction provided a slight gain in sensitivity for intervals. Compared to the other methods, there was a $\sim 30\%$ decrease in sensitivity for anchors reported with the inversion reconstruction. The specificity in reporting anchors dropped from 25% to 45% with reconstructions. There was a 38% increase in the precision of reporting anchors with the inversion and DCJ reconstructions for this regime, but there was a decrease observed with the breakpoint reconstruction.

Amongst the different reconstructions, the breakpoint and DCJ reconstructions had a higher coverage than the inversion reconstruction. NI and ND were lowest for the breakpoint and NB was lowest for the inversion reconstructions.

Prior to using the reconstructions, only level 2 multiplicons were detected. The reconstructions enabled the detection of level 3 multiplicons within the original data set.

Table 4.6 summarizes the $HD$ regime simulations where $A_G = 50$. Values reported in this table are summarized for 10 simulations.

The sensitivity for all the methods in reporting intervals is very low. There is an increase in sensitivity for both intervals and anchors with the reconstructions. The breakpoint reconstruction yielded the highest sensitivity in reporting anchors. As was observed for the $HI$ regime, there is a decrease in specificity using the reconstructions. However, while the decrease in specificity in intervals drops from $\sim$ 20-30% with reconstructions, in reporting anchors there is only a $\sim 1\%$ difference. There was no noticeable change in precision in reporting intervals. In reporting precision for anchors however, there was a 25-30 % decrease.

The breakpoint and inversion reconstructions had a higher coverage than that of the DCJ. All three measures of NI, NB and ND were higher than that observed in the $HI$ regime. Of the three reconstructions, the inversion reconstruction had the highest values of NI, NB and ND.

Table 4.6: Accuracy in synteny analysis with and without reconstructions under the $HD$ regime

| | **Before** | **After** | | |
|---|---|---|---|---|
| | | BP | INV | DCJ |
| **Intervals** | | | | |
| Sens | 0.051 | 0.072 | 0.069 | 0.074 |
| Spec | 0.438 | 0.323 | 0.351 | 0.339 |
| Prec | 0.022 | 0.028 | 0.027 | 0.028 |
| **Anchors** | | | | |
| Sens | 0.311 | 0.378 | 0.327 | 0.325 |
| Spec | 0.817 | 0.816 | 0.807 | 0.806 |
| Prec | 0.249 | 0.186 | 0.174 | 0.183 |
| **Performance measures** | | | | |
| Coverage | | 0.546 | 0.547 | 0.514 |
| NB | | 0.873 | 0.881 | 0.863 |
| NI | | 0.901 | 0.921 | 0.912 |
| ND | | 0.861 | 0.879 | 0.867 |
| **Multiplicon Counts** | **Level** | | | |
| 2 | 116.2 | 248.1 | 250.3 | 217.33 |
| 3 | 22.5 | 73 | 73.9 | 68.3 |
| 4 | 12.8 | 12.5 | 14.3 | 20.2 |
| 5 | 13.5 | 10.3 | 7.4 | 12.2 |
| 6 | 13.2 | 5.8 | 4.2 | 4.5 |
| 7 | 6.3 | 1.6 | 1.8 | 3.1 |
| 8 | 0.8 | 0.4 | 0.1 | 0.6 |
| 9 | 0.9 | 0.1 | - | 0.5 |

Table 4.7: Accuracy in synteny analysis with and without reconstructions under the $HL$ regime

| | **Before** | **After** | | |
|---|---|---|---|---|
| | | BP | INV | DCJ |
| **Intervals** | | | | |
| Sens | 0.306 | 0.433 | 0.393 | 0.403 |
| Spec | 0.507 | 0.359 | 0.465 | 0.382 |
| Prec | 0.436 | 0.411 | 0.476 | 0.424 |
| **Anchors** | | | | |
| Sens | 0.235 | 0.292 | 0.302 | 0.302 |
| Spec | 0.895 | 0.796 | 0.836 | 0.801 |
| Prec | 0.689 | 0.579 | 0.671 | 0.632 |
| **Performance measures** | | | | |
| Coverage | | 0.131 | 0.093 | 0.108 |
| NB | | 0.869 | 0.854 | 0.865 |
| NI | | 0.911 | 0.885 | 0.909 |
| ND | | 0.889 | 0.891 | 0.861 |
| **Multiplicon Counts** | **Level** | | | |
| 2 | 13 | 16.6 | 8.6 | 8.6 |
| 3 | 2.5 | 6 | 5 | 6 |
| 4 | - | 1 | 0.8 | 0.5 |
| 5 | - | 0.16 | 0.16 | 0.16 |

It is interesting to note that this regime had the highest level reported in its multiplicons for all methods.

There was an increase in levels 2, 3 and 4 observed in the multiplicons reported with the breakpoint, inversion and distance methods, with the exception of the breakpoint method for level 4. However, there was a decrease in levels 5-9 observed with the reconstructions.

Table 4.7 shows the results of the $HL$ regime with $A_G = 500$. Values reported in this table are summarized for 10 simulations.

As in the case of the $HI$ and $HD$ regimes, there was a gain in sensitivity for intervals

and anchors with using reconstructions. The largest gain in sensitivity in intervals was a $\sim 41\%$ increase with the breakpoint reconstruction. In reporting anchors, there was a $\sim 28\%$ increase in sensitivity. There is a decrease in reporting specificity, from a largest decrease of $\sim 29\%$ in intervals to $\sim 11\%$ in anchors. Precision for the breakpoint reconstruction was the lowest reported for all four methods. In reporting intervals, there was a $\sim 9\%$ increase in precision with the inversion method.

The coverage with all three methods was very low for this regime and is the lowest observed in all four regimes. The breakpoint method had the highest values of NB, NI and ND observed. The inversion method had the lowest values of NB and NI and the DCJ method had the lowest value of ND.

In the multiplicons reported, there was a decrease in level 2 with using reconstructions, but an increase in level 3 multiplicons. Also, multiplicons of levels 4 and 5 which were not observed with *i-ADHoRe* alone were observed with the reconstructions.

Table 4.8 shows the results for the $AR$ regime. $A_G = 50$ for this regime. Values reported in the table are summarized for five simulations.

There was an increase in sensitivity for intervals and anchors with using reconstructions, except in the case of the DCJ reconstruction for intervals. The breakpoint reconstruction had the highest values of sensitivity reported with an increase of 16% and 34% for intervals and anchors respectively, over *i-ADHoRe* alone. Specificity decreased with the reconstructions for both intervals and anchors. Among the specificity reported in intervals and anchors with the reconstructions the DCJ reconstruction had the highest values. Unlike in other regimes, the DCJ reconstruction yielded the highest precision in reporting intervals and anchors.

