Visualization Tools for Comparative Genomics applied to Convergent Evolution in Ash Trees

Josiah David Seaman

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

School of Biological and Chemical Sciences
Queen Mary University of London

Supervisor: Prof. Richard Buggs

December 2020

Statement of Originality

I, Josiah David Seaman, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Josiah David Seaman
Date: December 3, 2020

Details of collaboration and publications:

This research was partially funded under grant BBSRC BB/S004661/1.
Collaborations and co-authors are detailed at the beginning of their respective chapters.

Thesis Abstract

Assembly and analysis of whole genomes is now a routine part of genetic research, but effective tools for the visualization of whole genomes and their alignments are few. Here we present two approaches to allow such visualizations to be done in an efficient and user-friendly manner. These allow researchers to spot problems and patterns in their data and present them effectively.

First, FluentDNA is developed to tackle single full genome visualization and assembly tasks by representing nucleotides as colored pixels in a zooming interface. This enables users to identify features without relying on algorithmic annotation. FluentDNA also supports visualizing pairwise alignments of well-assembled whole genomes from chromosome to nucleotide resolution.

Second, Pantograph is developed to tackle the problem of visualizing variation among large numbers of whole genome sequences. This uses a graph genome approach, which addresses many of the technical challenges of whole genome multiple sequence alignments by representing aligned sequences as nodes which can be shared by many individuals. Pantograph is capable of scaling to thousands of individuals and is applied to SARS and *A. thaliana* pangenomes.

Alongside the development of these new genomics tools, comparative genomic research was undertaken on worldwide species of ash trees. I assembled 13 ash genomes and used FluentDNA to quality check the results and discovered contaminants and a mitochondrial integration. I annotated protein coding genes in 28 ash assemblies and aligned their gene families. Using phylogenetic analysis, I identified gene duplications that likely occurred in an ancient whole genome duplication shared by all ash species. I examined the fate of these duplicated genes, showing that losses are concentrated in a subset of gene families more often than predicted by a null model simulation. I conclude that convergent evolution has occurred in the loss and retention of duplicated genes in different ash species.

Acknowledgements

I have come as far as I have by involving and drawing strength from many exceptional and wonderful people and now I'd like to give them all thanks. To the greatest endurer and strong pillar in my life, my wife Sam Seaman, you have my deepest gratitude for coming with me on this long journey across continents and decades. I want to extend a great thank you to my advisor Richard Buggs who has stayed with me for four years and taught me much about academic thinking and writing. Thank you for walking with me through the gardens to discuss experimental designs during lunches. Thank you for reading all the versions of my thesis and giving me invaluable feedback. Thank you for constructing the lab and environment that enabled the research to happen.

I am deeply grateful to my parents, Mark and Stacie Seaman who have not only fulfilled the classical roles of supportive parents but have also been tremendously helpful in the battlefield roles as project managers during the pandemic. Thank you as well to Michael Chester and Ilia Lietch whose constant encouragement and fascinating conversations kept me interested in plant genomics and connected with the inner workings of Kew Gardens. I am indebted to Richard Nichols who provided excellent detailed feedback on my thesis chapters. He hosted weekly lab chats and was a constant source of statistical and practical wisdom. Thanks to Endymion Cooper for igniting my interest in paleopolyploidy and for charting the project that led to Chapter 4. Thank you to Laura Kelly and Carey Metheringham who were knowledgeable and helpful for discussion about *Fraxinus*. Thank you to Jonathan Wendel for reviewing Chapter 1 and giving me the confidence that one really can survey a field and understand it at some point. The stamp of approval of a giant in the field meant the world to me.

I'm very grateful to Yan Wong for all the fascinating discussions and for the tour of Oxford Big Data Institute. Your customization of space filling curves broke through a roadblock I had with them since 2009. Thank you to Tomasz Neugebauer and Bryan Hurst for help with the start of FluentDNA development. I'm thankful that Seth Richardson and Astrid Krumins were working at Kew Gardens with me because it was a pleasure working with them in outreach and education using FluentDNA. Thank you to Paul Kersey, Detlef Weigel, and Computomics for providing me with the intellectual freedom and three months of time to dream up the designs for pangenome schematics.

To the Pantograph Team, I owe so much for all the hours of camaraderie and brainstorming as we worked to develop the future and got to know each other in the process. I especially want to thank Simon, Erik, Toshiyuki, and Torsten for being there to kickstart the project. Thank you to Thomas Townsley, Christian Kubica, and Peter Amstutz for running support on datasets. I'm very thankful to have the privilege to have met Ben Busby, Pjotr Prins, Jerven Bolleman, Toshiaki Katayama, and Erik Garrison who made the 2019 Biohackathon a life changing event for me and invigorated my vision for academia when I needed it the most. I owe my warmest thanks to my London friends Rob, Victoria, Sarah, Ted, Matthew, Rogers, and Rosie for making London the longest place I've called home since leaving my childhood home in 2003. I will always have fond memories of my time here and I will treasure them for the rest of my life.

Contents

List of Figures

List of Tables

Chapter 1
Introduction to Polyploid Evolution

Author Contribution

I wrote the entire text of this chapter. Special thanks to Jonathan Wendel for reviewing an earlier draft of this chapter and references.

Abstract

1. Gene duplication plays a critical role in evolution because it provides the raw materials necessary to explore sequence space by allowing one or both gene copies to mutate down different paths (neofunctionalization and subfunctionalization). Without the branching options provided by duplication, genes necessary for cell survival would likely become trapped in local optima.
2. In plants, individuals with more genome copies than their parents (polyploids) are pervasive but most newly-formed polyploids do not survive. Polyploidy can have profound effects on phenotype, reproductive isolation, and evolutionary trajectory. Polyploids are frequently crosses of parents who cannot produce viable diploid offspring, creating hybrids with novel combinations of genes and gene regulatory sequences. This can sometimes include gene expression ranges outside those found in either parent (transgressive expression). Such novelties can be important to plant breeders looking for large changes in quantity or size of a trait as well as the potential for local adaptation. Polyploidization has many evolutionary consequences ranging from reproductive isolation (minority cytotype exclusion) and transcriptomic shock, but also a possibility of expanding into new ecosystems. The wide range of consequences means there is still much about polyploidy to be researched.
3. Sequence evidence across a wide range of plants suggests rounds of ancient duplications (paleopolyploidy) followed by gene copy loss (fractionation). From this we infer that all extant species of seed plants are the descendants of rare successful polyploids that overcame physiological and reproductive challenges associated with polyploidy. Whole genome duplication (WGD) represents a wide scale disruption to gene regulatory network dynamics, but at the same time may be more survivable than a partial genome duplication because all relative ratios are preserved and cytoplasmic volume is increased. The Boom and Decay model explains long term genome size staying relatively stable despite rounds of duplications, due to loss of the majority of duplicated genes.
4. Loss of duplicated genes after whole genome duplications could be random, or biased by biological factors which influence whether gene copies are retained versus lost during fractionation. These factors may include the toxic effects of incorrect absolute dosage of certain gene products and the stoichiometric constraints of protein complexes whose dosage is determined by the correct interaction and binding of all parts. It is also posited that early Biased Expression from subgenome dominance strongly influences long term retention of some duplicated genes. There may also be a “Lag Time” after WGD, where many homeologs (copies created in a WGD) are not lost for millions of years.
5. Studies of *Fraxinus* and *Olea* can serve as a model for testing these hypotheses because *Fraxinus* and *Olea* share a WGD at 26 million years ago (Mya) and *Olea* has a second WGD. The 28 species sequenced in the worldwide *Fraxinus* genome project

gives us a unique opportunity to test the reproducibility of fractionation outcomes and expand our understanding of whole genome duplication dynamics.

1.1 Gene Duplication in Evolution

1.1.1 Homologs and Paralogs

Plant and animal genomes show extensive evidence of duplications of coding and non-coding genetic material. The history of genes within an organism's genome can be understood through the lens of a history of duplications. When two genes are related through a duplication event, they are called paralogs and can occur in different positions within the same genome or across species (Thornton and DeSalle 2000). When two genes are related through a speciation event, they are called homologs and these often continue to maintain the same function. This history of evolution of one gene can be represented using a gene tree where branches in the tree represent either speciation events or duplications. The bioinformatic methods for identifying homologs are discussed in 3.2.5. Speciation is a rare event requiring a sustainable breeding population, but gene duplications can be specific to a single cell.

1.1.2 Sequence Space

Evolution is a gradual two step process requiring 1) the creation of random genetic diversity through mutation and 2) the elimination of genetic diversity through selective survival and breeding. In order to model similar gene sequences in a mathematically robust way it is useful to view two sequences as points in a high dimensional space. For example, a typical CDS of 900 nucleotides (Brocchieri and Karlin 2005) would be 'edit distance' of one from a homologous duplicate sequence with a single nucleotide changed. These two sequences could be considered adjacent neighbors in sequence space because a single change moves from one state to the other. Each nucleotide is a degree of freedom in this high dimensional space meaning that the typical 900 nucleotide CDS has 2,700 (900 * (4-1)) adjacent neighboring sequences that differ by only one nucleotide.

As a mathematical model, sequence space has properties which can be studied. The number of possible variations on a sequence increases exponentially with an increasing number of nucleotide changes from the base sequence. The number of possible sequences is roughly $(3N)^k$ where N is the base sequence length and k is the number of single nucleotide changes. The vast majority of sequence changes are neutral or nearly neutral and indetectable to selection (Kimura 1979).

Stable folding proteins are made from secondary structures called alpha-helices and beta-sheets (Bungard et al. 2017). These secondary structures are only formed from sequence motifs which make up a minority of sequence space (Carvunis et al. 2012), thus the majority of sequence space will not form stable protein folds (Nartey et al. 2017). *De novo* gene evolution is the term given for these rare instances where DNA which has not been selected as an amino acid sequence begins producing an expressed protein. Examples of *de novo* gene birth that are not rescued pseudo-genes are so rare that it's difficult to verify their existence. Bungard et al. (2017) advance Bsc4 as one such example and point to its apparent function despite lack of a stable quaternary fold as an intermediary evolutionary step to a stable globular protein. Typically, mutations accumulate in extant stable proteins. While mutations may be random, the starting sequence that each new generation mutates is, by definition, adjacent to a functional sequence that its parent used to survive.

Duplication of protein coding sequences provides branches along gene trees that allow new evolutionary possibilities that have the advantage of starting with a functional sequence and stable protein folds. Duplication allows genes to escape selection pressure to explore adjacent sequence space while one copy maintains survival critical functions. Sequence duplication may contribute to adaptation. For example, Faba bean necrotic stunt virus uses gene copy number changes to adapt while still maintaining small genomes (Sicard et al. 2016).

1.1.3 Subfunctionalization and Neofunctionalization

The most common outcome of duplicated sequence is for the duplicate to be lost (Lynch and Conery 2003). For duplicates that are retained, there are two possible routes to divergent functions. When a generalist ancestor protein has duplicates that each evolve to only carry out a subset of the ancestor's function, this is called subfunctionalization (Konrad et al. 2011). For example, given an enzymatic protein with several substrates or products, it is possible for different copies to lose enzymatic activity in different subsets of substrates or products: this subfunctionalization leaves behind two proteins with more specialized functions.

The second possible outcome of gene duplication is evolving a function not present in the ancestor, called neofunctionalization (Ohno 1970). Neofunctionalization can mean the development of new and more nuanced morphological features, as in the well studied case of flower morphology development through gene duplication. APETELA3 is one such gene family whose diverged copies control stamen identity in the flower. Duplications can lead to new petal parts (Kramer et al. 2003). Neofunctionalization can also be an enzymatic activity in a new substrate or higher catalytic activity (Mindrebo et al. 2016). Aharoni et al. were able to artificially evolve two protein variants that had 40-fold increase in hydrolizing activity and 2,000-fold increase in specificity (Aharoni et al. 2004). Neofunctionalization can also occur in the form of a transcription factor binding to a new sequence or protein (Teufel et al. 2018); for example, a duplicated KRAB zinc-finger protein with a higher binding affinity for a new transposon subfamily (Ecco, Imbeault, and Trono 2017; Imbeault, Helleboid, and Trono 2017).

This ongoing process of duplication and subsequent divergence is thought to have left its mark in genomes in the form of large gene families with varying sequences to support nuanced function. For example, the body plans of animals are determined by tandem repetitions of 39 HOX genes whose sequence variations allow them to specify different sections of the body plan (Kuraku 2011).

1.2 Polyploidy

Whole genome duplications (WGD) commonly occur in plants (Spoelhof et al. 2017; Adams and Wendel 2005; Ramsey and Schemske 1998). These events produce polyploid offspring with more genome copies than either parent and can frequently facilitate hybrid crosses that are unviable in diploid progeny, or sterile due to failure of chromosome pairing; these are sometimes called wide hybrids (Dodsworth, Chase, and Leitch 2016). Hybrid polyploids are known as allopolyploids and are often viable because chromosomes can pair in meiosis (Darlington et al. 1937).

In plant pest and pathogen systems, we are presented with an evolutionary paradox. If tree generation times are orders of magnitude slower than insect and fungal generational times how is it that trees have been able to keep up in the evolutionary arms race against pests and pathogens over millions of years? Wininger and Rank (2017) in a review of 117 studies found that plants primarily adapt to pathogens through gene for gene interactions. Whereas, plants predominantly adapt to herbivores through escape and radiate responses.

Whole genome duplication creates additional copies relevant to gene-for-gene adaptation races through subfunctionalization (1.1.3). Furthermore, WGD have historically been associated with increased radiation into new niches in times of changes such as the Cretaceous-Tertiary extinction event (Fawcett et al. 2009). We explore the evolutionary dynamics of plant whole genome duplications here.

1.2.1 Endopolyploidy

Endopolyploidy is a tissue specific polyploidy found widely in multicellular organisms where the ploidy level of a particular tissue is a multiple of the base ploidy level of the gametes. For example 60% of *Arabidopsis* tissue is endopolyploid (Galbraith, Harkins, and Knapp 1991), meaning that the tetraploid state is already exposed to selection pressure every generation (I. J. Leitch and Dodsworth 2001). Endopolyploidy also exists in animal tissue with protein production functions such as nourishment of embryos, secretion, and platelet formation (Neiman et al. 2017; Cross 2005; Flemming et al. 2000; Ravid et al. 2002). Endopolyploidy is achieved through modifications of the cell cycle and can result in polytene chromosomes that do not fully separate. Other endo-duplications can occur through partial amplification of a section of chromosome, rather than the entire genome (Lee, Davidson, and Duronio 2009). Endopolyploidy frequently occurs in cancers (Storchova and Pellman 2004).

Endopolyploidy means the polyploid state is already under selection pressure every generation. At a tissue level, the polyploid state must be a viable state for gene regulatory machinery to function in any organism that makes extensive use of endopolyploidy such as *A. thaliana*. Endopolyploidy then acts as a selectable intermediary to a viable polyploid organism. Since random changes to one gene are more likely to cause nonfunctionalization than neofunctionalization, statistically, making random changes to every gene in an organism is certain to be deleterious (Bataillon 2000; Belancio, Roy-Engel, and Deininger 2010; Böndel et al. 2019). This is because any extant organism occupies a small area of sequence space (1.1.2) that has been repeatedly tested to be more fit than every adjacent variant. The same is true of gene regulatory networks. However, polyploidy is not a random change. It is merely switching from one gene regulatory regime under selection (diploidy) in one tissue to another regulatory regime which was already under selection in the same organism albeit in a different tissue (endopolyploidy). While this change can cause large disruptions (1.2.2.2, 1.2.2.3, 1.2.2.4, 1.2.2.5) it is in no way comparable to the lethal genome rearrangements caused by high doses of ionizing radiation or aneuploid cancer lines like K562 (Zhou et al. 2018).

1.2.2 Immediate Consequences Of WGD

1.2.2.1 Minority Cytotype Exclusion

One common immediate consequence of polyploidy is reproductive isolation from the parental species. This may be advantageous as it frees a polyploid to evolve in new directions without gene flow from its parents (Pannell, Obbard, and Buggs 2004). On the other hand, it produces an immediate reproductive disadvantage and the possibility of producing no fertile offspring; this is called minority cytotype exclusion (Levin 1975). In order to establish a breeding population, a polyploid must overcome this barrier.

1.2.2.2 Phenotypic Effect of Genome Size

Polyploidy may cause immediate phenotypic novelties. Whole genome duplication may directly affect cell size, cytoplasmic volume versus surface area, and changes to the transcriptome and molecular transport due to the increased distances between cell parts. These affects in autopolyploids (polyploids which are not hybrids) of *Glycine, Arabidopsis,* and *Solanum* are reviewed in Doyle and Coate (2019). Phenotypic novelties may mean that new polyploids can occupy a new niche. There are many examples of new polyploids colonizing new ecosystems (Beest et al. 2012). Polyploid invasiveness was first proposed by Levin (1983) and explained by increased robustness e.g. drought and salinity tolerance, however, the cause is still debated.

1.2.2.3 Expression Changes

Another immediate consequence of polyploidy may be changes to gene expression regulation. When these changes are immediate and widespread it is called transcriptomic shock (Hegarty et al. 2006; Arrigo and Barker 2012; Yoo, Szadkowski, and Wendel 2013; Buggs et al. 2011). New polyploids may have numerous novel regulatory interactions that are beneficial or deleterious to the organism. These come from both gene copy number changes, and the interactions of two genomes which have diverged significantly in the case of allopolyploids (Kashkush, Feldman, and Levy 2002).

1.2.2.4 Hybridity and Parent Subgenome Dominance

In some allopolyploids, hybridity is the most important factor in gene expression changes. This seems to be the case in *Arabidopsis, Gossypium,* and *Brassica* allopolyploids but less so in maize and *Senecio* (Doyle et al. 2008). When a hybrid is formed, the gene expression and traits are not a 50/50 mix of the two parents. In allopolyploids, a bias towards the expression of one parent's genes over the other, called Parent Subgenome Dominance, is often observed (Emery et al. 2018). This dominance appears to be linked to gene regulatory machinery, where one subgenome has more positive feedback regulation or more completely silences the other subgenome. Intriguingly, this dominance can be tissue type specific as in the case of cotton which expresses different subgenomes per tissue type (Hovav et al. 2015). One can easily see how this could contribute to morphological complexity, though not necessarily adaptation, by giving the plant different copies of expression machinery per tissue, each which can continue to evolve complementary specializations for their given tissues.

1.2.2.5 Genome Instability

New polyploids often have severe genome instability resulting from unstable connections between homeologous chromosomes during meiosis. During cell division (meiosis or mitosis) matching chromosomes pair up across from each other along the fission plane and are separate by the centrioles. This delicate process can be complicated by the presence of homeologous chromosomes in polyploids which may imperfectly match if paired during crossing over of meiosis. A study of the 100-year-old neopolyploid *Tragopogon miscellus* revealed substantial fluctuations in the content of each chromosome between individuals (Buggs et al. 2012; Chester et al. 2012). Similar results can be seen in the 140-year-old allopolyploid *Senecio cambrensis* (Hegarty, Abbott, and

Hiscock 2012). In 2005 (Salmon, Ainouche, and Wendel 2005) found two independent hybrids of *Spartina* have undergone the same sequence region losses in the last 165 years. Similarly repeatable losses have been observed in wheat F1 hybrids (Shaked et al. 2001).

1.2.2.6 Transposon Instability in Neopolyploids

Transposons can likely explain this rapid repeatable gene loss in allopolyploids. Transposon silencing mechanisms can be disrupted in a hybrid which can lead to rapid sequence loss (Hawkins et al. 2009). In rice (*Oryza sativa*) LTR activity has been observed as a precursor of rapid DNA loss (Vitte, Panaud, and Quesneville 2007). Transposons create many similar sequences throughout the genome. This can lead to non-homologous gene conversion where one copy of a sequence is stamped over another similar copy, eliminating sequence diversity even in sites that are not alleles paired up during meiosis (Ellison and Bachtrog 2015).

Parental subgenome dominance information is often not available for very old polyploids because parental sequences are not available. In recent hybrids such as *Tragopogon*, *Senecio*, *Spartina*, and Cotton determining which sections of sequence belong to which subgenome is fairly straight forward because their parental population are still living. As we look at older polyploids, assignment becomes more difficult, especially if the parental populations have gone extinct. Phylogenetic analysis requires sequence from one of the parental populations in order to assign subgenomes (Edger et al. 2018). Without this, claims of subgenome dominance become circular reasoning since silenced DNA and tissue specific expression are also observed in species lacking a history of WGD, such as humans. From a comparison with the sea lamprey genome, it appears the human lineage last underwent a WGD 550 mya at the base of the vertebrate lineage (Smith and Keinath 2015).