The breakpoint reconstruction had the highest coverage for this regime and the inversion reconstruction has the lowest values of NB, NI and ND.

In the levels of multiplicons reported, there was a decrease in level 2 multiplicons detected with using reconstructions and an increase in level 3 multiplicons. Level 4

Table 4.8: Accuracy in synteny analysis with and without reconstructions under the simulated angiosperm data $AR$ regime

| | **Before** | **After** | | |
|---|---|---|---|---|
| | | BP | INV | DCJ |
| **Intervals** | | | | |
| Sens | 0.567 | 0.658 | 0.629 | 0.535 |
| Spec | 0.513 | 0.412 | 0.449 | 0.511 |
| Prec | 0.194 | 0.188 | 0.191 | 0.245 |
| **Anchors** | | | | |
| Sens | 0.284 | 0.383 | 0.321 | 0.302 |
| Spec | 0.848 | 0.746 | 0.712 | 0.798 |
| Prec | 0.181 | 0.207 | 0.151 | 0.258 |
| **Performance Measures** | | | | |
| Coverage | | 0.392 | 0.342 | 0.366 |
| NB | | 0.843 | 0.798 | 0.841 |
| NI | | 0.916 | 0.897 | 0.917 |
| ND | | 0.859 | 0.831 | 0.861 |
| **Multiplicon Counts** | | **Level** | | |
| 2 | 32 | 19 | 21 | 21 |
| 3 | 1.5 | 21 | 18 | 18 |
| 4 | - | 1 | 1 | - |

Table 4.9: Percentage of the Arabidopsis and rice genome in multiplicons from various synteny analyses

| | **Arabidopsis** | | | **rice** | | |
|---|---|---|---|---|---|---|
| Multiplicon Level | % genome | segments | anchors | % genome | segments | anchors |
| 2 | 41 | 83 | 2250 | 16 | 93 | 2504 |
| 3 | 5 | 14 | 232 | 4 | 18 | 272 |
| 4 | 1.8 | 4 | 62 | 0.7 | 2 | 36 |
| 5 | 1.2 | 2 | 64 | 0.09 | 1 | 46 |
| | | | **Combined** | | | |
| | | **Arabidopsis** | | | **rice** | |
| 2 | 54 | 308 | 2989 | 37 | 328 | 3243 |
| 3 | 13 | 119 | 647 | 8.7 | 112 | 701 |
| 4 | 5.6 | 59 | 236 | 2.5 | 33 | 244 |
| 5 | 2.4 | 25 | 125 | 1.2 | 15 | 91 |
| 6 | 0.7 | 9 | 52 | 0.09 | 9 | 32 |

multiplicons were detected with the breakpoint and inversion reconstructions that were not detected without reconstruction or for the DCJ reconstructions.

### 4.4.1 Angiosperm Data

I compared the synteny detected by using *eAssembler* reconstructions to that detected by *i-ADHoRe* alone with data from the plants Arabidopsis and rice.

Previously, *i-ADHoRe* had been used to detect multiplicons of level up to 4 in rice, and up to level 10 in Arabidopsis in a synteny analysis where it was used on each genome separately. In a combined analysis with both genomes, rice segments were present in multiplicons of level 5 and Arabidopsis in multiplicons of level 11 (86).

As the data set I used in this chapter is different from what was used in the previous study, I repeated this analysis for the Arabidopsis, rice and combined Arabidopsis-rice data sets.

Summarized in table 4.9 are the properties of the syntenic segments identified with *i-ADHoRe* in Arabidopsis alone, rice alone and in Arabidopsis and rice combined.

Table 4.10: Percentage of rice genome in multiplicons from various synteny analyses

| **Level** | Rice-Only | Rice-Arabidopsis | Breakpoint | Inversion | DCJ |
|---|---|---|---|---|---|
| 2 | 16 | 37 | 34.17 | 33.12 | 34.64 |
| 3 | 4 | 8.7 | 5.9 | 6.4 | 5.92 |
| 4 | 0.7 | 1.2 | 0.91 | 0.85 | 0.85 |
| 5 | 0.09 | 0.09 | 0.58 | 0.58 | 0.59 |

Synteny analysis on Arabidopsis and rice individually yielded level 5 multiplicons. Compared to the previous *i-ADHoRe* study where no level higher than 5 was found in rice-only synteny, an additional level of 5 was found here in the rice-only synteny analysis. 49% of the Arabidopsis genome was identified reported within multiplicons, compared to the previous estimate of 82.9% (86). The estimate of 20.8% of rice in multiplicons however is comparable to the previous estimate of 20.9% (86). This is probably due to the different and more current annotation of the Arabidopsis and rice genome data set used in this chapter as explained in the Methods.

Synteny analysis on the combined data set of Arabidopsis and rice identified many more syntenic segments and anchors in both genomes than in the single-genome analyses. The percentage of the genome assigned to multiplicons increased from 49% to 75.7% in Arabidopsis and from 20.8% to 49.5% in rice. Additionally, 3 multiplicons of level 6 were identified in the combined Arabidopsis-rice analysis.

The syntenic segments identified by *i-ADHoRe* in the Arabidopsis-only comparison were sent to *eAssembler* for reconstruction. The resulting reconstructions were combined with the rice genomic data set and the regions in Arabidopsis that were not identified in multiplicons with the *i-ADHoRe* Arabidopsis-only analysis. The syntenic segments and anchors identified in rice with this iteration of *i-ADHoRe* were then compared to the syntenic segments in rice identified in the rice-only and Arabidopsis-rice analysis.

The largest syntenic segment size in the Arabidopsis input data set to *eAssembler* was 594 genes in length. The largest reconstructed segment in Arabidopsis was 756 genes in length for the breakpoint and 688 genes in length for the inversion and DCJ

Table 4.11: Comparisons of the percentage of the rice genome identified in multiplicons in different synteny analyses

| **Reconstruction** | | **rice-only** | | **Combined** | |
|---|---|---|---|---|---|
| | % genome | % new | % undetected | % new | % undetected |
| BP | 27.3 | 2.79 | - | 0.61 | 2.7 |
| INV | 27.8 | 4.16 | - | 0.73 | 2.38 |
| DCJ | 27.6 | 3.98 | - | 0.78 | 2.61 |

reconstructions.