1.2.3 Longer Term Evolutionary Consequences

After WGD, most duplicated sequences are expected to be hidden from selection and gradually lost as mutations accumulate. In order for a gene to be retained in the long term, there needs to be selective pressure in favor of its retention. The process by which most of the genome gradually returns to a diploid (diploidization) whilst some regions remain in a duplicated state is known as fractionation. William Bateson first used the term “fractionation of factors” in 1915 to refer to a similar distillation of phenotypic traits in successive generations of inheritance, before the discovery of DNA or genes, making it possibly the oldest usage of the term fractionation of an organism through evolution (British Association for the Advancement of Science 1915). Langham et al. (2004) used fractionation specifically for diploidization to refer to how the fraction of genes under some criteria will be preserved in duplicate. Which genes are lost and which are retained depends upon a number of factors, which are still an active area of research (L. Flagel et al. 2008).

Table 1.1 The Link between Biased Fractionation and Subgenome Dominance

Taxon or taxa	Approximate WGD age (mya)	Biased fractionation?	Genome dominance?	Reference
Arabidopsis suecica	0.02	Yes	Yes	(Chang et al. 2010; Novikova et al. 2017)
Capsella bursa-pastoris	0.2	No	No	(Douglas et al. 2015)
Zea mays	8	Yes	Yes	(Schnable, Springer, and Freeling 2011; Swigoňová et al. 2004)
Glycine max	13	No	No	(Garsmeur et al. 2014)
Cucurbita spp.	3–26	No	No	(Sun et al. 2017)
Brassica rapa	15	Yes	Yes	(Cheng et al. 2016; Mandáková et al. 2017)
Arabidopsis thaliana	47	Yes	Yes	(Garsmeur et al. 2014; Thomas, Pedersen, and Freeling 2006; Hohmann et al. 2015)
Medicago sativa	58	Yes	N/A	(Garsmeur et al. 2014)
Gossypium spp.	60	Yes	Yes	(Renny-Byfield et al. 2015)
Musa acuminata	65	No	No	(Garsmeur et al. 2014; D'Hont et al. 2012)
Populus trichocarpa	65	No	No	(Garsmeur et al. 2014)
Poaceae	70	Yes	Yes	(Garsmeur et al. 2014)

Biased fractionation occurs when parental subgenome dominance leads dominant genes from one parent to be preferentially retained in descendants (Bottani et al. 2018). In the 12 species reviewed, parental subgenome dominance and biased fractionation always occurred together (yellow highlighting). Five of the 12 species showed no biased fractionation nor subgenome dominance. No correlation with WGD age and biased fractionation is evident. **Source**: Table 1.1 used with permission from Wendel et al. (2018).

It is posited this Subgenome Dominance is due to a difference in transposable element (TE) load and proximity to genes. TEs are targeted for silencing using small RNA silencing machinery and histone modifications which can affect neighboring genes (Vicient and Casacuberta 2017). A primary mechanism of DNA silencing is methylation, which can lead to a region becoming closed chromatin. This is consistent with the idea that Subgenome Dominance is driven by TE load and position. From *Tragopogon*, there is some evidence that dominance appears to be intrinsic to a parent genome. Genes from *T. pratensis* were preferentially retained over *T. dubius* genes, regardless of which was the maternal or paternal genome (Buggs et al. 2012). In maize, Woodhouse et al. 2010 found that biased fractionation occurs mainly through single gene loss in regions flanked by evidence of intra-chromosomal recombination. The subgenome with more losses also had more targets for transposon removal (Woodhouse et al. 2010). From all this we can potentially predict that if two genomes diverge in TE content and then produce a hybrid, the one with fewer TEs will have dominant expression and is more likely to retain genes in its descendants.

1.2.3.1 Evolutionary Advantages

There has been much conversation about the mix of advantages and disadvantages of polyploidy and the long term evolutionary consequences of duplication (Ohno 1970; L. E.

Flagel and Wendel 2009; I. J. Leitch and Leitch 2013; Douglas E. Soltis et al. 2014; Mayrose et al. 2011). Here we discuss several important examples.

Freeling and Thomas propose that preferential retention of transcription factors (TFs) and developmental genes after a WGD is sufficient to explain a predictable increase in morphological complexity in plants (Freeling and Thomas 2006). We note that in order for this to be true, the theory requires that: 1) morphological complexity actually requires subfunctionalization of TFs, 2) complexity results in higher fitness 3) fitness gains are large enough to offset the disadvantages of large, duplicated genomes.

It may be that patterns of fractionation could lead to adaptation. A polyploid's wide range of redundant genes could be silenced in order to rapidly adapt to new environments without the need to wait for nucleotide changes (Ohno 1970; Werth and Windham 1991). This rapid adaptation ability is seen even in polyploids of yeast (Selmecki et al. 2015). In theory, short term silencing through methylation after a WGD then removes selection pressure to maintain the gene copy which can then be lost in subsequent generations (1.2.2.4) (De Smet et al. 2013; Woodhouse et al. 2010, 2014). Angiosperms become polyploid far more frequently than gymnosperms which has possibly driven the increased number of species in angiosperms. This potential for rapid adaptation is postulated as the reason angiosperms dominated niches previously filled by gymnosperms in the ancient past (A. R. Leitch and Leitch 2012). However, two recent reviews concluded that WGDs themselves were not associated with higher species diversification in angiosperms at present (Landis et al. 2018; Mayrose et al. 2011).

1.3 Evidence for Paleopolyploidy

Given the extensive effects of polyploidy in modern day organisms, how extensively has polyploidy affected the ancestors of extant species? Here we look at the evidence for ancient polyploids (paleopolyploidy) and the challenges in analysis. As we have seen, new polyploids frequently undergo genome instability (Buggs et al. 2012) on the scale of 100 years. The artificial autotetraploid *Phlox drummondii* has been observed to lose 25% of its genome in the first two generations as the karyotype stabilizes with concomitant increase in seed viability. No further changes were observed in subsequent generations (Raina et al. 1994). Naturally, this means ancient evidence of paleopolyploidy does not include a complete copy of a genome. Instead, many duplicates which share the same age (Ks and phylogenetic analysis) or blocks of duplicated sequence in the same order (shared synteny) are the primary sources of evidence.

1.3.1 Ks

The level of sequence divergence in alignable regions is measured by the frequency of synonymous codon mutations, called Ks. Ks is used as an estimate of long ago the duplicate was created. A peak in the Ks plot is evidence that all the duplicates came from the same event (Jiao et al. 2011). A peak in the Ks plot is a clear indication of a WGD (Figure 1.1) in the last 80My given only a single genome assembly. More precise bioinformatics can separate multiple WGD or detect events deeper in the past. However, there are technical challenges as more ancient peaks blur together and require mixture modeling to separate (Ruprecht et al. 2017; Maere et al. 2005; Smith, Brown, and Walker 2017). 4DTv is a stricter version of Ks that only uses sites that are 4-fold synonymous codons for comparing the two genomes to ensure clock like behavior.

Ks plots are calculated using nucleotide sequence alignments between pairs of paralogous genes identified by sequence similarity. Ks is the synonymous substitution rate: the frequency at which the synonymous third position of codons differs between gene copies. Synonymous substitutions are assumed to be neutral to selection and thus a better clock-like estimate of the amount of time passed since the duplication event which created the second copy. A peak in the Ks plot would indicate a large number of genes which were all created around the same time. Currently, the only known cause for a significant fraction of the genome to be duplicated is a WGD. Duplication events which happened further back in time will have larger Ks values. More ancient WGDs are also expected to have wider peaks because every gene does not accumulate synonymous mutations at exactly the same rate. Paralogs with Ks values which do not group into a peak are assumed to be small scale duplications (SSD) (Maere et al. 2005).

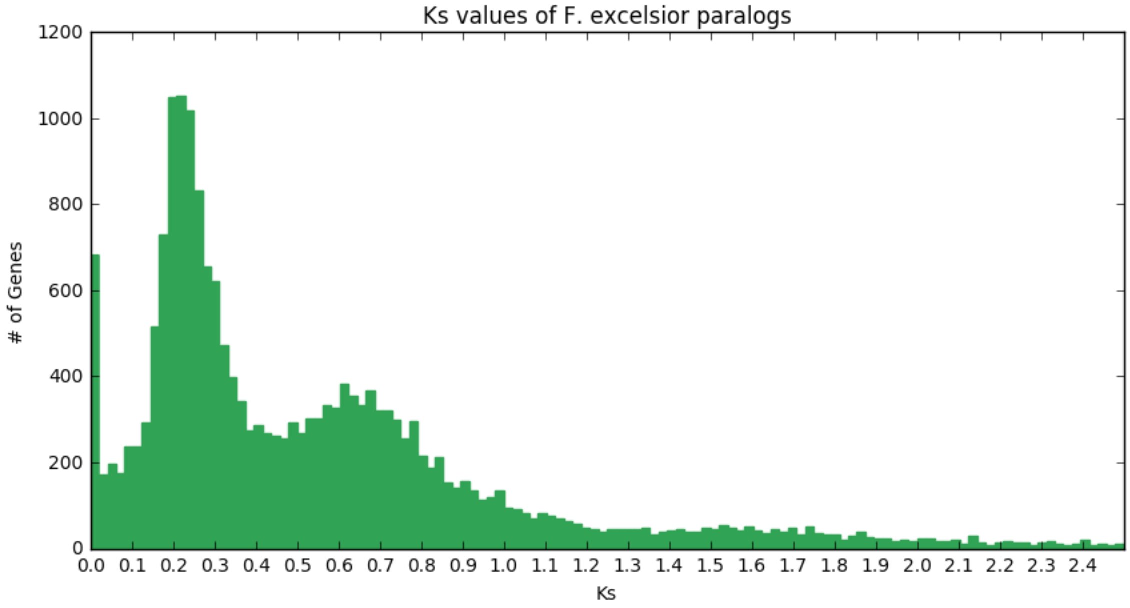


Figure 1.1 Ks Plots: are a graph of the number of synonymous substitutions in the wobble position of paralogous genes within an individual. Peaks are used to estimate at what time a group of genes were duplicated together in a WGD. In this graph, the peak at zero is likely caused by a recently duplicated SSD or assembly artifacts from the diploid phase. The peak at 0.25 is from the aWGD 25 million years ago shared by all of *Fraxinus* and *Olea*. An older duplication, around 60 Mya is also shared by Jasmineae. **Source:** Figure generated from raw data gathered in Sollars et al. (2017).

1.3.2 Phylogenetic Analysis

Phylogenetic analysis is a second method, which makes WGD timing more precise by noting whether two species share the WGD and so constrain the timing. To make WGD timing more precise, researchers use phylogenetic analysis to tell if two species share the WGD so that the timing can be constrained to before or after their species divergence (Rabier, Ta, and Ané 2014).

The Barker lab has developed an algorithm called MAPS which uses a more advanced form of phylogenetic analysis (Li et al. 2015). Their approach takes any number of species which may share multiple WGD. MAPS iterates through each species gene trees and analyzes which gene tree subsets would support a WGD at a particular node. This process is iterated until all probable WGD are identified in Newick format. The results are compared against a null model to avoid false positives.

1.3.3 Shared Synteny

Synteny is the co-occurrence of homologous genes at the same genomic locus in two separate species. Before genome assemblies, co-localization was defined at the level of chromosomes. With genome assemblies, synteny now often refers to a region of two assemblies that share the same ordering of homologous genes. This new understanding of synteny can be more specifically referred to as shared synteny.

Shared synteny evidence can be seen when plotting all hits of shared sequence between two genomes on a dot plot, with genome A as the x-axis and genome B as the y-axis. When two species share >10 Mbp regions of sequence in the same order, then it can be inferred they have been broken up by translocations and chromosome fusions after a WGD (E. H. Lyons 2008; E. Lyons et al. 2008). Synteny evidence has the advantage that it is improbable to occur by chance and may give additional information about which genes have been selected and retained together.

Unfortunately, fitness selection for sets of sequence to be colocated also means synteny is not a neutral clock-like indicator of history. The recently growing study of chromatin conformation capture demonstrates there are biological consequences for transcription regulation based on the location of genes (Lieberman-Aiden et al. 2009; Rao et al. 2014; Grob, Schmid, and Grossniklaus 2014; Delaneau et al. 2019). The fitness neutral assumption of location in synteny analysis is analogous to synonymous substitutions used in Ks plots to track clock-like accumulation of neutral mutations in homeologous pairs. Non-synonymous positions are not used as a molecular clock, since divergent bases are interpreted as positive selection and identical bases are interpreted as purifying selection regardless of evolutionary distance (Loewe 2008). To apply this comparison, discoveries of function in chromatin conformation shift synteny out of the domain of clock-like historical indicators and into the domain of shared (or divergent) functional indicators. The exact kilobase resolution of function for super enhancers, chromatin loops, and nuclear subcompartments (Rao et al. 2014) sets the upper bound on the feature size of micro-synteny that can confidently be labeled as neutral historical signal free from functional constraints.

1.3.4 Paleopolyploidy in Angiosperms

These techniques have identified evidence for paleopolyploidy throughout the tree of life, particularly in angiosperms (Li et al. 2016). Newer techniques allow researchers to detect events further back in evolutionary time. An expressed sequence tag (EST) study in Asteraceae first identified 5 different WGD among 18 species (Blanc and Wolfe 2004; Barker et al. 2008). Jiao et al. (2011) identified ancient polyploid ancestors further back in time for all seed plants and angiosperms. All three studies use phylogenetic analysis and Ks plots based on aligned sequences.

As sequencing increases, almost every new plant genome brings new reports of paleopolyploidy: for example in maize, tomato, and *Brachypodium* (Schnable, Springer, and Freeling 2011; Tomato and Consortium 2012; Gordon et al. 2017). This provides an opportunity to look at the long-term outcomes of theories of new polyploids and see how they compare with the ancient past. For example, Garsmeur et al. (2014) demonstrate that parental subgenome dominance can be detected in the ancient past by examining different signatures of gene copy loss and gene expression. They show that autopolyploids and allopolyploids can be distinguished because ancient autopolyploids have unbiased fractionation and equivalent gene expression between their subgenomes.

To support their claim they reconstruct the chromosome lineage of *Musa* through two WGD to trace duplication and loss.

1.3.4.1 Retained Percentages of Genomes

The observed and theoretically predicted drop in the percentage of the genome retained in duplicate after diploidization is notable because only a small minority of sequence remains behind as evidence (Figure 1.2). Ren et al. (2018) conducted a review of 105 angiosperms and found 17 new WGD. Their study included transcriptomes with no full assembly as a separate class of data. While useful, this approach prompted further discussion on data quality and possible confounding factors (Zwaenepoel et al. 2019; Wang et al. 2019). Figure 1.3 shows how the merging of these two lines of evidence begins to blur the line between false positive (grey dots) and true positive (red and purple dots) WGD detection.

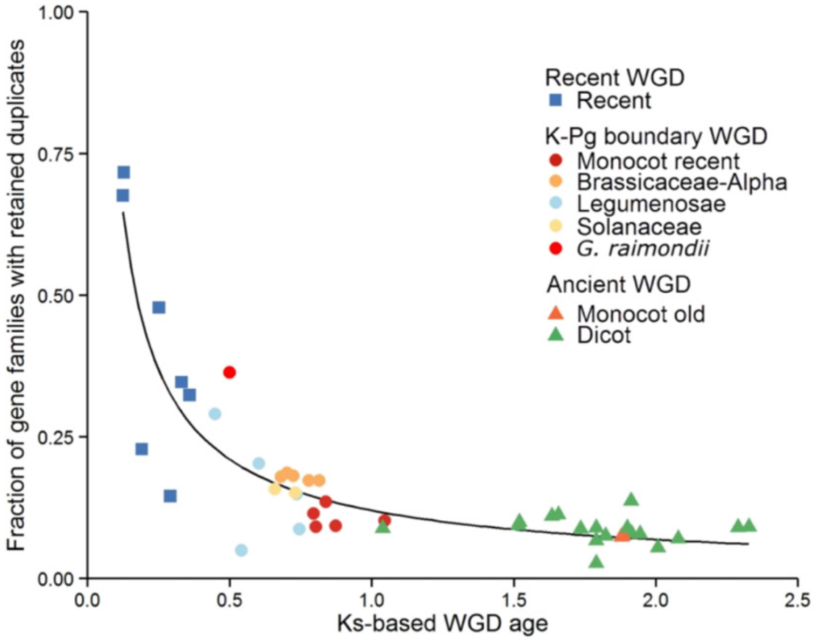


Figure 1.2: Duplicate gene retention as a function of time since WGD: Across all studied angiosperms, the percentage of genes retained levels off around 12% after a Ks based age of 0.75. This means the rate of loss is much higher directly after a WGD and gradually converges to a subset of gene families retained in duplicate. **Source:** Figure 3 of (Li et al. 2016) is reprinted here with permission.

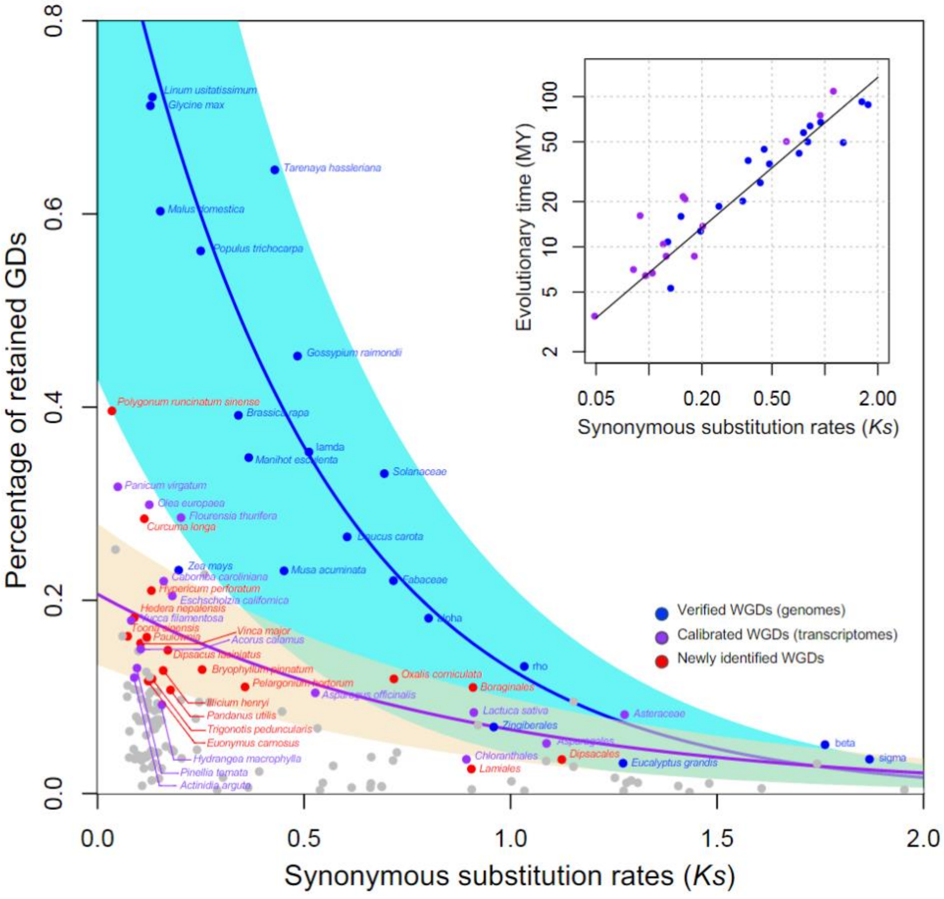


Figure 1.3: A Model for Exponential Decrease over Time of Number of Duplication Events with Actually Detected GDs: The authors present a scatter plot of 105 Angiosperms by percentage duplicates and synonymous substitution rates among inferred homeologs. **Source**: Figure 2 of (Ren et al. 2018) is reprinted here with permission.

1.3.5 Boom and Decay Model

The "Boom and Decay" model developed by (Wolf and Koonin 2013) proposes that occasional WGDs are followed by long periods of genome size reduction (fractionation). They postulate a gradual increase in genome size and concomitant morphological complexity over hundreds of millions of years as genes families which survive multiple rounds of WGD and fractionation can quickly reach 128x their original copy number. The Boom and Decay model explains several factors, including evidence for paleopolyploids, the current stable state of many plant genome sizes, and the disproportionate effect that WGDs have had on some gene families while others appear untouched. This model posits that most changes happen very early on after duplication and this can be contrasted with the Lag Time model.

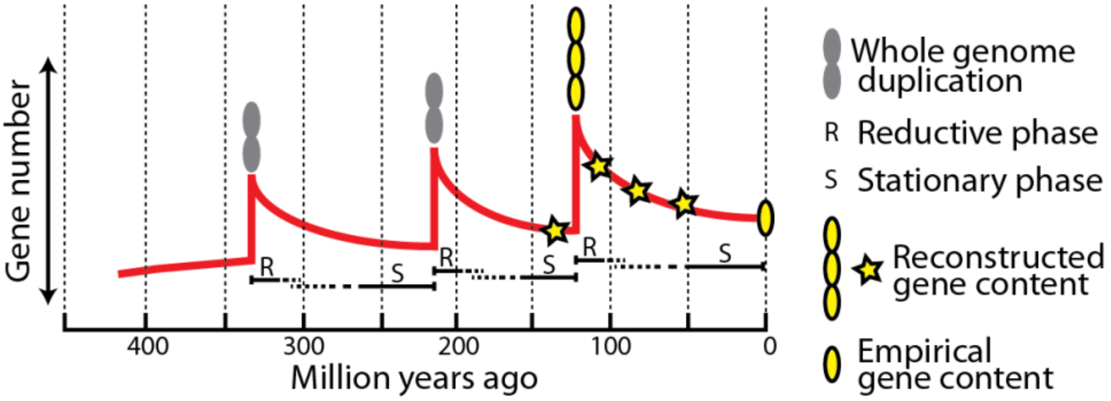


Figure 1.4: "Boom and Decay" model of nuclear gene content: ***Source**:* Figure from the Marie Curie grant proposal that generated the current PhD thesis topic (Cooper, 2014), used with permission.

1.4 Biological Factors Affecting Fractionation

The above discussion has focused on factors influencing entire subgenomes. In this section, we examine factors which determine why a specific gene or gene family is either retained in duplicate or quickly reduced to single copy after a WGD. Changing dosage of even a single gene can have a profoundly negative impact upon an organism. For example, Huntington's disease is caused by an excess of copies of a small repeat whereas Down Syndrome is caused by an extra copy of human's smallest chromosome (Bahlo et al. 2018; Makino and McLysaght 2010). While plant genomes are surprisingly robust to large changes compared to animals, they still include a complex gene regulatory network that can be disrupted (Kejnovsky, Leitch, and Leitch 2009).

1.4.1 Retention by Gene Function

Gene function influences which genes have been retained over deep time. For instance, De Smet et al. (2013) identify a set of singleton genes conserved across all eukaryotes which are resistant to duplication over deep time and possess many gene function specific attributes. They are over-represented in essential housekeeping genes expressed across a wide range of cell types, and highly expressed. In contrast, another notable review analyzes 37 angiosperms and finds that the set of gene families based on sequence and reflected in function is consistently retained in duplicate after WGD (Li et al. 2016). Specifically, they show that the number of gene families that remain single copy is much higher than null model simulations. The null model simulation and classification of single-copy gene families versus multi-copy gene families serves as the template for the *Fraxinus* analysis in Chapter 4 of this thesis.

1.4.2 Dosage Constraints

An individual gene may have absolute dosage constraints if protein abundance is the limiting step in critical metabolism. Evidence can be found in *Saccharomyces cerevisiae*, whose WGD increased the absolute dose of genes necessary to ferment glucose in the presence of oxygen (Chen, Xu, and Gu 2008). The ancient WGD-β (70 Mya) in *Arabidopsis* shows limited evidence for increasing the dose of individual genes, however they conclude the more recent WGD-α (20 Mya) has no such association (Bekaert et al. 2011). The evidence for constraints on relative dosage in a network of proteins is far more prevalent.

1.4.3 Stoichiometric Constraints

Genes that operate in a pathway or protein complex are sensitive to changes in their relative abundance. Stoichiometric constraints refer to ratios between genes, frequently called “Relative Dosage” in the literature. A survey of the minimal genome of *Arabidopsis thaliana* found clusters of “connected genes” which were retained after a duplication (Thomas, Pedersen, and Freeling 2006). This is seen as evidence of the Gene Balance Hypothesis, which states that the relative quantities, not absolute amounts of gene products, are under selection. Chief among gene families amplified by WGD are transcription factors which regulate hundreds of genes per protein and can have a profound effect on gene expression. Studies have repeatedly confirmed that TFs are retained in duplicate after WGD and show very little copy number variation outside of these events (Edger and Pires 2009).

Proteins which form complexes are a prime example of stoichiometric constraints and can also be found clustered in Yeast genomes (Teichmann and Veitia 2004). The relative proportions of each part of a complex determines their reaction rates and thus the final number of complexes formed. One protein can be involved in multiple complexes which creates the counterintuitive effect that organisms with more of one protein will create less of the final complex due to lack of availability of the other parts (Birchler et al. 2005). Oberdorf and Kortemme (2010) note that it is protein complex topology, not membership, which determines a gene’s dosage sensitivity because complexes assemble in a stepwise process. From this dynamic, it is predicted that genes in complexes will not be duplicated in small scale duplications (SSD), but commonly duplicated in WGD.

1.4.4 Retention Rate In Homeologs Is Different From Small Scale Duplications

The final piece of evidence for the Dosage Balance Hypothesis comes from how genes react to different types of duplications. In addition to WGD, genes can be duplicated in tandem, in large segments of chromosomes or through transposition. Experimental evidence points to dosage sensitive “connected genes” copy number variation only being non-deleterious when the gene is duplicated along with its partners such as in WGD and segmental duplications (Freeling 2009). This pattern of duplication tolerance in WGD appears to be a widespread principle as it occurs in *Arabidopsis, Oryza, Saccharomyces* and *Tetraodon* as well as other taxa (Paterson et al. 2006). In the soybean *Glycine max* it was discovered that genes in the photosystem are exclusively duplicated in WGD, while genes in the Calvin Cycle did not show this same dosage balance sensitivity (Coate et al. 2011).

1.4.5 Lag Time Model

Lag Time was proposed to explain observations in cases where one class of gene was retained for millions of years after WGD, while another class of genes were immediately reduced to single copy number (Schranz, Mohammadin, and Edger 2012; Tank et al. 2015). The Lag Time model proposes that these genes are essential and dosage sensitive when first duplicated, however subsequent mutation accumulation in both copies can gradually relax the dosage constraint, leading to the copy’s eventual loss (Dodsworth, Chase, and Leitch 2016; Robertson et al. 2017; Cheng et al. 2018; Clark and Donoghue 2017).

1.4.6 Lag Time Evidence

We can use Bekaert et al.'s study of *Arabidopsis thaliana* to estimate time scales for when Lag Time model begins and ends (Bekaert et al. 2011). Bowers et al. (2003) identified three WGD in the *A. thaliana* lineage: WGD-*α* (20 Mya), WGD-β (70 Mya) and gamma (170 Mya). Given two competing models, the researchers compare which model best predicts each WGD retained gene set. WGD-*α* duplicated genes are consistent with the Dosage Balance hypothesis, not the Lag Time model. In contrast, β duplicated genes are necessary for high metabolic flux, while the Dosage Balance model fails to predict which genes will be retained in duplicate from the WGD-β. From this they conclude that Lag Time restrictions are already fully resolved in the more ancient WGD-β. This would place Lag Time resolution between 20 to 70 million years after a WGD in plants with similar dynamics.

It is difficult to find a consensus on these dates since the *Arabidopsis* line has apparently had an anomalously high synonymous substitution rate and different studies using different plants will arrive at different dates (Jaillon et al. 2007). This review uses dates from (Unver et al. 2017) for *Arabidopsis thaliana*. Since the organism under study matters, next we'll discuss paleopolyploidy studies, specifically in *Fraxinus* and *Oleaceae*.

1.5 *Fraxinus* Model System

Oleaceae is the family of Asterids including Ash trees (*Fraxinus*), Olive trees (*Olea*), and Jasmine (*Jasmineae*). *Oleaceae* has economic significance as Ash trees are the most common hedgerow tree in the UK and olives are a food staple for Mediteranean diets. In Europe, Ash trees are dying off in unprecedented numbers due to an Ash Dieback fungus (*Hymenoscyphus fraxineus*) which spread from Asia in 2012 where it was approximately in equilibrium with the native *Fraxinus* species. In America, Ash trees are threatened by the Emerald Ash Borer (*Agrilus planipennis*). Some species of *Fraxinus* are more resistant to these threats than others, so a wide scale sequencing project was launched in 2013 in the hopes of discovering which sequence traits confer resistance. A total of 28 species of *Fraxinus* were sequenced for the world-wide Fraxinus genome project (see Chapter 2: Genome Assembly and Annotation). We also have high quality chromosome assemblies for *Fraxinus excelsior* (N50=103,995) and *Fraxinus pennsylvanica* (N50=27,152,721, Table 3.2); the key species for the UK and US respectively.

1.5.1 Ash Trees as a Model for Studying Fractionation

Fraxinus may be an ideal model organism for testing theories about paleopolyploidy because all 28 species share an ancient WGD 25 million years ago (Julca et al. 2018, Figure 1.1). Other studies reviewed here never go beyond a ratio of 3 species per single WGD. A 28:1 ratio in *Fraxinus* provides an opportunity to test predictions about the repeatability of evolution. Specifically, are the same gene families always lost during fractionation in independent lineages? What is the timing of these losses? Which gene families are preserved in duplicate? Can gene function or network interactions be used to predict gene copy number?

1.5.2 Evidence of Fraxinus WGD

The main evidence for an ancient whole genome duplication (aWGD) in *Fraxinus* are peaks in the Ks plots between paralogs, indicating a large number of genes that were all duplicated at the same time. With the sequencing of *Fraxinus excelsior* (Sollars et al.

2017) and *Olea europaea* evidence was found indicating these species share two WGDs (Figure 1.5). *Olea europaea* has a third WGD distinct to genus *Olea* (Julca et al. 2018).

1.5.3 Separating Two Fraxinus WGD

In order to be used as a model system, we need to first assess what resources are available to be able to separate different WGD events. When two WGDs happen very close together in time, it can be impossible to separate out genes which were duplicated in each event. Probabilistic inference or a separate genome that shares one, but not both events can be used to distinguish them (Rabier, Ta, and Ané 2014; Tiley, Ané, and Burleigh 2016).

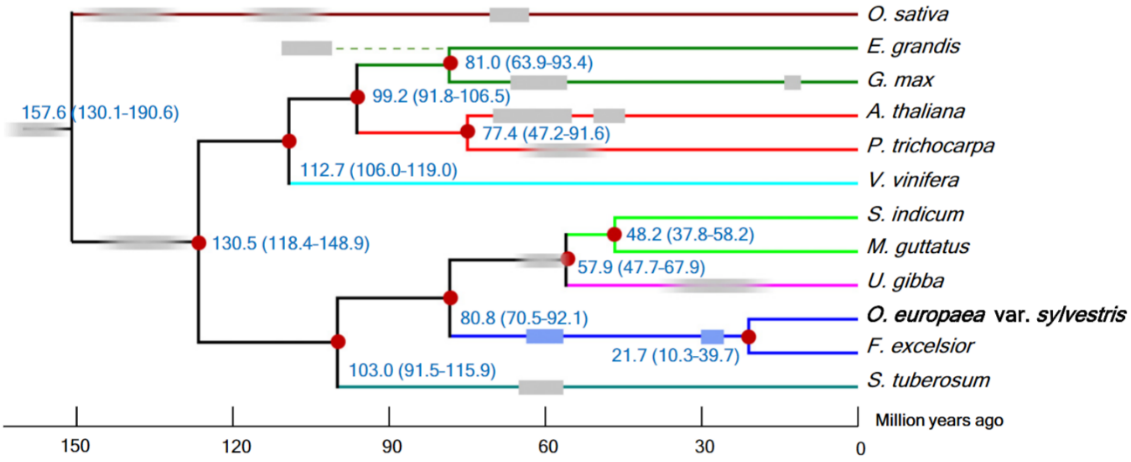


Figure 1.5: Angiosperm WGD in the last 150 million years are marked as grey rectangles. Species divergence times are marked by red dots with confidence intervals for their age. *Fraxinus excelsior* and *Olea europaea* share two WGD (blue rectangles). **Source:** Figure 2 of (Unver et al. 2017) used with permission from publisher.

In *Fraxinus*, there are two WGD events estimated at 26 Mya and 60 Mya (Fig. 5). *Olea europaea* also has a WGD not shared with *Fraxinus* (Fig. 6). The older WGD shared by *Fraxinus* and *Olea* does not include *Jasmineae* (Figs. 5 & 6). *Jasmineae* could be used to separate the *Fraxinus* WGD from the older *Oleaceae* WGD. However, as of the time of writing there was not a full genome assembly available for *Jasminus sambac* (Y.-H. Li, Zhang, and Li 2015) and is not used in this study. Also, *Phillyrea angustifolia* (sister species to *O. europaea* in Figure 1.6) is an allotetraploid (Olofsson et al. 2019). This unfortunately means there's no diploid species outgroup to contrast with the *Fraxinus* WGD. This is addressed further in Chapter 2 using Gene Tree Reconciliation.

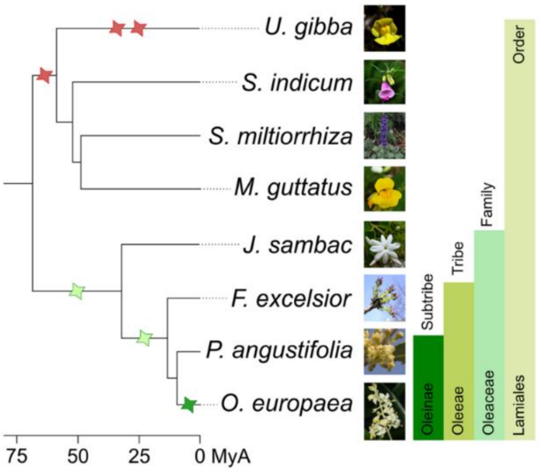


Figure 1.6: *Jasminus* and *Olea* WGD: This species tree from Julca 2017 shows the relevant species for this thesis along with three WGD events (stars). Red stars are WGD not in the *Fraxinus* lineage. The second green star is a WGD at 26 Mya and the subject of study in *Fraxinus excelsior* for this thesis. **Source:** Fig. 5 of (Julca et al. 2018) is used under the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

1.5.4 Timing of Shared Duplications

It is very likely that the most recent *Fraxinus* WGD is shared with *Olea* because there is evidence the same set of duplicated genes is shared between species. This evidence, in the form of Ks (4DTv) graphs, has been calculated using paralogs between *Fraxinus* and *Olea* to verify shared timing of the same WGD (see Julca et al. 2018 Fig. 3 and Fig. S7). Unver et al. (2017) estimate the most recent WGD to 26 Mya and the divergence of *Olea* and *Fraxinus* to 5 million years later. Given the confidence interval on both the species divergence time and the timing of the WGD, it is possible that the WGD contributed to the divergence of the *Olea* and *Fraxinus* lineages.

The exact dates of lineage divergence in the Oleaceae differ among published phylogenetic trees. For the purposes of methodological consistency, we have identified a single source that includes all species under study in this thesis. (Zedane 2016) uses plastid and rDNA sequences to construct a time tree with over 136 species in Oleaceae. Using organellar DNA has the advantage that WGD nuclear genome instability will not cause artifacts in the dating methods. Zedane's time tree is used for all future calibration dates in this thesis (see Table 3.1, 3.2.3).

1.6 Conclusion

The dynamics of how polyploidy affects evolutionary outcomes is worthy of further study. Polyploids, both past and present, have played a role in the geographic expansion of plants into new regions and the expansion in copy number and diversity of some gene families. Ample evidence in the form of shared timing of gene duplicates indicates seed plants are the descendants of multiple rounds of ancient whole genome duplication and fractionation. Despite minority cytotype exclusion, genome instability, and transgressive expression drawbacks these paleopolyploid organisms outcompeted all of their diploid relatives to become the ancestor of all future generations.

The process of fractionation of genes to a diploid state is still not fully understood. There is broad evidence that some gene families are preferentially retained after WGD including transcription factors and developmental genes. WGDs provide a rare opportunity for the duplication of gene sets which are constrained by relative dosage in protein complexes or pathways.

However, there are still causal relationships to untangle. Given that most WGD events studied so far have only a few species per event, how repeatable are these outcomes? Even if fractionation is repeatable, is it driven by positive selection of fitness or neutral evolution attributes like the chromosome position of a gene? The Worldwide *Fraxinus* genome project provides a novel opportunity to test the repeatability of fractionation using 28 diploid species that all share the same whole genome duplication event.

1.7 Future Chapters

In the next chapters, we will apply the methods discussed here to use 28 new *Fraxinus* genomes to address these questions. Chapter 2 will introduce genome sequence visualization and genome alignments, which serves as the introduction for the second aspect of this bioinformatics thesis. I introduce FluentDNA, a new tool for genome assembly construction and exploration. Chapter 3 details the methods used to construct the 28 *Fraxinus* genome assemblies. Gene families with Reconciled Gene Trees are constructed to identify gene pairs (homeologs) from the most recent *Fraxinus* WGD.

Chapter 4 will leverage this data to answer questions on the repeatability of fractionation after a Whole Genome Duplication. Chapter 4 uses *Fraxinus* to test three key theories about paleopolyploidy from the literature: i) the repeatability of fractionation is tested using convergent evolution in 28 *Fraxinus* species ii) Gene Balance Hypothesis is tested using GO term enrichment analysis on fractionated families iii) Lag Time Model is tested by constructing a detailed timeline of gene losses over the *Fraxinus* lineage.

Chapter 5 broadens the scope of comparative genomics from gene families to whole genome alignment. We address the technical challenges in multiple sequence alignment for entire genomes with many rearrangements using Graph Genomes. Chapter 5 introduces Pantograph, the first Graph Genome browser capable of scaling to thousands of individuals. Chapter 6 discusses the challenges and future directions for comparative genomics. Transitioning from comparing a pair of genomes to analyzing large sets of complete genomes requires a new set of genomic tools and methods.

Chapter 2
FluentDNA: Nucleotide visualization of whole genomes, annotations, and alignments

Author Contributions and Collaborations

This chapter was previously published separately in the journal *Frontiers in Genetics* in April, 2020, co-authored with Richard Buggs. It is included here with minor edits to fit the thesis. FluentDNA started as a Python rewrite of DDV implemented by Tomasz Neugebauer which could visualize a column layout for single sequences of length up to 300Mbp. Initial FluentDNA source code was co-authored with Bryan Hurst and this was subsequently massively extended to support whole genome visualization. The authors thank Yan Wong for identifying Peano curves could be used in chromosomes and creating a JavaScript prototype showing a path with configurable radices.

Abstract

Researchers seldom look at naked genome assemblies instead, the attributes of DNA sequences are mediated through statistics, annotations and high-level summaries. Here we present software that visualizes the bare sequences of whole genome assemblies in a zoomable interface. This can assist in detection of chromosome architecture and contamination by the naked eye through changes in color patterns, in the absence of any other annotation. When available, annotations can be visualized alongside or on top of the naked sequence. Genome alignments can also be visualized, laying two genomes side by side in an alignment and highlighting their differences at nucleotide resolution. FluentDNA gives researchers direct visualization of whole genome assemblies, annotations and alignments, for quality control, hypothesis generation, and communicating results.

2.1 Introduction

An intrinsic part of the analysis of genomic data is the summarization of large sequence datasets. This accomplishes three primary tasks: (1) quality checking an output, (2) understanding a sequence in context and (3) communicating about sequence data in talks, posters and articles. This summarization is commonly achieved via metrics or by visualization. Simple metrics have the advantage of being precise, concise and easy to transmit, for example: N50, GC content, the mean size of exons and introns, and percent alignments. Tables of metrics can be used to convey information about, for example, overrepresented k-mers, or the location of low complexity regions or gene annotations. On the other hand, visualizations can give a broad, spatially explicit overview of sequence data.

Many software tools exist to visualize DNA sequence data, but in those that do include the bare sequence, it is only shown at the smallest scales. Genome browsers display nucleotide sequence only when zoomed to sub-kilobase scales, but not in broader overviews, and usually show annotations as linear blocks or line graphs in parallel tracks (Robinson et al. 2011; Kuhn et al. 2013; Buels et al. 2016). Multiple-Sequence Alignment

(MSA) editors such as Jalview have zoomable depictions of nucleotides or amino acids as colored blocks allowing variation between vertically organized samples to be picked out by the naked eye (Waterhouse et al. 2009; Katoh et al. 2017). Chromosome painting gives large scale summaries of genome structure, for example showing translocations between chromosomes using different colors (Serov et al. 2005; Kemkemer et al. 2006; Rasmussen et al. 2014). Circos plots visualize large scale rearrangements, such as syntenic blocks, with arcs (Krzywinski et al. 2009). SynTView uses heat-maps to depict variation among sequences (Lechat et al. 2013). To investigate tandem repeats and the subtle repeat pattern of codon bias, the tool SpectroFish uses a vertical axis to represent frequency (Sussillo et al. 2004; Sánchez and Lopez-Villasenor, 2006). DNA Walk visualizes sequence in terms of spatial steps (Arakawa et al. 2009). Ensembl, VisGenome and BugView all offer a browser view for aligned genomes, though they focus on larger features such as genes (Leader, 2004; Jakubowska et al. 2007; Zerbino et al. 2018) or gene presence/absence. These approaches do not show the negative space of intervening sequence (Hennig et al. 2015). In contrast, dot plots do show negative space and can handle densely connected or noisy data well; they are used for synteny analysis by duplicating the x-axis to form a square matrix of matching sequences (Lyons, 2008). More abstract visualizations which still use sequence are CGR, which shows k-mer representation (Deschavanne et al., 1999; Joseph and Sasikumar, 2006). BioJS sequence viewer (Yachdav et al. 2015; Paladin et al. 2020), and Genome Projector (Arakawa et al. 2009) provide multiple ways of viewing genomic information and sequence variation at a range of scales. For genome assembly and pan-genome studies, visualization is used for quality control, for example, in Pan-Tetris (Hennig et al. 2015), Blobtools (Laetsch and Blaxter, 2017), Hawkeye and AMOS (Schatz et al. 2013).