The differences in the multiplicons obtained in terms of intervals and anchors reported are shown in Table 4.11. The first column is the percentage of the rice genome detected in syntenic blocks. The % age of genes that are newly detected in syntenic segments in comparison with the previous reports of rice-only and Arabidopsis-rice combined data sets are shown here as are the % of genes that are not detected from these previous reports. As all the intervals detected in the rice-only synteny analysis are detected with the reconstruction analysis, the third column in Table 4.11 is empty.

Using the breakpoint, inversion and DCJ reconstructions, new syntenic segments and anchors were found in rice when compared to what was found in the rice-only and Arabidopsis-rice analysis. The percentage of the rice genome reported within syntenic blocks $\sim$ 27 % is comparable for the three reconstructions and was higher than the 20.8% of the rice genome that was detected in the rice-only synteny analysis. Among the different reconstructions, more of the rice genome was detected in syntenic blocks with using the inversion reconstruction in comparison to the rice-only analysis. The DCJ reconstruction synteny analysis detected a higher percentage of rice genome that was not detected by the previous combined synteny analysis. $\sim$ 2.7% of the genome that was found previously in the combined synteny analysis was not detected using the reconstructions.

Table 4.12 shows the distributions of multiplicons of different levels identified within

Table 4.12: Difference in Multiplicon Levels between the different synteny analyses

| **Experiment** | **Multiplicon Levels** | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| Rice-Only | 93 | 18 | 2 | 1 |
| Arabidopsis-rice | 120 | 18 | 2 | 2 |
| Breakpoint | 278 | 87 | 15 | 5 |
| Inversion | 74 | 86 | 14 | 5 |
| DCJ | 80 | 89 | 14 | 5 |

the rice genome. Many more multiplicons were reported for levels 3, 4 and 5 with the reconstruction that with either the rice-only or Arabidopsis-rice combined data sets. No new levels of multiplicons were identified in this analysis from what was reported before. It is important to recollect that these increases account for between 0.5% - 0.7% of new regions in the genome that are identified with the reconstructions. It is interesting to note that there were 278 level 2 multiplicons reported with the breakpoint reconstruction, more than 3.5 times that reported for the other reconstructions.

Within the orphan gene intervals of Arabidopsis, 0.7% of them were identified in multiplicons with rice. In comparison with the combined Arabidopsis-rice synteny analysis, 4.83%, 4.69% and 4.81% of the Arabidopsis genome were additionally identified with using the breakpoint, inversion and DCJ reconstructions.

## 4.5 Discussion

Using simulated data sets, I have demonstrated that use of *eAssembler* reconstructions in conjunction with the synteny detection program *i-ADHoRe* can provide a gain in sensitivity in identifying true multiplicons. With reconstructions, the sensitivity in reporting intervals was the highest for simulations of the angiosperm data regime and lowest for the high dispersed duplication regime. Sensitivity reports were lower for the

inversion regimes than for the high loss regimes. The sensitivity in reporting anchors were comparable for all the regimes when reconstructions are used. There was an increase in the level of multiplicons detected in all the simulation regimes, except for the regime with high dispersed duplications. Dispersed duplications had the most adverse impact on synteny detection with and without reconstruction, at $\lambda_d = 2$. Studies have estimated that the rate of dispersed duplications in wheat range from 2.5 x $10^{-3}$ per gene per Myr to 5.2 x $10^{-2}$ per gene per Myr. This corresponds to values of 0.375 - 7.8 in my simulations. Genomes like wheat, therefore, can be expected to show low synteny. Among the other rearrangement processes, the high inversion rate affected synteny prediction more than the high loss rate. The quality of reconstruction did not reflect the accuracy of prediction of synteny. The highest coverage for the reconstructions was obtained in the high dispersed duplication regimes and the lowest for the high loss regime. The lowest normalized induced distances were observed for the high inversion regimes.

Within the combined reconstruction-synteny detection analysis, different reconstruction methods yield different estimates of synteny, as observed in the Results section. In systems like those of the cereals, maize is inferred to have undergone a high number of inversions since its divergence from rice (46; 134). The maize-rice comparison would correspond to the *HI* regime tested here. Therefore, the DCJ method should be used for the highest gains in synteny reported. A recent study has inferred a higher rate of dispersed duplications within Arabidopsis than previously suspected (39). For a comparison of synteny within Arabidopsis, the breakpoint reconstruction-synteny method should be used for the highest gains in synteny reporting, as inferred from the tests in the *HD* regime. For a system that corresponds to the high loss regimes tested here, as for the angiosperms (112) or the yeast genomes (1), there is no one method that can be suggested for the highest yields for all measures of performance in synteny reports. However, among them, the inversion method yielded the highest reports for specificity and precision. From the results for the angiosperm regime simulations themselves, the

DCJ method is expected to yield the highest reports of synteny in terms of specificity, precision and quality of reconstruction when used within an angiosperm synteny analysis; for example, a comparison with rice, Arabidopsis, poplar, maize, grape, mimulus, etc.

Using the reconstructions in synteny analysis increased the amount of synteny reported in rice genomic data. Based on the simulations, the reconstruction of the Arabidopsis ancestor was estimated to have a coverage of 34 - 40 % of the Arabidopsis ancestor genes with an estimated normalized induced distance of 0.8 - 0.9 from it. In comparison with the simulations, the increase of 0.61 - 0.78 % of the genome detected in the synteny analysis with the reconstructions corresponds to a 16% and 34% increase in sensitivity in reporting syntenic intervals and anchors, respectively. In contrast to the simulations, no increase in levels of multiplicons reported were detected with reconstructions for the rice genome. The increase in level 3, 4 and 5 multiplicons in rice detected with the reconstruction synteny analysis in comparison to the combined rice-Arabidopsis analysis and the additional presence of the previously 'orphan' regions of Arabidopsis might suggest additional evidence of an older duplication event in the rice genome than previously inferred in (86). However, this increase corresponded to 0.7% of the rice genome and $\sim$ 5% of the Arabidopsis genome, which cannot be considered substantial evidence for an additional polyploidy event.

Most studies in synteny detection that reported increases with multi-genome comparisons (86; 28) report an increase in the number of genomic segments that are found to be syntenic. The simulator used in this chapter incorporates key biological assumptions and is useful in evaluating the accuracy of the synteny detected. Local and segmental duplications occur at different rates in both vertebrate and angiosperm genomes (118). Homology that issues from these duplication events can interrupt synteny from older WGD events. Polyploidy, a parameter in this simulation, is usually not modeled in simulations that have been generated to test genome reconstruction algorithms and synteny

detection programs. The impact of polyploidy and subsequent gene loss can be studied from these simulations as I have demonstrated and is very important in understanding angiosperm synteny. Evaluating the accuracy of synteny detection programs using these simulations can help us in interpreting the homologies detected by these programs. For the regimes tested in this chapter, the simulations were useful in assessing the accuracy of *i-ADHoRe* and contribution of reconstructions in identifying true genomic synteny.