In several areas of information technology, direct visualization of big data has accelerated data analysis. This has been key to the success of the company Palantir, whose software enables humans to work out complex interrelations in data (Khurana et al. 2009; Wright et al. 2009; Hossain et al. 2011). Other companies use the visualization of raw binary data representing executable code in computer security research to seek the location of passwords, encryption, obfuscation and malware. One approach uses Hilbert space filling curves to calculate the entropy of programs (Conti et al. 2010; Cortesi, 2011). The software Cantor Dust uses this approach together with k-mer representation graphs (https://sites.google.com/site/xxcantorxdustxx/). Cantor Dust was acquired by Batelle, a think-tank for the CIA (Miller et al. 2001), though some features are available in open source derivatives Veles and Senseye (Stahl; Rombouts, 2014).

Given the success of raw sequence visualization in other areas of big data analysis, it is reasonable to ask whether these techniques would also aid in genetic research and communication. A simple way to visualize large sequence files has been pioneered by DNA Rainbow (Bierkandt and Bierkandt, 2009), DNASkittle (Seaman and Sanford, 2009) and DDV (Neugebauer et al. 2015). These depict single DNA sequences as colored pixels (like an MSA editor), but introduce line breaks which wrap long sequences into 2D blocks. DNA Rainbow has a single raster column per chromosome with a fixed width of 3,500 pixels; this makes all but very large features difficult to discern by eye. DNASkittle has a variable column width optimized for tandem repeats and a suite of visualizations for exploring sequence similarity features in detail; this single column layout and 1D zoom is not ideal for use on large datasets, and it handles draft genomes and multiple chromosomes poorly. DDV introduces a more intuitive 2D zoom feature using sets of columns in a single layout, but does not support annotations.

In this paper, we present the tool FluentDNA, which visualizes sequence data with nucleotides as colors in a 2D layout with a zoomable interface. The layout can scale to accommodate any number of chromosomes and scaffolds. Individual nucleotides are visible when zoomed in and colors are averaged in zoomed out images. Even in the absence of any annotation of a genome, FluentDNA allows the human eye to pick out key features of a genome assembly by size and nucleotide composition. With practice, major features of chromosome architecture including centromeres, isochores, telomeres, and tandem repeats can be identified from the naked sequence because changes in k-mer usage cause changes in color and texture. Contamination is visible because of G/C content and coverage differences. FluentDNA expands on DDV's visual paradigm with a suite of features such as the ability to handle multi-part FASTA files and whole genome assemblies, output different layout types, and visualize annotations, repeats, and alignments. It works on Windows, Mac and Linux. FluentDNA thus gives researchers direct visualization of their data files, for quality control, hypothesis generation, and communicating results. It can also promote the public understanding of science through public webpages and interactive museum displays.

2.2 Methods

We designed our software to use the following conceptual methods for an easy-to-use whole genome visualization tool.

2.2.1 Nucleotides As Pixels

Nucleotide sequences can be depicted as a series of pixels where the four bases are represented by four colors. The ideal color palette will conform to the following criteria: (1) high contrast; (2) friendliness to color-blindness; (3) typical nucleotide compositions should be viewable for over 20 minutes without causing discomfort (i.e. greens and blues should predominate)(Kaya and Epps; Mehta and Zhu, 2009).

2.2.2 Depiction of One-Dimensional Locality In Two Dimensions

To visualize long nucleotide sequences meaningfully in two dimensions, locality in the second dimension of the visualization must approximate locality in the one-dimensional source data. The simplest way to do this is a linear sequence with frequent line breaks, ordered into a set of nested tiles. In this Tiled Layout, horizontally-neighboring pixels are true neighbors in the source data, whereas vertical neighbors are spaced in the source data by the size of the column width. Another approach, referred to here as an Ideogram Layout, uses space filling curves. These are fractal shapes which fold a one dimensional continuous path to fill a 2D (or higher) area (Bially, 1969; Haverkort and van Walderveen, 2010) with no line breaks. One type, the Peano curve, is made of spirals of spirals, continually wrapping back in on itself to occupy the available space nearest to its origin. This process is recursive so locality is preserved at all scales. Peano curves approximate more closely the arrangement of nucleotide sequences in the interphase nucleus than do tiled arrangements (Lieberman-Aiden et al. 2009). However, it is impossible for the human eye to trace exact nucleotide sequences in a Peano curve: their utility is mainly restricted to broad overviews of data.

2.2.3 Pan-And-Zoom Functionality

Eukaryotic genomes tend to be hundreds of megabases and even gigabases in length. When visualizing them in two dimensions, rapid and seamless pan-and-zoom functionality is essential. When zoomed out, pixel colors should be merged together to give an approximate representation of the nucleotide content by color.

2.2.4 Mouseover Functionality

To move from visualization to analysis of specific genomic features with other software, users should be able to retrieve the sequence at any given point in the visualization simply by hovering over it. It should be possible to export snippets of DNA sequence as letter codes for further analysis.

2.2.5 Annotations

Annotations can be visualized in two ways: (1) by directly highlighting nucleotides which are present in a genome feature; this works for both tiled and ideogram (see above) visualizations or (2) by a side-by-side column in a tile layout showing the location of features.

2.2.6 Whole Genome Alignments

Whole genome alignments are commonly available as liftOver files. Using these, reference and query genome sequences can be visualized in side-by-side tile layouts where indels are depicted as gaps in one or the other genome. To highlight differences due to SNPs, indels, and rearrangements, extra columns can be added showing nucleotide differences between the two genomes, making them visible at a wide range of zoom scales. Different background colors can be used to indicate different types of rearrangements, though rearrangements within rearrangements will be hard to portray.

2.3 Implementation

These methodological concepts are implemented by FluentDNA in a Python code base with Javascript for browsing and mouseover. Python code handles the rendering of fasta files, annotations, and genome alignments as well as a file server. Javascript code depends on OpenSeadragon 2.4, Biojs Sequence 1.0, and jQuery 1.7 (OpenSeadragon; Resig et al. 2006; Yachdav et al. 2015). FluentDNA is available on MacOS and Windows as an executable command line tool or a GUI. It is available on all platforms as a python standalone library. The logical framework on FluentDNA is shown in Figure 2.1.

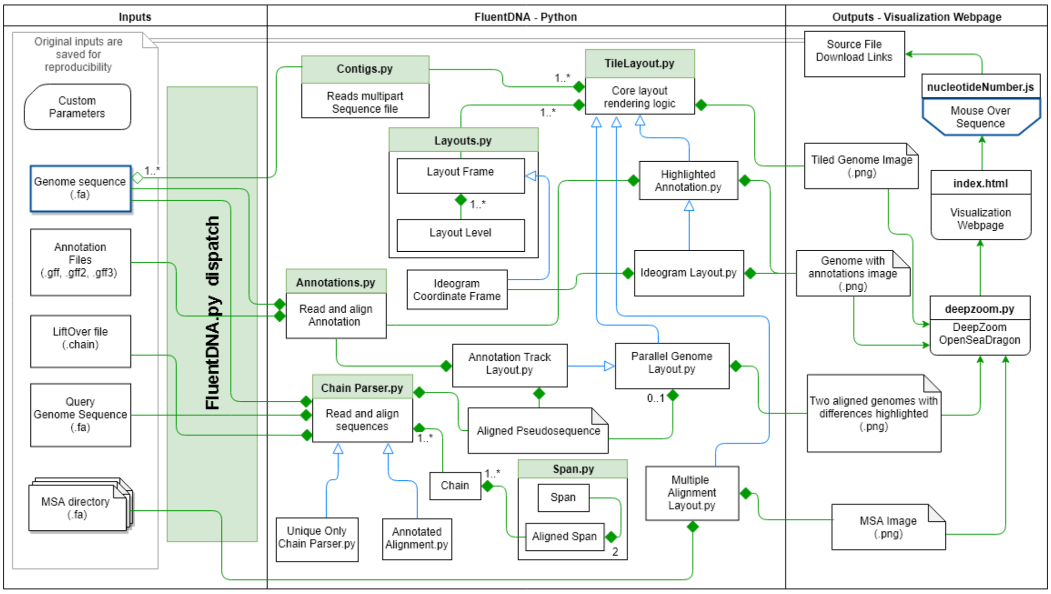

Inputs
FluentDNA - Python
Outputs - Visualization Webpage
Original inputs are saved for reproducibility
Custom Parameters
Genome sequence (.fa)
Annotation Files (.gff, .gff2, .gff3)
LiftOver file (.chain)
Query Genome Sequence (.fa)
MSA directory (.fa)
FluentDNA.py dispatch
1..*
Contigs.py
Reads multipart Sequence file
TileLayout.py
Core layout rendering logic
1..*
1..*
Layouts.py
Layout Frame
1..*
Layout Level
Highlighted Annotation.py
Annotations.py
Read and align Annotation
Ideogram Coordinate Frame
Ideogram Layout.py
Annotation Track Layout.py
Parallel Genome Layout.py
0..1
Chain Parser.py
Read and align sequences
1..*
Aligned Pseudosequence
Chain
1..*
Span.py
Span
Aligned Span
2
Multiple Alignment Layout.py
Unique Only Chain Parser.py
Annotated Alignment.py
Source File Download Links
nucleotideNumber.js
Mouse Over Sequence
Tiled Genome Image (.png)
index.html
Visualization Webpage
Genome with annotations image (.png)
deepzoom.py
DeepZoom OpenSeaDragon
Two aligned genomes with differences highlighted (.png)
MSA Image (.png)

Figure 2.1: FluentDNA Implementation UML showing the relationship between objects in the program. Green diamonds mean an object has one or more of the connected objects. Blue arrow mean one object inherits all the properties of another object. Based on the input files provided (left), FluentDNA uses different rendering modes specified by the user, routed through FluentDNA.py. A FASTA sequence can be rendered in Tile or Ideogram style, each of which can also have gene annotations (GTF/GFF2/GFF3) overlaid with HighlightedAnnotation.py. Whole Genome Alignments handled by ChainParser.py require two FASTA files and a LiftOver file. Annotation Track Layout and Alignments both use the Parallel Genome Layout module and provide pseudosequences for further analysis. On the right, FluentDNA produces a web directory containing all input files and parameters (for reproducibility). The Visualization Webpage (top right) requires no installation and provides mouseover sequence retrieval. A glossary of files is listed in Supplemental 2.

2.3.1 Input Data

FluentDNA reads single or multiple sequence FASTA files of any size that the host machine's memory can accommodate. For annotations, it reads GFF, GFF2 and GFF3 files. Visualizing whole genome alignments requires input of two genome assemblies in FASTA format and a liftOver file describing their alignment. The FluentDNA dispatch selects the appropriate layout based on input data and user parameters entered through the command line or GUI.

2.3.2 Tile Layout

A FASTA file of any size can be visualized by FluentDNA in a tile layout (Figure 2.2). The default layout is arranged in powers of ten: rows of 100 pixels (each pixel representing one base), in columns of 1,000 rows containing 100 Kbp. One hundred columns are arranged in 10 Mbp mega-rows. Chromosomes occupy mega-columns composed of enough mega-rows to accommodate the largest chromosome (default 260 Mbp). Chromosomes are laid out side by side and several smaller chromosomes can share a single mega-column. In the default layout there is no white space within and between rows, 3 pixels of white space between columns, 9 between mega-rows, 700 pixels between chromosome columns. This default layout is defined in FluentDNA by a list of radices followed by a list of padding sizes: i.e. ([100, 1000, 100, 26, 999],[0, 0, 3, 9, 700]). Users can change this using the --custom_layout option.

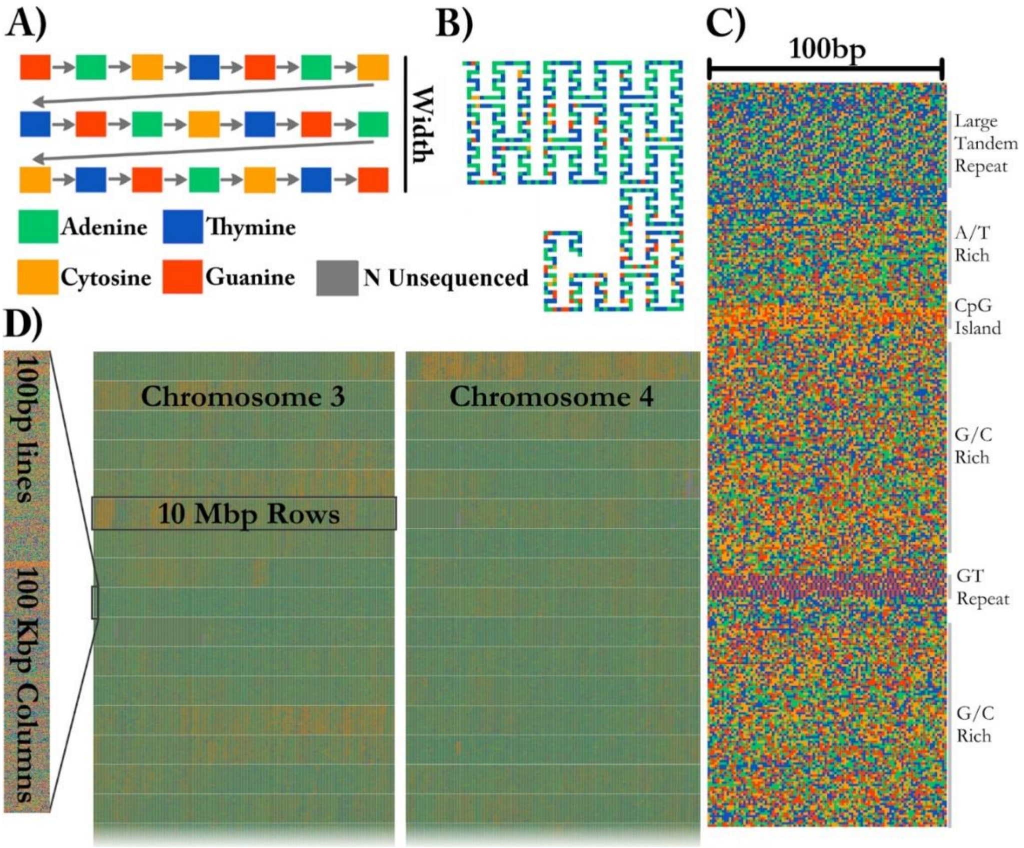


Figure 2.2: **Visualization Method.** A) In the Tile Layout sequence reads left to right like English text. B) The Ideogram Layout uses Peano curves painted with sequence colors. In this example, the radices are x=(3,3,3) y=(5,3,3) and scale=2 to insert whitespace around the curve. FluentDNA uses scale=1, meaning the same path shape is present but there is no whitespace separating disjointed nucleotides. C) The default width of one column is 100bp. Features visible from bare sequence are annotated on the right. D) In the default Tile Layout, 100 x 1000 bp columns are arranged in rows within mega-columns that represent chromosomes.

2.3.3 Ideogram Layout

In Ideogram Layout, FluentDNA depicts the linear DNA sequence as a Peano curve (Figure 3.2B). This has an overall bounding box defining the 2D space filled by the curve, and internal bounding boxes that define how frequently the curve bends. The bounding boxes are defined internally by a set of (x, y) radices (Sagan, 1994).

2.3.4 Pan-And-Zoom Functionality

The basic output of FluentDNA is a single master image file depicting the input DNA sequence. This file is inevitably very large for long sequences, making panning and zooming very memory intensive using direct image viewers. FluentDNA therefore automatically precomputes a “zoom stack” using the DeepZoom library, and sets up a local HTTP server which uses the OpenSeadragon platform (OpenSeadragon; Khouri-Saba et al. 2013) to view the zoom stack as a website using a web-browser. Interactive zooming can be disabled with the --no_webpage command line option. The position of the viewport, combined with the zoom level, generates a small list of tiles to be streamed to the browser. This allows for constant time performance on any device with any size dataset.

2.3.5 Mouseover Algorithm

FluentDNA allows the selection of small sequence snippets in browser using mouse clicks over the image. Users can save 300bp snippets of sequence using a keyboard shortcut which will add the coordinates and sequence to a log. This is often useful for BLAST or manually checking a result. Since the image is not itself a text object, FluentDNA uses an inverse function of each layout transformation to retrieve the original sequence position in the fasta input file and output the snippet's DNA sequence in letter codes.

2.3.6 Annotations

Annotation information from GTF, GFF2 or GFF3 files are visualized by FluentDNA as highlighted sequences within tiled or ideogram layouts, or in an annotation track next to a tiled layout. Currently, VCF and BED annotations are not supported.

Highlighted annotations are painted directly on top of the sequence using lightening, darkening, or outlines. Up to three different annotation files can be rendered with a different appearance. Gene annotations are specified with --ref_annotation and appear as lightened areas of the sequence, with lower opacity for introns and higher opacity for exons. Overlapping annotations are visible as doubly highlighted areas. Particular genes of interest can be highlighted with a drop shadow by specifying a second gene set with --query_annotation. Set intersections are used to detect shadows that collide with other annotated regions so that they can be adjusted to look natural. Repeat annotations specified with --repeat_annotation are rendered as dark regions. Gene name labels are rendered directly onto the rectangular bounding box of the annotated region. Label font size scales up for larger annotation areas. In the Tile Layout, gene labels are always placed at the start of the gene respective of strand: genes on the positive strand have their label at the top of the bounding box, while genes on the negative strand have labels at the bottom of the bounding box. In the Ideogram Layout, gene name labels are placed in the geometric centroid of the annotated nucleotides. The maximum and minimum x and y coordinates are used to determine a bounding rectangle to approximate the size of the gene region. Label font size and opacity is determined in a lookup table so larger genes get larger, more transparent labels painted onto them.

Annotations in a parallel track are depicted as a pseudosequence based on the GFF file. Only one annotation type can be present at any given location in the annotation track, so priority is given in order: CDS, exon, mRNA, gene. The annotation pseudosequence is interlaced side-by-side with the nucleotide sequence columns. As the annotation sequences are less information dense than the DNA sequence, the number of horizontal pixels in the annotation column can be set to a lower value than in the sequence column. The display width of the annotation column can be set with the --annotation_width parameter. When an annotation spans multiple columns, the median point is used to identify the column to fill with a label.

Gene names are rendered and stored as pixels, which leads to a key limitation of FluentDNA's design: the lack of seek functionality. In the current implementation it is not possible to use the name of a gene or other query to jump to a location in a genome. The information needed for this jump is not present in the pixels. However, FluentDNA stores all input files, including annotations in the /sources/ directory for reproducibility. Seek functionality would require javascript access to indexed annotation static files in a manner similar to how Jbrowse serves annotations without the need for a server (Buels et al. 2016).

2.3.7 Whole Genome Alignments

FluentDNA can visualize whole genome alignments when provided with FASTA files for a reference and a query genome, and a liftOver file defining the genome coordinates of regions aligning between the two genomes. The liftOver file must have been previously generated using external whole genome alignment software. FluentDNA generates two gapped sequences from the reference and query genomes, using information from the liftOver file. It outputs a tiled layout with four columns: the reference genome, variants unique to the reference genome, variants unique to the query genome, and finally the query sequence (Figure 2.3). The two middle sequence columns highlight inversions, transpositions and translocations using background color (white: syntenic, blue: intrachromosomal transposition, red: interchromosomal translocation). In this way, differences between the two genomes in terms of SNVs, indels, inversions and translocations are visible at a range of zoom scales. FluentDNA also outputs a table quantifying these differences.