There are many directions for future work. Other processes that I have not incorporated into my simulator may impact conservation of synteny. Modeling lineage-specific rates at different nodes in the phylogeny could model angiosperm data better. The estimates I considered for inversion had a large range, varying from 3 per 150 million years for rice to 817.5 per 150 million years for *H. annus*. The The model I developed in Chapter 2 differentiated between the rate of gene loss following polyploidy and a background rate of loss. The simulations here can be modified to model different sets of rates on branches with parent WGD labels. Rearrangements rates were inferred to increase after WGD in teleost fish with additional variability in the rates across species (145). Modeling different rates for different genomes after WGD in the simulations could better describe the disruption of synteny post WGD. It is also not known how incompleteness in genetic mapping of the genome might affect synteny analysis. Explicitly modeling incomplete segments of synteny where anchor genes present are simulated to be absent could uncover how obscured synteny is in current-day plant genomes.

# Chapter 5

# Conclusions

In this dissertation, I set out to study the evolution of genome structure by modeling how synteny is preserved in genomes and how far back we are able to detect it in time. In particular, I examined synteny evolution in polyploid genomes.

In Chapter 2, I assessed the effects of two processes that contribute to synteny rearrangement: gene loss (immediately following WGD and otherwise) and gene transposition. I developed probabilistic models that account for the effects of different sets of these processes simultaneously. Using these models, for both simulated and genomic data, I found that gene content within syntenic regions of unsequenced genomes can be predicted with high accuracy for a non-trivial amount of input data. Of the factors examined, the largest increase in accuracy of prediction came with an increase in the number of input segments in the data set. Among the different models tested, accounting for the two kinds of gene loss and gene transposition yielded gains in sensitivity of gene content prediction for values of specificity that were higher than 0.8. For other ranges of specificity and for different ranges of input data parameters tested, the differences in predictions between models were not that profound.

With these models, I was able to model synteny evolution of gene content following WGD events. Building on the assumptions made in these models can yield deeper insight into the mechanisms by which different genome rearrangement processes impact

synteny. The framework of the model allows for adding more realistic components that have been observed in genomic data, like rate variation among different lineages (115), gene- and branch-specific distribution of rates (57), etc.

In Chapter 3, I evaluated the use of alternative distance measures in *eAssembler* (2), a heuristic method that reconstructs ancestral gene order and content for syntenic genomic segments, particularly for segments derived from multiple rounds of WGD events. Such reconstructions are useful in deducing the divergence in synteny between related genomes. I also evaluated using data-derived clustering parameters for the algorithm over user-defined ones. Alternative distance measures are thought to capture the divergence between species by accounting for specific rearrangements than the distance currently used in *eAssembler*. Using simulations that included all of the rearrangements considered in this dissertation, I generated syntenic genomic segments for which I used *eAssembler* using different distance measures to reconstruct the starting pre-WGD ancestral gene order and content. By measuring how close the reconstructions were to the ancestral configurations used in the simulations, I found that different distance methods produce differences in the quality of reconstruction. This implies that the use of distance measure for deriving reconstructions should be chosen based on the properties of rearrangements within the underlying syntenic genomic regions. I also found that data-derived clustering parameters yielded the highest quality reconstructions over arbitrary choices for these parameters.

Reconstruction algorithms should closely account for the properties of the genomic segments for which the reconstructions are to be assembled, as accounting for one or a few rearrangements does not capture the effects of the underlying biological processes. Tailoring the reconstruction to underlying genomic properties might prove to be computationally expensive; however, the costs in obtaining the reconstructions could be offset by the accuracy in the reconstructions.

In Chapter 4, I evaluated the differences in accuracy of synteny detected in genomes

that are descendants of WGD events between two approaches. The first approach is through methods that detect profiles of synteny among multiple pairwise comparison like in the program *i-ADHoRe* (86) and the second is through using the reconstructions (using *eAssembler* (2)) of the ancestors of genomic segments that are identified as syntenic through pairwise comparisons. I used simulated data that incorporated all of the rearrangements considered in this dissertation to test these two approaches and also applied them to a set of angiosperm genomic data. I found that using the second approach, i.e. using reconstructions, can provide a gain in identification of true genomic synteny. The measures of accuracy of synteny varied depending on the regimes of rearrangements tested, as in regimes that predominantly experienced dispersed gene duplications, for example. The use of different distance methods in the reconstructions also contributed to differences in the accuracy of synteny detected. On the angiosperm data set of rice and Arabidopsis, the synteny detected by the two approaches were comparable to previously reported measures of synteny (86). Both approaches led to identification of synteny within the rice genome that was novel to each approach, but did not correspond to a WGD event that had not been inferred in previous studies (86).

Using simulations that incorporate different rearrangements is very useful in evaluating different approaches to synteny detection, particularly incorporating WGD events. Therefore, modeling additional rearrangements not accounted for in this dissertation like inter-chromosomal inversions, for example, or more sophisticated modes of the rearrangements themselves, like lineage-specific gene loss, could help in better estimates of the accuracy of synteny detected by the approaches considered in Chapter 4.

Through the studies in this dissertation, I have demonstrated that modeling the complexities involved in synteny rearrangement can improve our understanding of the evolution of genome structure. Extending these models and approaches outlined here by incorporating more biologically realistic assumptions and sophisticated rate approximations, for example, in future research can further our understanding of the underlying

mechanisms that shape genome structure.