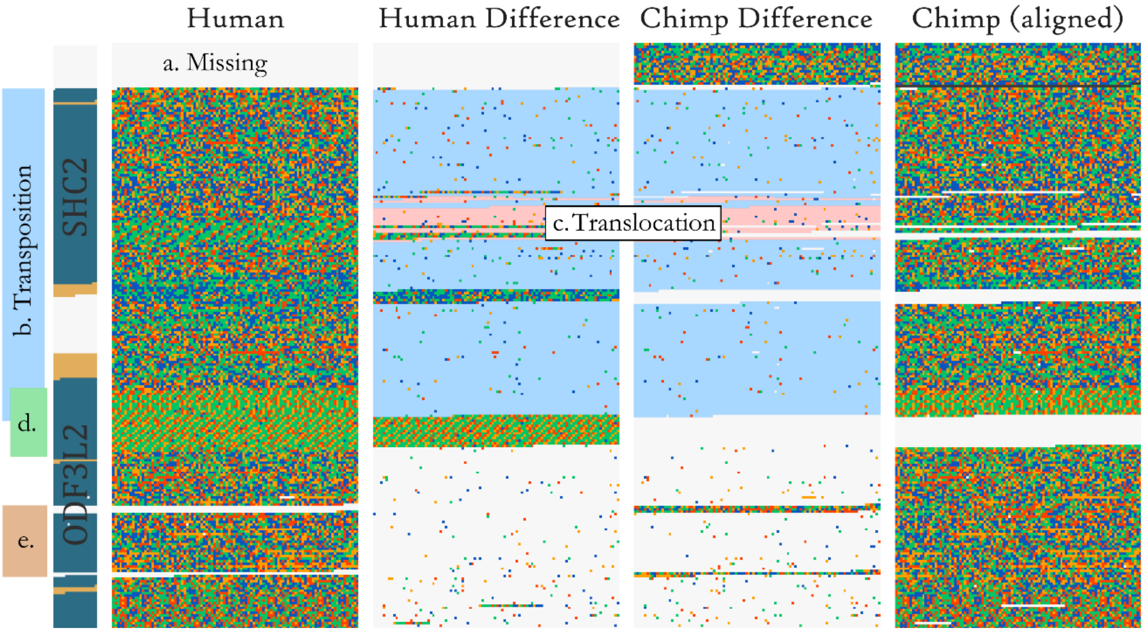


Figure 2.3: Design of An Alignment Visualization. An example from a whole genome alignment visualized in FluentDNA. 25Kbp of Homo sapiens (Hg38) chr19:458,731 and Pan troglodytes (panTro5 2017). From left to right: gene annotation, human sequence, human unique sequence, Chimp unique sequence, and aligned Chimp sequence. Genome elements in the sequence can be seen without an annotation because of changes in nucleotide composition. Simple and tandem repeats appear as a texture. The two center "difference" columns generated by FluentDNA show the differences between the two sequences. Background colors indicate the source of the aligned region: syntenic (white), inversion and transposition (blue), or interchromosomal translocation (red). Two human gene annotations for ODF3L2 and SHC2 appear on the left with blue introns and yellow exons. a) Chimpanzee has a ~1700bp sequence not present in human. b) The blue background indicates a transposition within the same chromosome covering half this figure. Sequence in the center left is human specific, for example the highly A/T rich region that overlaps the end of SHC2's exon annotation. c) A small translocation from another chromosome marked in red. d) The transposition ends in a AAAC tandem repeat where human has twice as many copies as chimpanzee. e) Human and chimpanzee share a syntenic region where Chimpanzee has 300bp and 130bp inserts.

Whole genome alignment liftOver files, such as those available for many species pairs and assembly versions on the UCSC genome download site, contain a list of chain objects defined by a start position and strand in the reference and query genomes. Each chain is a series of entries with a contiguous alignment punctuated by gaps in the query or reference. Where two genomes are assembled to chromosomal level and highly similar, a single chain may cover much of the sequence data for each chromosome. Translocations and inversions introduce new chains. Multiple translocations from the same chromosome in the same orientation may be netted together depending on a distance cutoff. Ideally, a liftOver file will aggregate the alignment into as few chain objects as possible.

In order to turn the list of chains in a liftOver file into a visualization, it is necessary to linearize the alignment, pull in the sequence, and rearrange translocated sequences. FluentDNA sorts all chain entries in a UCSC Chained LiftOver file into a single list on the reference positive strand. The first large chain (referred to as the master chain) is used to establish a shared coordinate frame with the query genome. Other chains are then inserted into position, meaning all chains become intermixed. The reference genome stays in the same order and copy number but gaps may be inserted. The query genome sequence is rearranged to match the ordering and copy number of the reference genome (though if the liftOver file is for a reciprocal best alignment each sequence in the query genome will only be represented once). Each nucleotide index range tracks information about the source sequence: syntenic, intrachromosomal, or interchromosomal. New query sequence is brought in to fill unaligned gaps in the initial master chain alignment until all known alignments are composited into a single visualization.

When the master chain covers a large proportion of the query chromosome, unaligned query sequence is brought in with the master chain, introducing gaps in the reference and allowing the user to see sequence that is unique to the query genome. However, if the master chain covers only a small proportion of the query chromosome (for example because the genomes are highly divergent, or the query genome assembly is highly fragmented), then a design limitation of FluentDNA will become apparent. Little to zero unaligned query sequence can be included in the visualization and few gaps will be introduced into the reference genome. It will thus appear that the query genome is a subset of the reference genome because regions of the query genome that cannot be aligned to the reference genome will not be placed within the visualization. This limitation could only be ameliorated with a search of the remaining alignment file to verify that adjacent sequence was not allocated elsewhere, requiring additional compute.

The background of the columns in the four-column alignment layout are colored to show which query alignments come from the master chain (shown by a white background: these are syntenic alignments), secondary chains with the same chromosome label as the master chain (shown by a blue background: these are normally due to inversions or translocations within a chromosome) and secondary chains with a different chromosome label (shown by a red background: these are normally due to translocations among chromosomes). FluentDNA can also output an image that only shows the nucleotides unique to the reference genome, using the option --layout=unique. The script AlignmentStats.ipynb can be used to aggregate genome alignment statistics for a whole genome.

2.3.8 Phylogenomic Multiple Sequence Alignments

FluentDNA can visualize many multiple sequence alignments (MSA) in a single field of view, such as a set of genes aligned for a phylogenomic study. This allows users to, for example, pick out poorly aligned sequences. This function requires a directory of FASTA files as input. Each file in the directory contains multiple aligned sequences, representing one MSA. The file name is rendered as a text label over the sequence block. Files are either rendered in alphabetical order or in descending count of FASTA entries if --sort_contigs parameter is used. In the rendering engine, each MSA is listed as a separate layout with its own width and height. Mouseover sequence is handled by storing the origin point of each layout in the HTML.

2.3.9 Image Generation For Publications

FluentDNA produces PNG visualizations at different scales for publications. The script Image_resize_script.py allows the user to set the level of magnification for any image output without introducing aliasing artifacts. Vector graphics are a proxy for providing super resolution which removes pixel artifacts around text, curves etc. Since FluentDNA directly provides super resolution through a zooming interface and the export of images of any size, it does not provide vector output as well.

2.3.10 Publishing Results on Public Web Pages

Each genome visualized is stored in /results/ inside of the FluentDNA installation folder. Visualization webpages can be published by placing this folder on any public facing server. No special FluentDNA server is required. For example, a visualization with --outname="HumanHg38", the user would copy the folder results/HumanHg38 to the server then link to HumanHg38/index.html. Javascript runs on the client's machine and downloads for all the source files are available through links to HumanHg38/sources/. Image browsing requires a small amount of traffic per user regardless of the size of the genome. Sequence mouseover generates more server traffic but can be disabled by deleting the /chunks/ directory. Similarly, source downloads can be disabled by deleting the /sources/ directory to protect private data.

2.3.11 Museum Display

FluentDNA can support an interactive museum display allowing visitors to explore a whole genome assembly. A large poster is printed showing a tiled or ideogram image of a whole genome assembly, and overlaid with a touch sensitive screen. A flat screen monitor is built into the display. When visitors touch a point on the genome poster, a zoomed in image of that region is shown on the flat screen, together with annotation information and the DNA letter-code sequence. Detailed instructions for setting up such a display are given in Supplemental 1.

2.4 Results

Here, we show the various outputs of FluentDNA for the latest version of the human genome, and its alignment to the chimpanzee genome. We also show how FluentDNA can be used to make a museum display. The commands to generate these visualizations are available with the published version of this chapter, found here: https://www.frontiersin.org/articles/10.3389/fgene.2020.00292/full. The time and memory required to render specific figures is listed in Table 2.1.

Table 2.1 . FluentDNA Time And Memory Requirements

	Input Size	Time	Memory
HG chr18 Tile Layout without Annotation	80 Mbp	0:01:18	920 MB
HG chr18 Tile Layout with Annotation (Figure 2.4A)	80 Mbp	0:06:00	14 GB

HG chr18 Ideogram with Annotation (Figure 2.4B)	80 Mbp	0:13:16	18 GB
Hg38 whole genome without Annotation	3 Gbp	1:18:54	55 GB
Hg38 whole genome with Annotation (Figure 2.5)	3 Gbp	4:52:22	110 GB
Human-Chimp chr19 Alignment Visualization (Figure 2.3)	58 Mbp	00:08:00	18 GB

Time performance is roughly linear with respect to input data size. Chr18 processed 0.975 Mbp per minute. Rendering a dataset 37.5x larger, the human genome processed 1.16 Mbp per minute. A Highlighted Annotation takes three times as much memory as the nucleotide sequence alone since two images are created to make the overlay image. Compared to Tile Layout, Ideogram Layout takes additional time to compute the Peano curve coordinates.

2.4.1 Visual Analysis of The Human Genome

A tile layout was generated for the Hg38 version human genome assembly chromosome 18 (Hg38) with default settings and highlighted gene annotations (Figure 2.4A). An ideogram (see Figure 2.4B) was made for the same chromosome with highlighted gene annotations for comparison. In both layouts the centromere is clearly visible as a homogenous gene free region. One advantage of the Ideogram Layout is that gene names can be drawn larger than the Tile Layout columns.

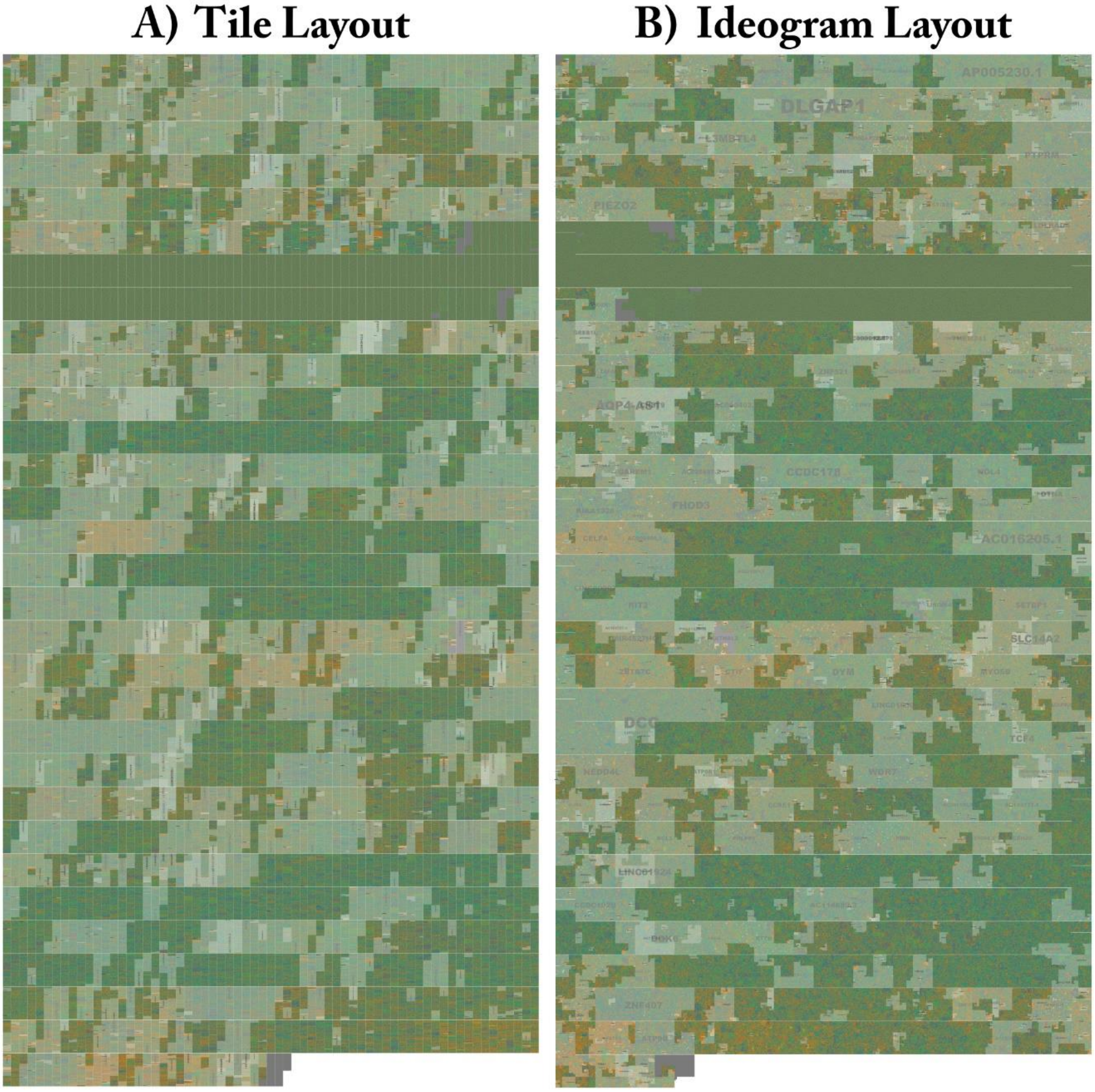
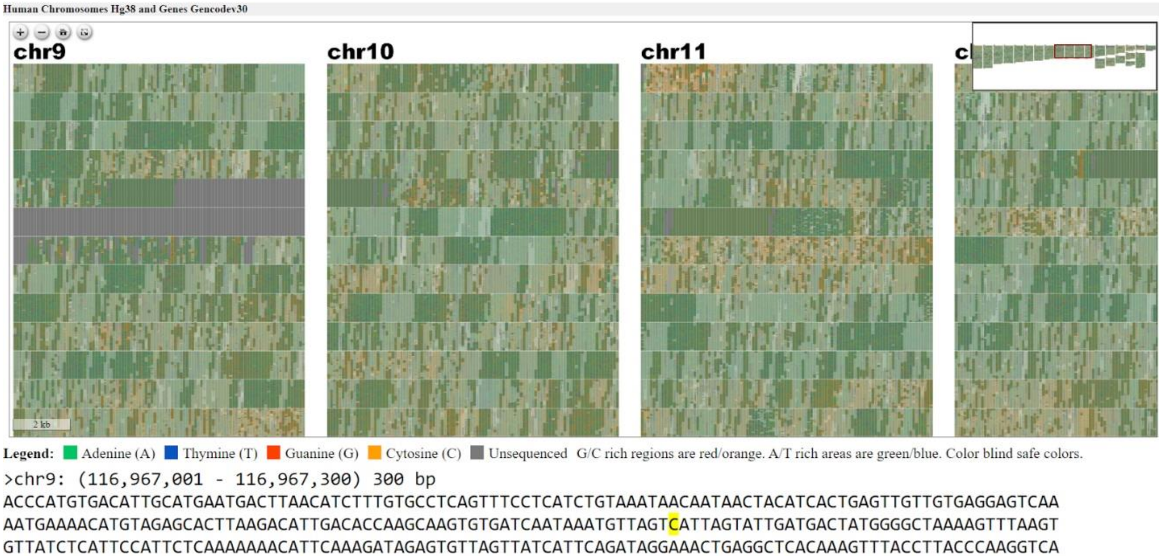


Figure 2.4: Side by Side Comparison of Tiled and Ideogram Layouts. The same sequence is shown in two different layouts. Genes on human Chr18 Tile and ideogram layouts side by side with highlighted annotation. A) The whole structure of human Chr18 with GenCode v30 Genes Highlighted rendered in the Tile layout. Gene labels are drawn in the center and scaled to the size of the region. Centromeres can be clearly seen as a two-row region devoid of gene annotations. Isochores defined by changes in G/C content can be seen as changes in the background color. FluentDNA's zooming interface allows users to see the whole chromosome then zoom in on areas of interest to see smaller features. B) The whole structure of human chromosome 18 is rendered in the Ideogram layout. The gene label DLGAP1 can be seen in the upper chromosome arm. Large gene name labels are drawn with greater transparency so that they can overlap with smaller opque gene labels which may be embedded in the same region. The Peano curve snakes from left to right then right to left and is padded by a small amount of whitespace to mimic a chromatin fiber. Live versions: https://FluentDNA.com/Human_Genome_Hg38_chr18_with_Gencode_v30/ and https://FluentDNA.com/Human_Ideogram_Hg38_chr18_with_Gencode_v30/

For example, DLGAP1 at the top of Figure 2.4B is visible at chromosome scale because of its size. At high magnification, the advantage shifts to Tile Layout. Tandem repeats appear as coherent vertical lines in Tile Layout or are at least recognizable as a diagonal slope. However, in Ideogram they are much more difficult to spot as little more than an unusually homogenous area or quilt pattern. Tile Layout also renders much faster which is important for whole genome renders like Figure 2.5. At the whole genome level, users can see entire chromosome structures as well as prominent features like isochores and gene deserts (Figure 2.5).



Legend: 🟩 Adenine (A) 🟦 Thymine (T) 🟥 Guanine (G) 🟧 Cytosine (C) ⬛ Unsequenced G/C rich regions are red/orange. A/T rich areas are green/blue. Color blind safe colors.

```
>chr9: (116,967,001 - 116,967,300) 300 bp
ACCCATGTGACATTGCATGAATGACTTAACATCTTTGTGCCTCAGTTTCCTCATCTGTAAATAACAATAACTACATCACTGAGTTGTTGTGAGGAGTCAA
AATGAAAACATGTAGAGCACTTAAGACATTGACACCAAGCAAGTGTGATCAATAAATGTTAGTCATTAGTATTGATGACTATGGGGCTAAAAGTTTAAGT
GTTATCTCATTCCATTCTCAAAAAAACATTCAAAGATAGAGTGTTAGTTATCATTCAGATAGGAAACTGAGGCTCACAAAGTTTACCTTACCCAAGGTCA
```

Figure 2.5: Webpage View of FluentDNA Visualization of the entire human genome (hg38) in tiled layout with overlaid gene annotations. Users can zoom in on an element of interest to investigate in more detail. Users can see any sequence currently under the mouse pointer and save 300bp snippets to a log including the scaffold name and position of each snippet. The exact nucleotide under the mouse is shown in a BioJS sequence component (Gómez et al. 2013). The live version with sequence retrieval alongside annotations is available: https://FluentDNA.com/Human_Genome_Hg38_and_Genes_Gencodev30/

In Figure 2.6A, we use FluentDNA to visualize the repeat content of human chromosome 19 using a multiple sequence alignment gallery. RepeatMasker annotation positions downloaded from UCSC were used to extract the sequence for every non-simple repeat from Hg38, clustered by name, and aligned using the repEnd coordinate. This shows several families of LINES all with the same characteristic enrichment in 3' ends. Alu repeats also have a distinctive dimer structure where often only one L or R monomer is found in the genome. The result is equivalent to Figure 2.6B copied from Imbeault et al. (2017) which made it clear L1 has many more copies of the 3' end than the 5' end due to its copying mechanism.