# Bibliography


[1] J. Gordon, K. P. Byrne, and K. H. Wolfe, "Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome," *PLoS Genet*, vol. 5, p. e1000485, 05 2009. xi, 29, 31, 39, 44, 45, 46, 49, 67, 97

[2] J. Huan, J. Prins, W. Wang, and T. Vision, "Reconstruction of ancestral gene order after segmental duplication and gene loss," in *CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics*, (Washington, DC, USA), p. 484, IEEE Computer Society, 2003. xii, 51, 53, 54, 55, 101, 102

[3] L. et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921. 1

[4] V. et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001. 1

[5] G. et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, 1996. 1

[6] A. et al., "The gneome sequence of drosophila melanogaster," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000. 1

[7] C. elegans Sequencing Consortium, "Genome sequence of the nematode c. elegans: a platform for investigating biology," *Science*, vol. 282, no. 5396, pp. 2012–2018, 1998. 1

[8] A. G. Intitiative, "Analysis of the genome sequence of the flowering plant arabidopsis thaliana," *Nature*, vol. 408, no. 6814, pp. 796–815. 1

[9] M. Gale and K. Devos, "Comparative genetics in the grasses," *Proceedings of the National Academies of Sciences*, vol. 95, pp. 1971–1974. 1, 48

[10] S. Ahn and S. Tanksley, "Comparative linkage maps of the rice and maize genomes," *Proceedings of the National Academy of Sciences of the United States*, vol. 90, pp. 7980–7984, 1993. 1

[11] G. Moore, K. Devos, Z. Wang, and M. Gale, "Cereal genomeevolution, grasses, line up and form a circle," *Current Biology*, vol. 5, pp. 737–739, 1995. 1, 11, 12, 16

[12] P. et al., "Convergent domestication of cereal crops by independent mutations at corresponding genetic loci," *Science*, vol. 269, no. 5231, pp. 1714–1718. 1, 11

[13] D. et al, ed., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999. 2, 72

[14] M. Ferguson-Smith and V. Trifonov, "Mammalian karyotype evolution," *Nature Reviews Genetics*, vol. 8, pp. 950–962, 2007. 3

[15] K. Byrne and K. Wolfe, "The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species," *Genome Research*, vol. 15, 2005. 3

[16] K. Vandepoele, C. Simillion, and Y. Van De Peer, "Detecting the undetectable: uncovering duplicated segments in arabidopsis by comparison with rice," *Trends in Genetics*, vol. 18, no. 7161, pp. 606–608, 2002. 3, 16, 32, 74

[17] M. et al., "The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus)," *Nature*, vol. 452, pp. 991–996, 2008. 3

[18] M. Gale and K. Devos, "Plant comparative genetics after 10 years," *Science*, vol. 282, no. 5389, pp. 656–659, 1998. 3, 72

[19] H. Tang, J. Bowers, X. Wang, R. Ming, M. Alam, and A. Paterson, "Synteny and collinearity in plant genomes," *Science*, vol. 320, pp. 486–488, 2008. 3, 12, 17

[20] S. Ohno, *Evolution by gene duplication*. Springer-Verlag, 1970. 3

[21] S. et al., "Polyploidy and angiosperm diversification," *Am. J. Bot.*, vol. 96, no. 1, pp. 336–348, 2009. 3, 17, 49, 67

[22] A. McLysaght, K. Hokamp, and K. H. Wolfe, "Extensive genomic duplication during early chordate evolution," *Nature Genetics*, vol. 31, pp. 200–204, 2002. 3, 16, 17

[23] G. I. Woods, C. Wilson, B. Friedlander, P. Chang, D. Keyes, R. Nix, P. Kelly, F. Chu, J. Postelthwait, and W. Talbot, "The zebrafish gene map defines ancestral vertebrate chromosomes," *Genome Research*, vol. 15, pp. 1307–1314, 2005. 3

[24] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome ," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997. 3, 16, 32

[25] J. Bowers, B. Chapman, J. Rong, and A. Paterson, "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events," *Nature*, vol. 422, pp. 433–438, Mar 2003. 10.1038/nature01521. 3, 4, 12, 17, 32, 67, 73, 82

[26] T. Vision, D. G. Brown, and S. D. Tanksley, "The Origins of Genomic Duplications in Arabidopsis," *Science*, vol. 290, no. 5499, pp. 2114–2117, 2000. 3, 32, 49, 73, 82

[27] L. Cui, P. Wall, J. Leebens-Mack, B. Lindsay, S. D.E., J. Doyle, and P. Soltis, "Widespread genome duplications throughout the history of flowering plants," *Genome Research*, vol. 16, 2006. 3

[28] J. et al, "Genome duplication in the teleost fish *tetraodon nigroviridis* reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, 2004. 3, 98

[29] A. et al., "Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia," *Nature*, vol. 444, pp. 171–178, 2006. 3

[30] P. Otto and J. Whitton, "Polyploid incidence and evolution," *Annual Reviews in Genetics*, vol. 34, pp. 401–437, 2000. 3, 84

[31] S. Otto, "The evolutionary consequences of polyploidy," *Cell*, vol. 131, no. 3, pp. 452–462, 2007. 3

[32] J. Wendel, "Genome evolution in polyploids," *Plant Molecular Biology*, vol. 42, pp. 225–249. 4

[33] Y. Van de Peer, S. Maere, and A. Meyer, "The evolutionary signficance of ancient genome duplications," *Nature Reviews Genetics*, vol. 10, 2009. 4

[34] T. Vision, "Gene order in plants: a sure but slow shuffle," *New Phytologist*, vol. 168, pp. 51–60, 2005. 4, 5, 44

[35] H. Ku, T. J. Vision, J. Liu, and S. Tanksley, "Comparing sequenced segments of the tomato and arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny," *Proceedins of the Natural Academy of Sciences of the United States of America*, vol. 97, no. 16, pp. 9121–9126, 2000. 4, 17, 67

[36] M. et al, "Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of arabidopsis thaliana," *Genome Research*, vol. 11, pp. 1167–1174, 2001. 4, 5, 17, 67

[37] L. Flagel and J. Wendel, "Gene duplication and evolutionary novelty in plants," *New Phytologist*, vol. 183, pp. 557–564, 2009. 4, 5

[38] M. Hurles, "Gene duplication: the genomic trade in spare parts," *PLoS Biology*, vol. 2, 2004. 5

[39] M. Freeling, E. Lyons, B. Pedersen, M. Alam, R. Ming, and D. Lisch, "Many or most genes in arabidopsis transposed after the origin of the order brassicales," *Genome Research*, vol. 18, Oct 2008. 5, 17, 46, 82, 97

[40] C. Rizzon, L. Ponger, and B. Gaut, "Striking similarities in the genomic distribution of tandemly arrayed genes in arabidopsis and rice," *PLoS Computational Biology*, vol. 2, 2006. 5

[41] K. Kashkush, M. Feldman, and A. Levy, "Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat," *Nature Genetics*, vol. 33, pp. 102–106, 2003. 5

[42] A. Paterson, B. Chapman, J. Kissinger, J. Bowers, F. Feltus, and J. Estill, "Many gene and domain families have convergent fates following independent whole-genome duplication events in arabidopsis, oryza, saccharomyces and tetraodon," *Trends in Genetics*, vol. 22, pp. 597–602, 2006. 5