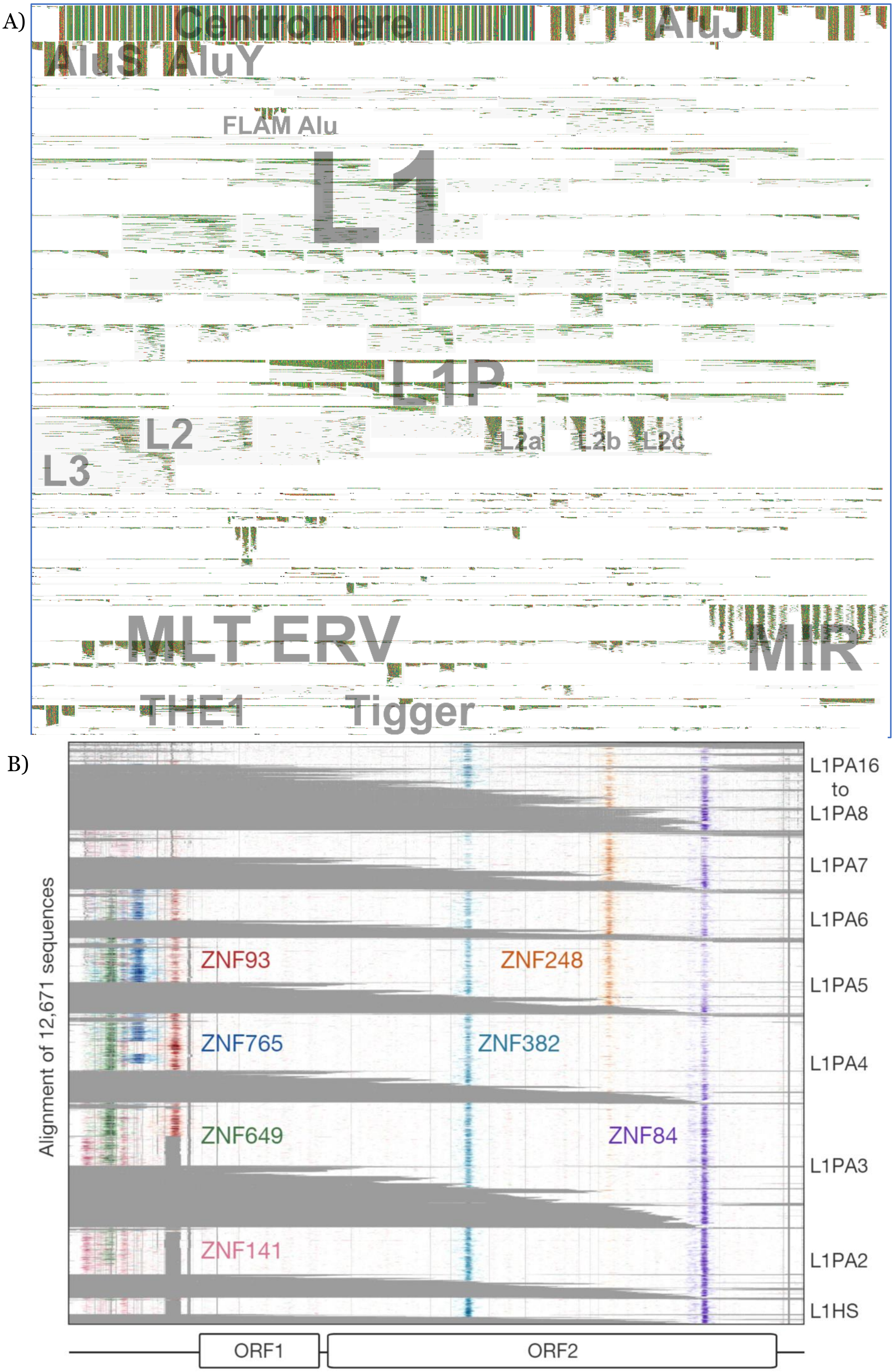

A)
Centromere
AluJ
AluS
AluY
FLAM Alu
L1
L1P
L2
L2a
L2b
L2c
L3
MLT ERV
MIR
THE1
Tigger
B)
Alignment of 12,671 sequences
L1PA16
to
L1PA8
L1PA7
L1PA6
L1PA5
L1PA4
L1PA3
L1PA2
L1HS
ZNF93
ZNF248
ZNF765
ZNF382
ZNF649
ZNF84
ZNF141
ORF1
ORF2

Figure 2.6: Multiple Sequence Alignment Gallery Visualization. A) This figure is a panoramic view of all instances of repeats on Human chr18 annotated by RepeatMasker. FluentDNA adjusts the layout width to match the consensus length of the repeat family. Starting in the Upper Left, major features are ALR centromere, Alu broken into subfamilies. Dominating the middle are long green repeats of L1, followed by the less conserved L2, then a collection of less abundant repeat families. RepeatMasker annotation positions downloaded from UCSC were used to extract the sequence for every non-simple repeat from Hg38, clustered by name, and aligned using only the repEnd coordinate (Kuhn et al. 2013; Smit et al. 2015). Live version: https://FluentDNA.com/Human_Hg38_Chromosome_18_Repeats_-_alphabetical/ B) All copies of L1 from the whole human genome are aligned in this figure copied from Imbeault et al. (2017). Each of the 12,671 horizontal lines is one copy of L1. Zinc Finger protein occupancy is painted onto the individual copies. This visualization clearly illustrates the deletion of the ZNG93 target site in L1PA3, L1PA2, and L1HS. The same preference for L1 3' retention can be seen in both visualizations as a ragged edge on the left side. This figure was the inspiration for the MSA Gallery view, by making it clear it was possible to visualize all sequences, not by their chromosome position, but by their sequence similarity coordinates. FluentDNA's approach favors quantity over quality by displaying all repeat types in a single collage. This could be further refined with preprocessing to group and align similar repeat subfamilies as in Figure 2.6B to improve clarity.

2.4.2 Human and Chimpanzee Comparison

As an example of using FluentDNA for inspecting whole genome alignments, we used (Human (Hg38, Dec. 2013) and Chimpanzee (PanTro6, Jan. 2018) assemblies available at UCSC and their corresponding liftOver file https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToPanTro6.over.chain.gz accessed March 2018. The full browsable alignment at nucleotide resolution is available at: https://FluentDNA.com/Human_Hg38_vs_Chimpanzee_PanTro6/.

We can visually compare across species two chromosomes in the "Alignment" layout. Figure 2.3 shows human chr18 compared with the rest of the chimpanzee genome. We can tell from the white background color in the central two columns that the entire lower chromosome arm is covered by a single syntenic alignment chain, indicating that Chimpanzee has an equivalent syntenic chromosome 18. The upper arm background color is blue, indicating the same chromosome, but not the master chain. This can be caused by an inversion in chimp or, more likely, because the chaining algorithm has not joined chains from the upper and lower chromosome arms. Around the telomeres and centromeres we see smaller regions with a red background: this indicates these alignments are pulled in from other chimpanzee chromosomes. These patches can be due to biological translocations or spurious alignments from another chromosome. Two obvious examples of this are in chunks 14,900,000 and 15,000,000 where regions brought in from other chromosomes show a markedly lower sequence identity in the middle difference columns.

Finally, sequence unique to either the reference genome or query show up as interruptions in the four column layout when zoomed out. This allows users to quickly get a sense of how much the alignment covers and where. Users can zoom in on unique sequences of interest. For example, chunk 30,100,000 contains 50 Kbp of non-repetitive unique human sequence whereas the chunk before 30,000,000 contains 10 Kbp where the aligner simply failed to cover two regions which are visibly similar.

In addition to generating a visualization for each chromosome, FluentDNA calculates alignment statistics to quantify alignment coverage, sequence identity, and the distribution of gap sizes in the alignment (Table 2.2). Since centromeres and unsequenced regions correspond to biological features, calculations including N's and centromeres are listed in parentheses. Initial alignment coverage is 95.57% (90.9%) of the Hg38 reference, and identity within the alignment of 98.65%. We used the "Unique" FluentDNA renderer to show the regions of Hg38 not covered by the alignment (Figure 2.7). This visualization immediately shows that over half the unique human sequence is actually centromere alpha satellite repeat, and sub-centromeric repeats, which are fully sequenced in humans but represented by Ns in PanTro6. FluentDNA allows us to quantify these regions with more customizable precision than a generic repeat-masking would: by visual inspection, we were able to make a custom annotation of centromere and sub-centromeric regions. This allowed us to calculate the total human unique sequence that is attributable to centromere repeats of all kinds. With centromeres excluded, coverage of the chimpanzee alignment is 97.9% of the human genome.

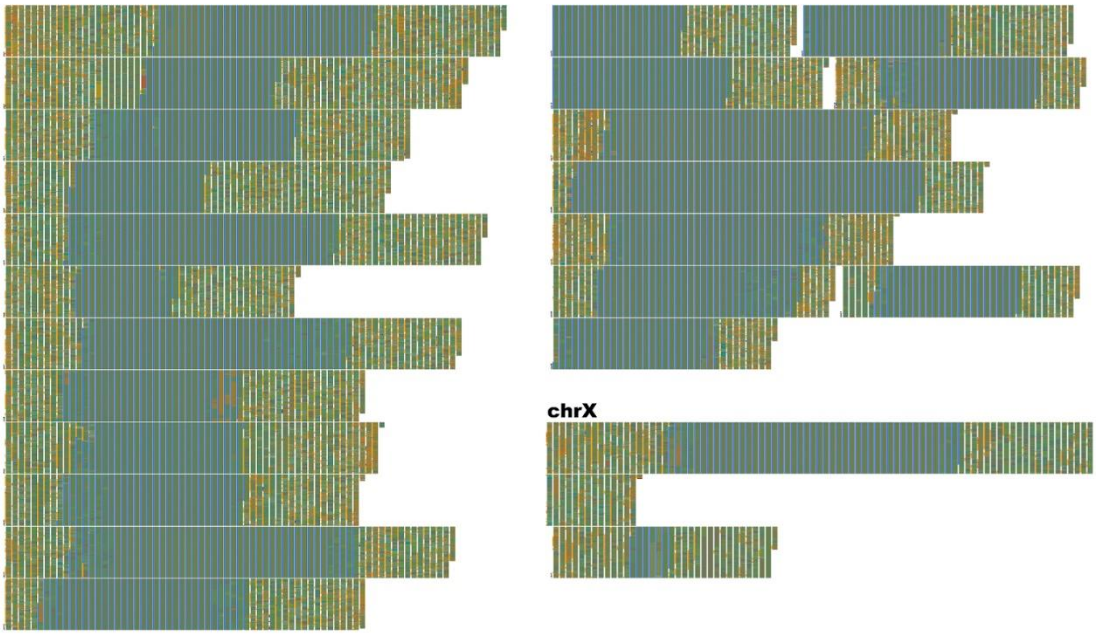


Figure 2.7: Human-Unique Sequence & Annotation Render of Hg38. FluentDNA's Unique layout allows researchers to subtract one genome from another, leaving only the difference for inspection. Human sequence not covered by the alignment between Chimpanzee and human is displayed in chromosome order within two layout pages. The unique portion is 135 MBp, about the same size as chr9. Each visible row is a concatenation of the unique sequence from one chromosome starting with a small label, while chrX takes up more than one row and has a large label. This is because chrX has more human unique sequence than any other chromosome. The grey-blue region in the middle of each chromosome is the sequenced centromere alongside a manual annotation (see Results and Table 2.2). Outside the centromeres, a variety of tandem repeats and non-repeat content are visible. In the annotation track, introns are orange, exons are blue, and CDS is red. The visualization shows approximately half of the human-unique sequence is intronic, while exons are a small minority and CDS unique to human are rarer still. Examples of human specific protein sequence are shown in Table 2.3. Live version: https://FluentDNA.com/Unique_Human_Genes_and_Centromere_vs_Chimpanzee_PanTro6/

Table 2.2.2 Alignment Statistics for Human Genome Hg38 Compared To The Entire Chimpanzee Pantro6 Genome.

Feature	Statistic
Reference Length (N's included) (independent calc)	3,088,269,832
Reference length (No N's) (independent)	2,937,639,113
Total alignment Length	2,807,378,393
Unaligned sequence within reference	149,173,906
Alignment length / Reference length	95.57%
Identical bases within alignment	2,769,610,997
Non-identical bases within alignment	37,767,396
Identical bases / Alignment length	98.65%
Number of gaps introduced in reference by alignment	2,139,409
Ref Gaps larger than 10bp	241,691
Ref Gaps larger than 100bp	54,321
Ref Gaps larger than 1000bp	18,911
Ref N to query bp	150,636,009
Query N to ref in bp	15,139,025
Number of gaps introduced in query by alignment	2,216,928
Query Gaps larger than 10bp	266,110
Query Gaps larger than 100bp	53,587
Query Gaps larger than 1000bp	17,145
Centromeric sequence length (manual annotation)	72,352,500
Reference length minus centromeres	2,865,286,613
Alignment length / Reference length minus centromeres	97.98%
Identical bases / Reference length minus centromeres	96.66%

The top 18 rows are available for any alignment processed with FluentDNA. Using the script Stats_Aggregator.ipynb, which collects statistics across many chromosomes and aggregates them into a summary of a whole genome alignment. Statistics for chromosomal alignment are generated automatically as a single file per reference chromosome by FluentDNA. In the Hg38 versus PanTro6 alignment,“Alignment length / Reference length” is lower than expected because Hg38 has almost fully sequenced centromeres, whereas the PanTro6 has largely unassembled centromeres. Using FluentDNA’s UniqueOnlyChainParser, we rendered the unalignable regions of Hg38 and used the tool to mark the beginning and end of sub-centromeric regions based on sequence. Using these coordinates, we calculated Alignment length / Reference length minus centromeres. Statistics for this manual analysis are shown in the final four rows.

FluentDNA also quantifies the source of aligned sequence for every nucleotide, shown by different background colors in the visualizations (see AlignmentStats.ipynb (Seaman and Buggs 2020)). In the UCSC whole genome alignment of PanTro6 to Hg38, the first chain for each chromosome covers 52.9% of Hg38 excluding Ns and centromeres (49.1% including Ns and centromeres) of the genome; 78.8% (73.1%) is covered by the first two chains and 87.3% (81.0%) is covered by the first three chains. In total, 95.7% (88.8%) of the Hg38 genome is covered by chains derived from the same homologous chromosome in chimpanzee (including chimpanzee chromosomes 2A and 2B as both corresponding to human chromosome 2). Another 2.2% (2%) of Hg38 is covered by chimpanzee chains that are not derived from homologous chromosomes for a total of 97.9% (90.8%).

We can use the visualization to find and explore the context of putative human-specific protein coding sequences. Browsing the webpage for Figure 2.7, one can find rare patches of red in the annotation column, indicating protein coding (CDS) sequence. For example, we identified 8 segments on Chr1 containing unique CDS and used the FluentDNA feature to clip and store each of the sequences (Table 2.3). This log was then submitted as a BLAST query against Hg38 which returns annotated features. Gene functions returned include: amiloride-sensitive sodium channel subunit delta, vascular cell adhesion protein, neuroblastoma breakpoint family member 19, mucin-1 isoform 19 precursor, HHIP-like protein 2 precursor, olfactory receptor 2T10 (Table 2.3). HHIPL2 (https://www.ncbi.nlm.nih.gov/gene/107970260) is related to HOX genes, possibly crucial, and deserves closer scrutiny. This result is caused by a 200bp segment of protein coding DNA in Hg38 that is not covered by the PanTro6 alignment. BLAST searches for this sequence in *Pan troglodytes* returns hits at the expected 98.66% identity, so it is safe to conclude that the sequence is present but the whole genome alignment is imperfect. In contrast, olfactory receptor 2T10 is genuinely missing from *Pan troglodytes* but present in *Gorilla gorilla* and *Pongo abelii* (Seaman and Buggs, 2020, Supplemental File: Hominidae - 7 unique genes from chr1.asn).

Table 2.3 Example Human Exons On Hg38 Chromosome 1 Showing No Alignment With Pantro6

Start position in Ch1 (Fig 9)	300 bp sequence from human CDS showing no chimpanzee alignment	BLAST Chimp	Gene Feature
122,005	GGCCCAGGGTAGGGAGGCCTGAGTGGGTGCAGGCCGGG CCCTGCTGAGGCCACTCTGCACACAGGCTGCAGCCCAG ACGCCCCCCAGGCCGGGGCCACCATCAGCACCACCACC ACCACCCAAGGAGGGGCACCAGGAGGGGCTGGTGGAGC TGCCCGCCTCGTTCCGGGAGCTGCTCACCTTCTTCTGC ACCAATGCCACCATCCACGGCGCCATCCGCCTGGTCTG CTCCCGCGGGAACCGCCTCAAGACGACGTCCTGGGGGC TGCTGTCCCTGGGAGCCCTGGTCGCGCTCTGCTG	present (79% cov.)	amiloride-sensitive sodium channel subunit delta
1,782,205	CAAGAATACAGTTATTTCTGTGAATCCATCCACAAAGC TGCAAGAAGGTGGCTCTGTGACCATGACCTGTTCCAGC GAGGGTCTACCAGCTCCAGAGATTTTCTGGAGTAAGAA ATTAGATAATGGGAATCTACAGCACCTTTCTGGAAATG CAACTCTCACCTTAATTGCTATGAGGATGGAAGATTCT GGAATTTATGTGTGTGAAGGAGTTAATTTGATTGGGAA AAACAGAAAAGAGGTGGAATTAATTGTTCAAGGTGAGT AGAATGTGAAAAAGGAATGATAAAGGTGCTGTCA	missing	vascular cell adhesion protein 1 isoform b precursor
5,892,205	TGAAATCTAGCTGGGGCTGTGTGGTTTCTGATTCCCCC TGGCTTATTCTTTACTTTTTCCCACTTTTCCAGGCTCA GCAGGGAGCTGCTGGATGAGAAAGGGCCTGAAGTCTTG CAGGACTCACTGGATAGATGTTATTCAACTCCTTCAGG TTGTCTTGAACTGACTGACTCATGCCAGCCCTACAGAA GTGCCTTTTACATATTGGAGCAACAGTGTGTTGGCTTG GCTGTTGACATGGATGGTGAGTACCTTTCTATGAAGGT GATAAGGATCCACTGAGTCTTCTGGTTAGGGTCA	present	neuroblastoma breakpoint family member 19
5,966,905	CTGTCCCCAGGTGGCAGCTGAACCTGAAGCTGGTTCCG TGGCCGGGGCCAGAGTGACATCCTGTCCCTGAGTGGTG GAGGAGCCTGAACCGGGGCTGTGGCTGGAGAGTACGCT GCTGGTCATACTCACAGCATTCTTCTCAGTAGAGCTGG GCACTGAACTTCTCTGGGTAGCCGAAGTCTCCTTTTCT CCACCTGGGGTAGAGCTTGCATGACCAGAACCCGTAAC AACTGTTGCGGGTTTAGGGGCTGTGGTAGCTGTAAGAA GTTAAAGTCATAGGGTTGG	present (92% cov)	mucin-1 isoform 19 precursor
7,216,105	GACTTCTGCCAGCTCGCTTCTGCTCTGCTGATGGCCTC ATCCTGCCACTGTGGCTTTTCAGGCTCTTCCTCCTCTT GCCCTGGCGGACGTGGGGCCCCACTCTGGCTTTCTTCT TTGTACCAGGCCCTCGCAATGTATTCTTGCTGCTTGTA GGAGAAGCCAGCTTCTTGGAGGAGCCTTTCTCAGACAA ACCCTGGGCTGGGCCAGAAGCTAAGGTTGCACTGGAAG ATTTTCTAGCAGCTTTCTCTGATTGTTCCTTTAGCAAG TCCAAGACTGTCTCTGAGAAATCAGTATTTATTT	present	HHIP-like protein 2 precursor
7,225,305	CCTGTGATTATCCGAGTTCTAGTAAGCAGAAATCAAAC AACTCTGTACATTTGTTACCTGCTTCATCTTCAGCAAA GGAGATGATGAACTTGCTATGGGTGCTGATCAGCCCTG GGAAGGCACAGGACGTGGTGCTGCCCAGGCAAAGATCC TGCTTCTTCCATTTCTTGTTTTTTCTATCTTCCTGCAA AGCCATAAGTCGACTAGACAAAAAATAAACCCTTATGT TTAGGAATCCATATATCCACTCTGCAGAATACTTTTTC TCCTAGAATCAGAGATCCCTGAGTACTAGGACTG	present	HHIP-like protein 2 precursor
7,711,805	CAGCACTCTGCACCTCCCAACTGCAGGTAGAAGTACAT CTGGGTGCCACACCCAAGGACCGAGATGGTCTTGTCTT TGGCCAGCTGGTTCACCAGCATTTTGGGGACAGTGACA GAAATATATGTCAAGTCTATGAGTGAGAGCTGGTTTAT AAAGAAGTACATGGGAGTATGCAGAGAGGAGTCAATGT GGATCAGAAGTATCAATGTAATATTCCAAGACACAGCC ATCAAAAATATACTGAAGATAAGCAAGCAGAGGCGGCC AGGGT	missing (73% ident homolog)	olfactory receptor 2T10

From the online version of Figure 9, 300 bp regions of human CDS regions with no alignment to chimpanzee sequence were snipped out (Column 2). These were searched for in the GenBank nr database, restricted to hominidae, and showing putative human-specific exons on Chr1.

2.4.3 Poster Images

FluentDNA images are useful for communication of genomic data. For example, Figure 2.8 shows a poster made displaying the entire malaria genome. In which images made using FluentDNA were arranged using desktop publishing software. The legend uses organelle genomes to demonstrate the shape of the Peano curve.