[43] E. Eichler and D. Sankoff, "Structural dynamics of eukaryotic chromosome evolution," *Science*, vol. 8, no. 5634, pp. 793–797, 2003. 5

[44] R. et al, "Comparative genome sequencing of drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution," *Genome Research*, vol. 15, pp. 1–18, 2005. 5, 83, 84

[45] A. Heesacker, E. Bachlava, R. Brunick, J. Burke, L. Rieseberg, and S. Knapp, "Karyotypic evolution of the common and silverleaf sunflower genomes," *The Plant Genome*, vol. 2, no. 3, pp. 233–246. 5, 67, 83, 84

[46] W. et al., "Physical and genetic structure of the maize genome reflects its complex evolutionary history," *PLoS Genetics*, vol. 3, no. 7, 2007. 5, 12, 67, 82, 83, 84, 97

[47] S. et al, "Prevalence of small inversions in yeast gene order evolution," 2000. 5

[48] B. C. Thomas, B. Pedersen, and M. Freeling, "Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes.," *Genome Research*, vol. 16, pp. 934–946, Jul 2006. 10.1101/gr.4708406. 5, 44, 46

[49] G. Fertin, A. Labarre, I. Rusu, E. Tannier, and S. Vialette, *Models Handling Duplications: Strings*, pp. 89–95. The MIT Press, 2009. 6, 51

[50] T. Dobzhansky and A. Sturtevant, "Inversions in the third chromosomes of wild races of drosophila pseudoobscura, and their use in the study of the history of the species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 33, no. 7, pp. 448–450, 1936. 6

[51] A. Sturtevant and E. Novitski, "The homologies of the chromosome elements in the genus drosophila," *Genetics*, vol. 26, no. 5, pp. 517–541, 1941. 6

[52] T. Jukes and C. Cantor, "Evolution of protein molecules," vol. 3, pp. 21–32, 1969. 6

[53] J. Felsenstein, "Evolutionary trees from dna sequences: a maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376. 6, 24, 46

[54] M. Kimura, "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences," *Journal of Molecular Evolution*, vol. 16, pp. 111–120. 7

[55] M. Kimura, "Estimation of evolutionary distances between homologous nucleotide sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 1, pp. 454–458. 7

[56] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial dna," *Journal of Molecular Evolution*, vol. 22, pp. 160–174, 1985. 7

[57] J. Thorne, H. Kishino, and I. Painter, "Estimating the rate of evolution of the rate of molecular evolution," *Mol Biol Evol*, vol. 15, no. 12, pp. 1647–1657, 1998. 7, 46, 101

[58] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, pp. 512–526. 7

[59] A. Siepel and D. Haussler, "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood," *Molecular Biology and Evolution*, vol. 21, p. 468, Mar 2004. 7, 18

[60] I. Holmes and G. M. Rubin, "An expectation maximization algorithm for training hidden substitution models," *Journal of Molecular Biology*, vol. 317, pp. 753–764, 2002. 7, 17

[61] D. Huson and M. Steel, "Phylogenetic trees based on gene content," *Bioinformatics*, vol. 20, no. 13, pp. 2044–2049. 7

[62] X. Gu and H. Zhang, "Genome phylogenetic analysis based on extended gene contents," *Molecular Biology and Evolution*, vol. 21, Jul 2004. 10.1093/molbev/msh138. 7, 17, 18

[63] C. S. McBride, J. R. Arguello, and B. C. O'Meara, "Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily," *Genetics*, vol. 177, no. 3, pp. 1395–1416, 2007. 7, 17, 18

[64] D. Barker, A. Meade, and M. Pagel, "Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes," *Bioinformatics*, vol. 23, pp. 14–20, Jan 2007. 7, 17

[65] J. Palmer and L. Herbon, "Plant mitochondrial dna evolves rapidly in structure, but slowly in sequence," *Journal of Molecular Evolution*, vol. 28, pp. 87–97, 1988. 7

[66] G. Watterson, W. Ewens, T. Hall, and A. Morgan, "The chromosome inversio problem," *Journal of Theoretical Biology*, vol. 99, no. 1, pp. 1–7, 1982. 7

[67] D. Sankoff, *Edit distance for genome comparison based on non-local operations*, vol. 644. Springer Berlin/Heidelberg, 1992. 8

[68] P. Pevzner and M. Waterman, "Open combinatorial problems in computational molecular biology," *Proceedings of the 3rd Israel Symposium on Thoery of Computing and Systems*, 195. 8

[69] V. Ferretti and D. Sankoff, "Phylogenetic invariants for more general evolutionary models," *Journal of Theoretical Biology*, vol. 173, no. 2, pp. 147–162, 1995. 8

[70] S. Hannenhalli and P. Pevzner, "Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals).," *Proc. 27th Ann. Symposium on Theory of Computing*, pp. 178–179, 1995. 8, 50, 52

[71] B. Moret, A. Siepel, J. Tang, and T. Liu, "Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data," *2nd International Workshop on Algorithms in Bioinformatics (WABI 2002)*, vol. 2452, pp. 521–536, 2001. 8, 50, 52, 58, 67

[72] D. Sankoff, D. Bryant, M. Deneault, B. Lang, and G. Burger, "Early eukaryote evolution based on mitochondrial gene order breakpoints," in *RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology*, (New York, NY, USA), pp. 254–262, ACM, 2000. 8, 50, 51, 54

[73] D. Sankoff and M. Blanchette, vol. 1276. 1997. 8

[74] V. Bafna and P. Pevzner, *Sorting permutations by transpositions*, pp. 614–623. Society for Industrial and Applied Mathematics, 1995. 8

[75] J. Kececioglu and R. Ravi, *Of mice and men: algorithms for evolutionary distances between genomes with translocation*. Society for Industrial and Applied Mathematics, 1995. 8

[76] S. Hannenhalli, *Polynomial algorithm for computing translocation distance between genomes*. Springer-Verlag, Berlin, 1995. 8

[77] S. Yancopoulus, O. Attie, and R. Friedberg, "Efficient sorting of genomic permutations by translocation, inversion and block interchange," *Bioinformatics*, vol. 21, no. 16, 2005. 9, 50

[78] C. Zheng, Q. Zhu, Z. Adam, and D. Sankoff, "Guided genome halving: hardness, heuristics and the history of the hemiascomycetes," *Bioinformatics*, vol. 24, no. 13, 2008. 9