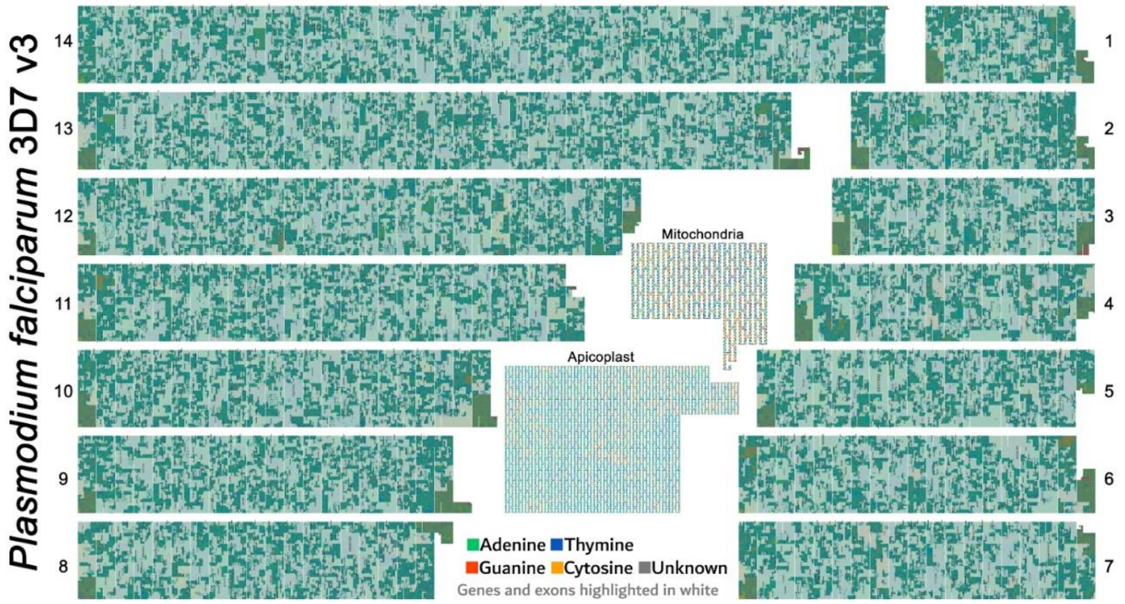


Figure 2.8: Poster Made Using FluentDNA Outputs. The entire malaria genome was rendered as a poster by arranging the 14 chromosomes in descending order of size. The legend uses organelle genomes to demonstrate the shape of the Peano curve by rendering at scale = 2 with whitespace and magnified. Rendered at standard scale, organelles would be tiny and indiscernibly covered in gene annotations. Major visible features include repetitive telomeres at the end of every chromosome. There are no obvious centromeres. Malaria is visibly much more gene dense than Anopheles gambiae or Homo sapiens. While Mitochondria are familiar to every geneticist, the Apicoplast organelle is specific to Phylum Apicomplexa protozoan parasites (Gardner et al., 1991; Egea and Lang-Unnasch, 1996).

2.4.4 Museum Display

The first FluentDNA display was set up in the visitor area of the Millennium Seed Bank as part of *Surviving or Thriving: An exhibition on plants and us* (March 2019 - October 2020). *Arabidopsis thaliana* was selected as the display organism because it is well annotated and has a small genome size. The poster acts as a macro navigation device while the monitor displays the GO Slim functional annotation of the gene as well as the sequence at the position selected (Swarbreck et al. 2008). Using the museum display, it is possible to locate a mitochondrial integration in the centromere of chromosome 2. By touching the visibly orange region (G/C rich) and dragging their finger around, visitors can see genes labeled “mitochondria”, “ATP synthesis”, “Transmembrane electron transport”, etc. Even without detailed knowledge of the technical terms used, every visitor may take away something learned from the display, from the basics of genetic code up to finding clusters of transfer RNA genes. Instructions for creating similar museum displays can be found in Supplemental 1.

2.5 Conclusions

Previous software tools have focused almost exclusively on rendering annotations and markers while the bare sequence is only visible at the smallest scales. We note that FluentDNA is not intended to replace standard genome browsers, but is a useful complement for quality assurance and genome comparison. FluentDNA places emphasis on nucleotides, while placing less emphasis on annotation direction and exon boundaries. Visualization of bare sequence can be informative because gene elements often introduce visible changes in k-mer usage. This is useful in genome assembly for quickly spotting artifacts. FluentDNA is a significant improvement on other direct sequence visualizations (e.g. DDV, DNASkittle) because it can handle multipart FASTA files and scale to viewing entire genomes at once. It also offers a range of capabilities for browsing annotations, protein families and aligned genomes. As a new tool, it does not support every possible file format but extensions are planned, including a VCF render already in development. Finally, FluentDNA allows the creation of posters and museum displays that can make genetic information more accessible to scientists and museum visitors alike.

2.6 Availability of Data and Materials

The FluentDNA software is available for download at https://github.com/josiahseaman/FluentDNA/releases under the Apache 2.0 open source license. The genome data and visualizations in this MS are available at https://FluentDNA.com/.

Chapter 3

Genome Assembly, Annotation, and Gene Families

Author Contribution and Collaborations

I produced new assemblies for 13 ash species and liaised with Dovetail over the *F. pennsylvanica* assembly. I did all visualizations, annotations and alignments of gene families. I visualized a whole genome alignment between the *F. pennsylvanica* and *F. excelsior* assemblies constructed by Carey Metheringham. Laura Kelly, Elizabeth Sollars, and Jasmin Zohren contributed to Ash genome data and assemblies used in this chapter.

Abstract

Comparative genomics is based upon a foundation of quality genome assemblies, gene trees, and alignments. This chapter focuses on improving 13 of the assemblies of the worldwide Ash genome project with new sequence data and generating the gene annotations and alignments necessary for studying convergent evolution in the genus (Chapter 4). My reassembly using 800bp libraries brought median scaffold N50 sizes from 2,997 to 5,583 for an improvement of 86.2%. This draft genome was then used to produce chromosome level scaffolds of *F. pennsylvanica* by Dovetail Genomics using their Hi-C based assembly method. I used FluentDNA visualization (Chapter 3) to investigate the new assemblies, finding several putative bacterial and fungal contaminants. I discovered a nuclear integration of the mitochondrial genome on chromosome 4 of the *F. pennsylvanica* assembly along with a bacterial endophyte *Sphingomonas*. All genomes were annotated using GeMoMa with a *F. excelsior* reference guided model. To facilitate studies of gene copy number evolution, I used OrthoFinder to group genes into gene families. OrthoFinder classified the majority of our *Fraxinus* genomes as diploidized, with 75.8% of genes being single copy per genome. We evaluated and visualized a whole genome alignment between *F. excelsior* and *F. pennsylvanica* to test if a reference-guided assembly method would be practical for further improvement of the 13 species genomes. We ultimately rejected this scaffolding approach due to the divergence between the two genomes and technical limitations of the pairwise alignment format. A superior alignment approach using graph genomes is explored in Chapter 5.

3.1 Introduction

3.1.1 Genome Assemblies

In order to study the repeatability of fractionation following a shared whole genome duplication (WGD), one must first start with a dataset of genome assemblies descended from the same WGD event. The *Fraxinus* worldwide genome project provides such a dataset. The full species list can be found in Table 3.2. The sequencing and assembly of 28 *Fraxinus* species genomes was first carried out and published by Kelly et al. (2019).

Previous sequencing work in *Fraxinus* left the assemblies at differing levels of quality. *Fraxinus pennsylvanica* received the most sequencing attention and is of particular interest because it is economically important in America and is threatened by the Emerald Ash Borer (Poland and McCullough 2006; Kelly et al. 2019). *Fraxinus excelsior*, as the first and highest priority genome assembled, is the draft genome with the largest scaffold size. Out of the 28 draft genomes assembled, there were six “clade exemplars” picked to be higher quality representatives of their respective six *Fraxinus* clades (Table 3.2).

3.1.2 Assembly Quality Control

During the assembly and quality control process, a number of tools can be used to help check results, these tools are outlined below. While the results include tables of assembly statistics, there was no browsable way to inspect the genome sequences themselves for anomalies, either biological or artefactual. Genome browsers are primarily designed to inspect genome annotations and experimental data painted onto the assembly coordinate frame. They do not provide a scalable method to inspect the nucleotide sequence itself across 28 separate 1 Gbp genomes. To this end, a direct sequence visualization was developed, called FluentDNA. Chapter 3 consists of a manuscript published on FluentDNA with detailed methods.

3.1.3 *Fraxinus* Species Tree Without Dates

The most highly-evidenced *Fraxinus* species tree to date used a filtered set of 272 genes found to be in single copy in every species was inferred via Bayesian concordance analysis with BUCKy (Kelly et al. 2019). This can be placed in the larger context of the complete family wide species tree calculated by Zedane (2016) using plastid and rDNA sequences across 136 Oleaceae species (See also section 1.5). Zedane's work included date estimates while Kelly's tree used branch lengths exclusively for concordance values, meaning no date information was estimated. However, Zedane's phylogeny contained fewer *Fraxinus* species and was based on many fewer loci. This study also makes use of Roalson and Roberts (2016) to place a date on *Erythranthe* divergence. The study covered 768 Gesneriaceae species and uses aligned sequence and fossil evidence to create a calibrated species tree.

3.1.4 Gene Annotation

There are many possible approaches for gene prediction depending on the biological question and materials available. *F. excelsior* was first annotated in section 3.4 of Sollars' thesis (2017) and later published in (Sollars et al. 2017) using AUGUSTUS with RNA-seq evidence to identify splice site locations. AUGUSTUS is a machine learning approach that uses generalized hidden markov models to model what gene sequences look like given a wide list of examples in other plant species (Stanke and Morgenstern 2005). When RNA-seq is not available but a closely related species has been previously annotated, a reference guided approach can be used.

GeMoMa is a reference guided gene annotation software that utilizes homology to annotate a target genome (Keilwagen et al. 2016). It was used by Kelly et al (2019) to annotate the 27 other species using *F. excelsior* as a reference genome. GeMoMa uses introns to increase the discovery power of divergent sequences and is particularly well-matched for identifying genes with surrounding sequence similarity from a large scale duplication as opposed to protein motifs conserved due to function, not ancestry (Keilwagen et al. 2016).

3.1.5 Gene Family Boundaries

Gene families are sets of similar genes that are likely to have arisen via duplications of a single ancestral gene. Estimating the size and boundaries of gene families require a judgement call depending on the biological question at hand. To ensure that there are no false negative gene counts, resulting in a lowered gene count in a species, it is best to define gene superfamilies as broadly as possible. However, if the goal is to identify only sets of genes with a high level of confidence then a minimal set will be preferable. If investigating homeologs (paralogous gene copies arising from a whole genome duplication), it is important these genes be in single copy number before the WGD under study lest a few large gene families skew the results.

OrthoFinder (2.2.5.2) is a gene clustering tool which uses a technical definition of a gene family called Orthogroups that is based on the scope of the species tree provided, not on sequence divergence. Orthogroups were defined as the clade of genes descended from a single gene at the time of the last common ancestor (LCA) (Emms and Kelly 2015; Tekaia 2016). An orthogroup could include functional paralogs (1.1) depending on the placement of the LCA.

3.1.6 Whole Genome Alignment

In some cases, a close relative genome assembly can be used as scaffolding to order the contigs of a genome with less coverage or quality. This practice was particularly common when sequencing was much more expensive (Chimpanzee Sequencing and Analysis Consortium 2005). A similar approach was considered in the case of the worldwide Ash genome project using a chromosome level assembly such as *F. pennsylvanica*. Such an approach would first need to verify that *Fraxinus* genomes do not have a prohibitive amount of structural rearrangements. Angiosperms have a much higher tolerance for repeat content and rearrangement than mammalian genomes (Kejnovsky et al. 2009).

3.1.7 Current Work

This chapter describes the improvement of 13 *Fraxinus* genome assemblies using new sequence data from 800bp libraries. In addition, a chromosome level *F. pennsylvanica* genome was assembled in collaboration with Dovetail Genomics using their HiRise/Chicago technology. Inspection in FluentDNA output revealed several surprises (3.3.2). A new gene annotation for all 28 species of the *Fraxinus* worldwide genome project (3.1.1) was reported. Annotation is based *Fraxinus excelsior* genes as a template for any number of highly similar genes in another assembly. These genes were then grouped into gene families, and gene trees were computed based on sequence differences both within species (paralogs) and across all species (orthologs). A whole genome alignment of *F. excelsior* and *F. pennsylvanica* was calculated to determine if there was sufficient structural similarity to justify using a scaffolding approach for other chromosome assemblies. I visualized this alignment using FluentDNA (Chapter 2) to determine the degree of synteny and whether heterozygosity would be enough to disrupt the alignment process.

3.2 Methods

3.2.1 Improving *Fraxinus* Assemblies

The Worldwide *Fraxinus* Genome Project has sequenced 31 genomes representing 28 species of Ash tree. Most are at draft-genome quality level. As this project is dependent on good assembly quality, it was necessary to improve the existing assemblies as much as possible. Additional 800 bp insert libraries were ordered from the Liverpool Center for Genomic Research for the 13 assemblies across 11 species that still lacked this data: *F. pennsylvanica* (FRAX09, FRAX10), *F. quadrangulata, F. ornus, F. mandshurica, F. dipetala, F. angustifolia, F. velutina, F. latifolia, F. paxiana, F. sieboldiana, F.*

caroliniana[1] (FRAX03), and *F. apertisquamifera* (Table 3.2). We selected six species to be exemplars of their clade and performed additional long-mate pair sequencing with 3 Kbp and 10 Kbp insert sizes. The clade exemplars are *F. mandshurica, F. ornus, F. pennsylvanica, F. quadrangulata, F. gooddingii,* and *F. excelsior*.

3.2.1.1 Assembly Process

Assembly methods adopted to improve these genomes are identical to methods described in (Sollars et al. 2017; Kelly et al. 2019). Python automation scripts for a pipeline across the various tools used in genome assembly were developed in order to allow reproducibility. These scripts are available in the DNA_Duplications github repo https://github.research.its.qmul.ac.uk/btx142/DNA_Duplications.

3.2.1.1.1 Trimming

Each genome was sequenced with a target coverage of 44x using Illumina NextSeq and HiSeq platforms. Initial library insert sizes were 300, 350, 500, and 550 bp. The 13 individuals were improved by a second round of sequencing with the addition of an 800bp insert library prepared at University of Liverpool. Paired reads were made from total genomic DNA. Clade exemplars also included 3 Kbp and 10 Kbp long mate pair (LMP) libraries with 125 nucleotides on each side from an Illumina HiSeq 2500 to a depth of c. 10x coverage of 1C genome size. LMP was prepared from sequence from the Centre for Genomic Research and University of Liverpool.

Reads were adapter trimmed and length and quality filtered using FastQC v0.11.5 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were clipped using the fastx_trimmer tool in the FASTX-Toolkit v.0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to remove the first 5-10 nucleotides and the last 5 nucleotides. Cutadapt v.1.8.1 was used to remove adapters with a minimum overlap of five bases on either end (Martin 2011). Sickle v.1.33 was used with parameters: -pe -t sanger -q 20, -l 50 to quality filter read pairs (Joshi, Fass, and Others 2011). Singletons were discarded.

3.2.1.1.2 CLC

De novo assembly of the filtered read pairs, with a minimum read length of 50 bp, was conducted in the CLC Genomics Workbench under the following parameter settings: automatic optimization of word (k-mer) size; maximum size of bubble to resolve = 5,000; minimum contig length = 200bp.

As total genomic DNA was sequenced and assembled, contigs in the assembly include those that originate from the organellar genomes, as well as those from the nuclear genome. All assemblies contained a single contig representing the Illumina PhiX control library. This contig was removed from the assemblies using a Python script provided by Illumina. SSPACE v3.0 was used to join contigs into scaffolds using default parameters. Library insert lengths were given an error range of ±40%. Gaps in the SSPACE scaffolds were filled using GapCloser v1.12 with default parameters. Average library insert lengths from SSPACE outputs were fed into GapCloser. Genome assembly metrics were generated using Assemblathon statistics script (https://github.com/ucdavis-

[1] FRAX03 was originally labeled as *F. caroliniana* but subsequently identified by Eva Wallander as *F. pennsylvanica* (see 2.3.1.2)

bioinformatics/assemblathon2-analysis/blob/master/assemblathon_stats.pl) which includes N50.

3.2.1.1.3 *F. pennsylvanica* Assembly Using HiRise

The above assembly process was applied to make the *F. pennsylvanica* FRAX09 draft genome of comparable quality and short contigs as all other assemblies in this chapter. This assembly was sent to Dovetail along with biological material for processing using their Chicago/HiRise technology built on Hi-C chromatin conformation capture (Lieberman-Aiden et al. 2009; Putnam et al. 2016). Hi-C uses the distance-dependent stochastic nature of cross-linking whole chromosomes followed by sequencing ligated pairs to estimate the distances between loci that may be >10MBp apart. This allowed production of chromosome level scaffolds for *F. pennsylvanica* by placing existing scaffolds in order along the chromosome. The technique only places existing scaffolds in the correct order with N gaps of approximately the correct size. Since each contig is likely separated by a gap, the technique does not improve contig size, but is able to improve scaffold length to the size of pseudochromosomes. *F. pennsylvanica* was the first genome assembly visualized in FluentDNA, since chromosome scale scaffolds make it easier to find structural features. Once visualized, contamination was discovered in the assembly (3.3.2.1). I designed filtering criteria for removing the contaminating reads. Dovetail then built a new assembly using my filtering criteria.

3.2.2 Visualization of Assemblies

A visualization tool was developed to quickly compare different genome assemblies and quality check for anomalies in the process. This tool, FluentDNA, is covered in detail in Chapter 2. FluentDNA turns the four nucleotides into four colors; A: Green, T: Blue, G: Red, C: Gold, N: Grey. Pixels read left to right, just like text, and wraps to the next line at 100bp intervals. Larger scales are wrapped and stacked together in powers of 10: 100Kbp columns, 10Mbp rows, 100Mbp pages, 1.2 Gbp tiles, etc. (Figure 2.2). This allows comparing genomes and gauging size because the layout itself acts as a scale bar. The program uses a satellite image style zooming interface that allows browsing images of 4 GB or more with web technology which can run on a smartphone. FluentDNA can visualize genomes, whole genome alignments, and annotations.

After genome assembly, the full FASTA file, containing one entry for each scaffold (or chromosome) was input as the --fasta= parameter to FluentDNA and visualized using the default Tiled Layout. At this point in the process, gene annotation is not available. FluentDNA is used to visually scan for anomalies such as:

- Anomalous distribution of scaffold sizes, e.g., one large scaffold or a profusion of <10 bp scaffolds.
- Distinctive GC and kmer usage concentrated in a minority of scaffolds. (Organellar DNA as well as contaminants may have very different kmer signatures.)
- Distribution of N's, which can indicate problems in the assembly process, e.g., scaffolds with 100bp at beginning and end separated by >10 Kbp of N's indicate long mate pairs that were never placed inside a larger scaffold.
- Long homo-polymers such as 1 Kbp of T's, potentially caused by jams in the sequencing machine.

- Species with repetitive or gene poor centromeres are visible in bare sequence.
- Tandem repeats including monomers of >1 Kbp which are missed by other tools
- Isochores, chromosome scale changes in GC usage in some species, which are readily visible in chromosome assemblies

After gene annotation, assemblies were re-rendered to check for anomalies in the annotation process, such as:

- Abnormally large genes or overlapping exons.
- Tandem arrays of genes of the same size and composition.
- Areas of chromosomes with no gene annotations, potentially centromeres.
- Annotations that include the beginning or end of the scaffold, which may indicate genes that were truncated by draft genome assembly quality.

Human reasoning and knowledge of expected sequence features is applied to identify any biological surprises. Any such anomalies were followed up with BLASTN (Altschul et al. 1990; Camacho et al. 2009) searches in the selected contigs against the nr (non-redundant) NCBI database and noted in Supplemental File 1: Assembly Inspection Notes.

3.2.3 Constructing A *Fraxinus* Time Tree Using r8s

The species tree developed by Kelly et al. (2019) had a topology but no branch lengths (3.1.3). Here, I augment the species tree with estimated timing for the divergence of each speciation event, called a Time Tree. Time estimates use a combination of calibration points listed in Table 3.1, such as estimated dates of speciation events speciation events and fossil species, plus degree of sequence divergence to estimate branch lengths using the r8s program. The final tree is made ultrametric, meaning that all branch lengths are normalized so every path sums to the same present day time point. In the Orthofinder run, RAxML produces relative branch lengths based on sequence differences in the multiple sequence alignment (MSA). Branch lengths for the tree were calculated using RAxML GAMMA model over 265,591 phylogenetically informative sites from 25,182,399 sites (SpeciesTreeAlignment.fa). These relative branch lengths were then fed into r8s 1.8 to estimate the actual dates using calibration dates taken from the literature (Table 3.1) (Sanderson 2003). The time tree produced by r8s was modified by hand to round dates to the nearest whole integer with a minimum branch length of 1 million years while maintaining the ultrametric property, that is all paths from root to tip sum to 79 million years. The resulting file was saved as Species_tree_corrected_root_ultrametric_integers.tre in Newick format.