[79] D. Sankoff, C. Zheng, P. Wall, C. dePamphilis, J. Leebens-Mack, and V. Albert, "Towards improved reconstruction of ancestral gene order in angiosperm phylogeny," *Journal of Computational Biology*, vol. 16, no. 10, pp. 1353–1367, 2009. 9

[80] M. et al, "Dupcar: reconstructing contiguous ancestral regions with duplications," *J Comput Biol*, vol. 8, pp. 1007–1027, Oct 2008. 9, 51, 67

[81] G. Tesler, "Grimm: genome rearrangements web server," *Bioinformatics*, vol. 18, Mar 2002. 10.1093/bioinformatics/18.3.492. 9

[82] G. Bourque and P. Pevzner, "Genome-scale evolution: Reconstructing gene orders in the ancestral species," *Genome Research*, vol. 12, pp. 26–36, 2002. 9

[83] B. Moret, A. Siepel, J. Tang, and T. Liu, "Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data," *Proc. 2nd Int'l Workshop on Algorithms in Bioinformatics (WABI'02)*, 2002. 9, 50, 56

[84] H. Luo, J. Shi, W. Arndt, J. Tang, and R. Friedman, "Gene order phylogeny of the genus prochlorococcus," *PLoS ONE*, vol. 3, no. 12, 2008. 9

[85] E. Kejnovsky, J. Leitch, and R. Leitch, "Contrasting evolutionary modes between angiosperm and mammalian genomes," *Trends in Ecology and Evolution*, vol. 24, no. 10, 2009. 10

[86] C. Simillion, K. Vandepoele, Y. Saeys, and Y. van de Peer, "Building genomic profiles for uncovering segmental homology in the twilight zone," *Genome Research*, vol. 14, Jun 2004. 10.1101/gr.2179004. 10, 54, 72, 73, 74, 80, 82, 85, 93, 94, 98, 102

[87] P. Calabrese, S. Chakravarty, and T. J. Vision, "Fast identification and statistical evaluation of segmental homologies in comparative maps," *Bioinformatics*, vol. 19, pp. 74–80, 2003. 10, 30, 51, 54, 72, 73

[88] L. et al, "Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape: Coge with rosids," *Plant Physiology*, vol. 148, pp. 1772–1781, 2008. 10

[89] S. Hampson, P. Baldi, and B. Gaut, "Closeup: Statistical detection of chromosomal homology using density alone  a comparative analysis," *Bioinformatics*, vol. 21, pp. 1339–1348, 2005. 10, 72, 73

[90] J. Bennetzen and M. Chen, "Grass genomic synteny illuminates plant genome function and evolution," *Rice*, vol. 1, pp. 109–118, 2008. 10, 16

[91] M. Kohn and H. Pelz, "A gene-anchored map position of the rat warfarin-resistance locus, rw, and its orthologs in mice and humans," *Blood*, vol. 96, pp. 1996–1998, 2000. 11

[92] S. Chien, L. Reiter, E. Bier, and M. Gribskov, "Homophila: human disease gene cognates in drosophila," *Nucleic Acids Research*, vol. 30, no. 1, pp. 149–151. 11

[93] C. Potter, G. Turenchalk, and T. Xu, "Drosophila in cancer research. an expanding role," *Trends in Genetics*, vol. 16, no. 1, pp. 33–39, 2000. 11

[94] P. Leopold and N. Perrimon, "Drosophila and the genetics of the internal milieu," *Nature*, vol. 450, pp. 186–188, 2007. 11

[95] T. et al, "Conservation of synteny between the genome of the pufferfish (fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial alzheimer disease (ad3 locus)," *PProceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 4, 1996. 11

[96] K. Oh, K. Hardeman, M. G. Ivanchenko, M. Ellard-Ivey, A. Nebenfuhr, T. J. White, and T. L. Lomax, "Fine mapping in tomato using microsynteny with the arabidopsis genome: the diageotropica (dgt) locus," *Genome Biology*, 2002. 11, 16

[97] F. Gallego, C. Feuillet, M. Messmer, A. Penger, A. Graner, M. Yano, T. Sasaki, and B. Keller, "Comparative mapping of the two wheat leaf rust resistance loci lr1 and lr10 in rice and barley," *Genome*, vol. 41, no. 3, pp. 328–336, 1998. 11, 16

[98] R. Tuberosa, S. Giuliani, M. Parry, and J. Araus, "Improving water use efficiency in mediterranean agriculture: what limits the adoption of new technologies?," *Annals of Applied Biology*, vol. 150, no. 2, pp. 157–162. 11, 16

[99] R. Grube, E. Radawanski, and M. Jahn, "Comparative genetics of disease resistance within the solanaceae," *Genetics*, vol. 155, pp. 873–887, 2000. 11, 16

[100] S. et al, "Genome evolution in mushrooms: Insights from the genome and assembled chromosomes of coprinopsis cinerea (coprinus cinereus)," *Proceedings of the National Academy of Sciences of the United States of America*, 2010. 12

[101] A. H. Paterson, J. E. Bowers, and B. A. Chapman, "Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 9903–9908, Jun 2004. 10.1073/pnas.0307901101. 16

[102] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae," *Nature*, vol. 428, pp. 617–624, Apr 2004. 10.1038/nature02424. 16

[103] S. Jackson, S. Rounsley, and M. Purugganan, "Comparative sequencing of plant genomes: Choices to make," *The Plant Cell*, vol. 18, pp. 1100–1104, 2006. 16

[104] V. Benedito, "Time to crop: jumping from biological models to crop biotechnology," *Crop Breeding and Applied Biotechnology*, vol. 7, pp. 1–10, 2007. 16

[105] S. et al., "A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics," *Theoretical Applied Genetics*, vol. 114, pp. 823–829, 2007. 16

[106] M. Lynch and J. S. Conery, "The Evolutionary Fate and Consequences of Duplicate Genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000. 17, 82

[107] A. Paterson, J. Bowers, D. Paterson, J. Estill, and B. Chapman, "Structure and evolution of cereal genomes," *Current Opinion in Genetics & Development*, vol. 13, pp. 644–650, 2003. 17

[108] J. Demuth, T. De Bie, J. Stajich, N. Cristianini, and M. Hahn, "The evolution of mammalian gene families," *PLoS One*, vol. 1, p. e85, 2006. 17

[109] M. Hahn, J. Demuth, and S. Han, "Accelerated rate of gene gain and loss in primates," *Genetics*, vol. 177, pp. 1941–1949, 2007. 17

[110] O. Zhaxybayeva, C. Nesbo, and W. F. Doolittle, "Systematic overestimation of gene gain through false diagnosis of gene absence," *Genome Biology*, vol. 8, no. 2, p. 402, 2007. 17, 18

[111] M. Hahn, T. De Bie, J. Stajich, C. Nguyen, and N. Cristianini, "Estimating the tempo and mode of gene family evolution from comparative genomic data," *Genome Research*, vol. 15, pp. 1153–1160, 2005. 17, 18

[112] S. Maere, S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer, "Modeling gene and genome duplications in eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 5454–5459, Apr 2005. 10.1073/pnas.0501102102. 18, 67, 83, 84, 97

[113] D. Gilkes, W.Rand Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996. 27, 28

[114] S. M. Hedtke, T. M. Townsend, and D. M. Hillis, "Resolution of phylogenetic conflict in large data sets by increased taxon sampling," *Systematic Biology*, vol. 55, no. 3, pp. 522–529, 2006. 31

[115] D. R. Scannell, A. C. Frank, G. C. Conant, K. P. Byrne, M. Woolfit, and K. H. Wolfe, "Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication," *Proceedings of the National Academy of Sciences*, vol. 104, no. 20, pp. 8397–8402, 2007. 31, 46, 84, 101

[116] J. Raes, K. Vandepoele, C. Simillion, Y. Saeys, and Y. Van De Peer, "Investigating ancient duplication events in the arabidopsis genome," *Journal of Structural and Functional Genomics*, vol. 3, pp. 117–129, 2003. 32

[117] K. Vandepoele, C. Simillion, and K. Van de Peer, "Evidence that rice and other cereals are aneuploids," *The Plant Cell*, vol. 15, pp. 2192–2202, 2003. 32, 74, 82

[118] M. Freeling and B. Thomas, "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity," *Genome Research*, vol. 16, pp. 805–814, 2006. 44, 98

[119] C. Cunningham, K. Omland, and T. Oakley, "Reconstructing ancestral character states: a critical reappraisal," *Trends in Ecology and Evolution*, vol. 13, no. 9, pp. 361–366. 48

[120] D. Swofford and W. Maddison, *Parsimony, character-state reconstruction and evolutionary inferences*, pp. 186–223. Stanford University Press. 48

[121] M. Pagel, "Inferring evolutionary processes from phylogenies," *Zool. Sc.*, vol. 26, pp. 331–348. 48

[122] G. e. a. Rubin, "Comparative genomics of the eukaryotes," *Science*, vol. 287, pp. 2204–2215, 2000. 48

[123] A. e. a. Paterson, "Comparative genomics of plant chromosomes," *THE PLANT CELL*, vol. 12, pp. 1523–1540, Sep 2000. 10.1105/tpc.12.9.1523. 49

[124] G. Blanc, K. Hokamp, and K. Wolfe, "A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome," *Genome Research*, vol. 13, pp. 137–144, 2003. 49

[125] M. Zhang, W. Arndt, and J. Tang, "An exact solver for the dcj median problem," *Proceedings of the Pacific Symposium on Biocomputing*, vol. 14, pp. 138–149, 2009. 51, 52, 67

[126] "The abcs of mgr with dcj — libertas academica." 51

[127] D. Sankoff, C. Zheng, C. dePamphilis, J. Leebens-Mack, A. V. Albert, and P. Wall, *Comparative Genomics*. Springer, 2008. 51

[128] S. Yancopoulos and R. Friedberg, *Sorting Genomes with Insertions, Deletions and Duplications by DCJ*. Springer, 2008. 52, 69

[129] N. Raghupathy and D. Durand, "Gene Cluster Statistics with Gene Families," *Mol Biol Evol*, vol. 26, no. 5, pp. 957–968, 2009. 56, 57, 66

[130] D. Sankoff and L. Haque, "The distribution of genomic distance between random genomes," *Journal of Computational Biology*, vol. 13, no. 5, pp. 1005–1012, 2006. 57, 66

[131] L. Szekely and Y. Yang, "On the expectation and variance of the reversal distance," *Acta Univ. Sapientiae, Mathematica*, vol. 1, no. 1, pp. 5–20, 2009. 57, 66

[132] M. Steel and A. McKenzie, *The 'shape' of phylogenies under simple random speciation models*. Springer Berlin/Heidelberg, 2002. 60

[133] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, 1977. 60

[134] L. et al., "Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in triticeae," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 37, pp. 15780–15785, 2009. 67, 82, 83, 84, 97

[135] K. Yogeeswaran, A. Frary, T. York, A. Amenta, A. Lesser, J. Nasrallah, S. Tanksley, and M. Nasrallah, "Comparative genome analyses of arabidopsis spp.:inferring chromosomal rearrangement eventsin the evolutionary history of a. thaliana," *Genome Research*, vol. 15, pp. 505–515, 2005. 67, 82, 83, 84

[136] H. Zhao and G. Bourque, "Recovering genome rearrangements in the mammalian phylogeny," *Genome Research*, vol. 19, pp. 934–942, 2009. 67

[137] D. Sankoff, C. Zheng, P. Wall, C. DePamphilis, J. Leebens-Mack, and V. Albert, "Towards improved reconstruction of ancestral gene order in angiosperms," *Journal of Computational Biology*, vol. 16, no. 10, pp. 1353–1367, 2009. 68, 69

[138] J. Salse and C. Feuillet, *Comparative Genomics of Cereals*, pp. 177–205. Springer Netherlands, 2007. 72

[139] J. Flowers and M. Purugganan, "The evolution of plant genomes scaling up from a population perspective," *Current Opinion in Genetics and Development*, vol. 18, pp. 565–570, 2008. 72

[140] M. e. a. Brudno, "Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna," *Genome Research*, vol. 4, pp. 21–31, 2003. 72

[141] M. Lynch, *Genomic Expansion by Gene Duplication*. Sinauer Associates, Inc. Publishers, 2007. 73

[142] J. et al, "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, no. 7161, pp. 463–46, 2007. 73

[143] Y. Van de Peer, J. A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele, "The flowering world: a tale of duplications," *Trends in Plant Science*, vol. 14, no. 12, 2009. 74

[144] S. Nee, "Inferring speciation rates from phylogenies," *Evolution*, vol. 55, no. 4, pp. 661–668, 2001. 84

[145] M. Semon and K. Wolfe, "Rearrangement rate following the whole-genome duplication in teleosts," *Molecular Biology and Evolution*, pp. 860–867, 2007. 99