Table 3.1 **Calibration Times**

Clade 1	Clade 2	LCA Mya	Source
Fraxinus	*Solanum*	79	Zedane 2016
Fraxinus	*Erythranthe*	72	Roalson 2016*
Jasmium	*Olea*	54	Zedane 2016
Fraxinus	*Olea*	36	Zedane 2016
F. quadrangulata	*F. ornus*	19	Zedane 2016
F. angustifolia	*F. ornus*	14	Zedane 2016
F. quadrangulata	*F. americana*	11	Zedane 2016

Calibration times for the species tree were taken from Zedane (2016, Figure 2.4). This study included plastid and ribosomal genes from Oleaceae species. This source was used exclusively for calibration dates to preserve methodological consistency and to avoid conflicts that often comes from different inferred rates of evolution in different studies.
**Note:* In order to include *Erythranthe* I collated all conflicting date estimates using timetree.org *("TimeTree :: The Timescale of Life" n.d.).* Roalson and Roberts (2016) estimate Oleaceae-*Erythranthe* split (87 Mya) further back in time than the median estimates for Oleaceae-*Solanum* split (79 Mya) (Schneider et al. 2004; Naumann et al. 2013; Barreda et al. 2015). To reconcile for this study, I used Roalson and Roberts (2016)'s estimate which includes both *Erythranthe* and *Solanum* in one study to derive an age ratio. That ratio was then calibrated with the 79 Mya date from Zedane (2016) to ensure it was on the same scale as the rest of the *Fraxinus* dates. Using cross multiplication to maintain proportions, we obtain 79 * 87 / 96 = 71.59 Mya for the last common ancestor of *Erythranthe - Olea*. This method ensures the proportions are maintained even if the absolute rates are less certain in more ancient nodes.

3.2.4 Gene Annotation Methods

Annotation transfer of all *Fraxinus* species using the latest assemblies was carried out with GeMoMa v1.5.0 using *F. excelsior* as reference genome (Keilwagen, Hartung, and Grau 2019; Keilwagen et al. 2016). Settings allow up to 10 prediction models per reference transcript, which is crucial to allow copy number variation for this study. This setting differs from the results in Kelly et al. (2019). The source GFF was Fraxinus_excelsior_38873_TGAC_v2.longestCDStranscript.gff3 (available from http://www.ashgenome.org/transcriptomes) with the longest splice variant for each gene model. File formatting was performed with Extractor with the following parameter settings: v=true f=false r=true Ambiguity=AMBIGUOUS. GeMoMa uses tblastn to index align genome fragments. Tblastn was used with the following parameter settings: -num_threads 24 -db ./blastdb -evalue 1e-5 -outfmt "6 std sallseqid score nident positive gaps ppos qframe sframe qseq sseq qlen slen salltitles" -db_gencode 1 -matrix BLOSUM62 -seg no -word_size 3 -comp_based_stats F -gapopen 11 -gapextend 1 -max_hsps 0.

3.2.4.1 Redundant Annotations

GeMoMa 1.5.0 was run for each assembly and each *F. excelsior* gene model had the potential to create multiple genes. Many different models could annotate the same gene region, especially in the case of large multi-gene families. GeMoMa's tool, GAF, was used to select only the longest non-overlapping transcript for each gene region. Annotations were then quality checked using GAF with default settings.

Genome size was estimated using flow cytometry for each accession before sequencing was performed. These estimates were used as a first pass to exclude polyploids and to calibrate expected coverage. There is no reason to think that *F. excelsior* would be anomalous in genome size or gene density, therefore an anomalously high gene count would indicate reference bias in annotation. Reference bias was checked by comparing the number of annotated genes in *F. excelsior* with all other species. A similar check is carried out in Chapter 4 to measure whether the reference creates an inflated gene retention rate compared with other species.

3.2.5 Gene Family Assignment with OrthoFinder

In this study, 28 *Fraxinus* specimens plus *Olea europaea*, *Erythranthe guttata* and *Solanum lycopersicum* were included as outgroups that do not share the same whole genome duplication events. This gene family definition allows us to capture genes that were in single copy number before the two rounds of WGD in *Olea*.

3.2.5.1. Outgroup Species

OrthoFinder 2.2.6 is a pipeline of tools which allows researchers to study orthologous genes. The input is a proteome for each species. Orthofinder produces orthogroups (a.k.a. gene families), multiple sequence alignments, reconciled gene trees, and lists of orthologous genes using a set of tools. The script translation_from_annotation.sh uses the GeMoMa gene annotations and assemblies to generate a proteome for each species. The proteome for each *Fraxinus* was extracted from the annotations (see 3.2.4) using Cufflinks v.2.2.1 gffread utility (Trapnell et al. 2012). Jalview was used to inspect the protein multiple sequence alignment (MSA) of *Fraxinus* orthogroups to ensure two dissimilar families had not been mistakenly fused together. The Jalview overview window feature was very useful for quickly identifying species with large deletions (Waterhouse et al. 2009).

Outgroup species were included to set the scope of orthogroups. Proteomes were collected from *Solanum lycopersicum,* and *Olea europaea* (Tomato and Consortium 2012; Julca et al. 2017). *Erythranthe guttata* was downloaded from Phytozome 12 using Mguttatus_256_v2.0.protein_primaryTranscriptOnly.fa from the v2.0 of the genome assembly (Hellsten et al. 2013).

3.2.5.2 OrthoFinder Pipeline

The internals of the OrthoFinder pipeline are as follows: DIAMOND uses proteome files to perform an all-against-all sequence search (Buchfink, Xie, and Huson 2015, Figure 1). The MCL clustering algorithm is used to group genes into families generating the Orthogroups.csv file (Enright, Van Dongen, and Ouzounis 2002; Van Dongen 2000). From these orthogroups, mafft 7.310 is used to produce MSAs of the amino acid files which are then converted to gene trees using raxml 8.2.11 (Stamatakis 2014; Katoh and Standley 2013). Unless otherwise noted, parameters used were those contained inside the OrthoFinder pipeline for external tools, which can be found in OrthoFinder's config.json.

OrthoFinder used hybrid overlap and the DLC algorithm to reconcile this species tree with the gene trees generated by sequence alone (Emms and Kelly 2019). The species tree used in this study was created by Laura Kelly using phylogenetically informative genes in an earlier version of GeMoMa annotations published in (Kelly et al. 2019). The protein MSA and reconciled gene trees are used for homeolog identification and analysis in Chapter 4.

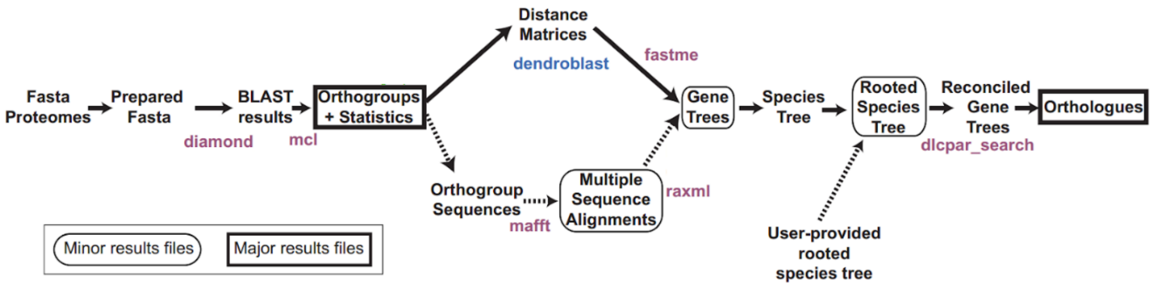


Figure 3.1: Diagram of OrthoFinder: From left to right: Gene membership in gene trees is determined by sequence-driven clustering using DIAMOND and MCL on protein sequences. MAFFT is used to construct one MSA from the protein sequences of each orthogroup. RAxML is used to calculate maximum likelihood gene trees from the MSA. Finally, each gene tree is reconciled with the user-provided species tree (Emms and Kelly 2015, 2019). **Source:** OrthoFinder User's Manual with command line arguments edited out. Used with permission (Emms and Seaman 2020).

3.2.6 LiftOver Whole Genome Alignment of *F. pennsylvanica* and *F. excelsior*

Given 28 related genomes and a chromosome-level assembly of *F. pennsylvanica,* this study constructed a whole genome alignment of *F. excelsior* and *F. pennsylvanica* to determine if there was sufficient structural similarity as well as overlap in alignable regions to justify using *F. pennsylvanica* as a scaffold. Further, I sought to measure how distantly related a *Fraxinus* genome could be before this was no longer a valid approach.

LiftOver whole genome alignment files were generated by following the Methods described in UCSC LiftOver Tutorial (http://genomewiki.ucsc.edu/index.php/LiftOver_Howto). LASTZ with default settings was used to align scaffolds of *F. excelsior* BATG 0.5 to *F. pennsylvanica* (fraxinus_pennsylvanica_26Jul2017_uXjxm.fa) chromosome level assembly (Harris 2007). Each of the 29 *F. pennsylvanica* scaffolds was run against all *F. excelsior* scaffolds and then the results were combined by selecting the best match for each *F. excelsior* scaffold based on total alignment scores (Figure 3.2). This process discards the possibility of duplications existing in *F. pennsylvanica* that are not present in *F. excelsior* but it also removes off-target alignments. This is a crucial step to get an accurate measure of alignment coverage.

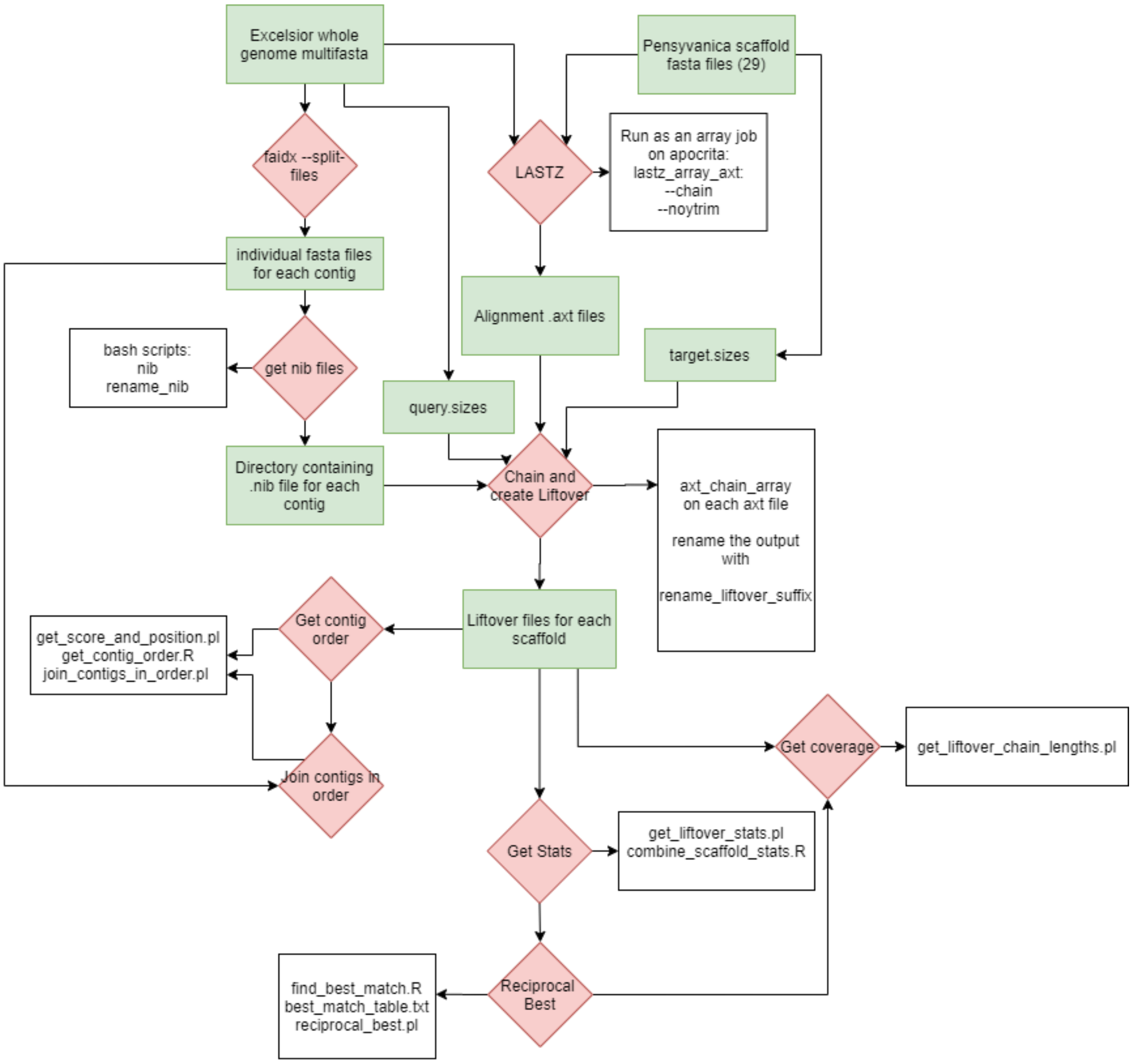

Excelsior whole genome multifasta
Pensyvanica scaffold fasta files (29)
faidx --split-files
LASTZ
Run as an array job on apocrita: lastz_array_axt: --chain --noytrim
individual fasta files for each contig
Alignment .axt files
target.sizes
bash scripts: nib rename_nib
get nib files
query.sizes
Directory containing .nib file for each contig
Chain and create Liftover
axt_chain_array on each axt file
rename the output with
rename_liftover_suffix
get_score_and_position.pl
get_contig_order.R
join_contigs_in_order.pl
Get contig order
Liftover files for each scaffold
Get coverage
get_liftover_chain_lengths.pl
Join contigs in order
Get Stats
get_liftover_stats.pl
combine_scaffold_stats.R
find_best_match.R
best_match_table.txt
reciprocal_best.pl
Reciprocal Best

Figure 3.2: Genome Alignment Pipeline: The whole genome alignment process starts with a series of small alignment seeds from LASTZ. It uses a series of PERL scripts under UCSC Kent Utilities to sort alignments and aggregate them into chains. These chains are netted together into syntenic sections that can be minimally expressed in the LiftOver format. Source: Carey Metheringham 2018 Queen Mary University PhD Progress Report.

3.3 Results

3.3.1 Assemblies and Quality Metrics

3.3.1.1 N50

In all, 13 *Fraxinus* genomes were improved using 800bp inserts including two *Fraxinus pennsylvanica* genomes. The assembled genomes show a wide range of quality. Scaffold N50 (the size of the scaffold necessary to cover 50% of the genome starting with the largest scaffolds) is used as a proxy for quality. After adding the 800bp insert libraries, scaffold N50 improved between 20% and 115% for an average improvement of 60%. This number excludes Dovetail's *F. pennsylvanica* assembly, which improved by a factor of 1,455x. *F. sieboldiana* was the least well assembled diploid with an N50 of 3,085 bp and *F. quadrangulata* was the best non-reference assembly with an N50 of 89,782 bp (Table 3.2). *F. apertisquamifera* and *F. uhdei* were classified as polyploids after sequencing.

Table 3.2 **Global *Fraxinus* species genomes with assembly statistics.**

	Note	Scaffold N50	Improved	Size	Genes
FRAX09 *F. pennsylvanica (Dovetail)*	Exemplar	27,152,721	145420.77%	961,215,495	37,125
FRAX00 *F. excelsior*	Exemplar	103,995		867,496,965	38,949
FRAX11 *F. quadrangulata*	Exemplar	89,782	77.63%	690,460,750	36,439
FRAX06 *F. mandshurica*	Exemplar	41,613	38.05%	851,290,057	36,628
FRAX10 *F. pennsylvanica (PE_48)*		35,102	1588.41%	922,824,354	N/A
FRAX19 *F. goodingii*	Exemplar	26,688		737,681,323	36,410
FRAX07 *F. ornus*	Exemplar	25,551	21.52%	881,712,383	36,925
FRAX31 *F. cuspidata*		16,666		615,746,290	34,490
FRAX27 *F. anomala*		10,033		647,623,079	34,693
FRAX04 *F. dipetala*		8,701	115.26%	621,665,989	33,402
FRAX33 *F. platypoda*		7,689		601,849,709	35,121
FRAX26 *F. albicans*		7,428		704,954,495	35,385
FRAX21 *F. griffithii*		7,160		724,907,964	32,786
FRAX28 *F. baroniana*		6,176		705,684,858	34,856
FRAX32 *F. floribunda*		5,677		700,666,799	34,842
FRAX23 *F. nigra*		5,611		599,092,042	32,819
FRAX01 *F. angustifolia subsp. angustifolia*		5,583	71.05%	744,598,201	34,424
FRAX25 *F. xanthoxyloides*		5,552		612,207,739	31,474
FRAX30 *F. chinensis*		5,515		694,194,979	34,595
FRAX13 *F. velutina*		5,090	69.84%	693,278,749	33,807
FRAX05 *F. latifolia*		5,023	82.32%	789,353,522	35,005
FRAX14 *F. americana*		4,583		641,723,771	32,428
FRAX03 *F. caroliniana*		4,361	54.10%	761,576,543	33,674
FRAX16 *F. angustifolia subsp. syriaca*		4,357		584,583,645	31,464
FRAX20 *F. greggii*		4,303		678,125,092	31,543
FRAX29 *F. bungeana (sp. 1973-6204)*		4,220		908,819,680	35,625
FRAX15 *F. angustifolia subsp. oxycarpa*		3,952		713,783,305	32,452
FRAX08 *F. paxiana*		3,724	40.74%	674,813,152	32,799
FRAX12 *F. sieboldiana*		3,085	55.26%	858,547,996	33,583
FRAX02 *F. apertisquamifera*	Polyploid	2,665	40.56%	1,138,437,068	N/A
FRAX34 *F. uhdei*	Polyploid	2,413		750,530,219	N/A

Highlighted assemblies were improved in this thesis. Clade exemplars were additionally sequenced using long mate pair libraries. The average genome had 34,419 genes. Some gene counts are unavailable because the assemblies were later dropped from the study.

3.3.1.2 Notable Species

Species that appeared to be recent polyploids based on genome size estimates and read heterozygosity were excluded from this study. These species are: *F. apertisquamifera, F. lanuginosa, F. profunda, F. uhdei, F. chinensis* subsp. *rhyncophylla, F. biltmoreana* (syn. *F. americana* var. *biltmoreana*). The remaining 23 genomes are draft quality, with large enough scaffolds to identify gene presence but not chromosome structure. Species names with subspecies and varietals can become unwieldy as filenames, so all assemblies were labeled FRAX00 through to FRAX34 for the purposes of consistency (Table 3.2).

When selecting representative *Fraxinus* species, polyploids were identified based on estimated genome size and excluded due to additional difficulties in assembly. After sequencing, we updated the classification of several species. Sequenced individuals were identified by Eva Wallander using ITS sequence data and morphology based on her previous classification methodology of the genus (Wallander 2012). *F. apertisquamifera* was excluded as a likely hybrid based on its heterozygosity, and on the basis of preliminary phylogenetic analyses using ITS and plastid genome sequence data, which suggest it may be a diploid hybrid between section *Ornus* and section *Fraxinus*. *F. uhdei* was later excluded from analysis as a polyploid after assembly. There are a few special considerations for assemblies. FRAX09 and FRAX10 are both *F. pennsylvanica* that are biological replicates taken from different individuals (accessions). FRAX03 was originally labeled as *F. caroliniana* but subsequently identified by Eva Wallander as *F. pennsylvanica*. Our sample of *F. bungeana* was determined to be a probable hybrid between the *F. ornus* lineage and another lineage in the same sect, and so was designated as *Fraxinus* sp. 1973-6204.

3.3.2 Visual Inspection of *Fraxinus pennsylvanica* Reveals Endophyte and Nuclear Integration

The Chicago/HiRise technique employed by Dovetail genomics placed all our existing scaffolds in their estimated positions along the chromosomes with N gaps of approximately correct size between each scaffold. This did not significantly improve contig N50, but scaffold N50 went from 18.6 kbp to 27,152 kbp (whole chromosomes). The assembly contains 29 megabase scale scaffolds which roughly correspond to the ash tree's 23 haploid chromosomes (Figure 3.3). Using an optical map alignment, the orientation of the six chromosomes with two scaffolds were placed by hand, while the other 17 chromosomes are represented by a single scaffold. Visual inspection revealed two notable features: a mitochondrial integration and a bacterial endophyte.

3.3.2.1 *Sphingomonas* Discovered as *F. pennsylvanica* Endophyte

FluentDNA was used to inspect the genome assembly before annotation. One GC rich scaffold immediately stands out as high-coverage, large, and a clear outlier (also 3.3.2.2). Smaller GC rich contigs that do not match the *F. pennsylvanica* nuclear background sequence model appear to come from the same source distribution of kmers. Using BLAST, the scaffold matches best to *Sphingomonas* sp. LK11 genome (Asaf et al. 2018). The scaffold covers 47% of the genome at 80% identity. This may be sufficient divergence to classify it as a different species than LK11, but further study is required. *Sphingomonas* has been observed as a plant endophyte in crops associated with improved growth and salinity tolerance (Ottesen et al. 2013; Halo et al. 2015). Using these observations, *Sphingomonas* contigs could be prefiltered by kmer usage into a separate sequencing pipeline before HiRise scaffolding to obtain an assembly without mixing the two sources.

3.3.2.2 Contaminant Survey of Fraxinus Using FluentDNA

Next, I used visual checks of other *Fraxinus* assemblies to see if *F. pennsylvanica* contamination was unusual. Other contaminants could be identified because of their varying kmer composition and searched for in BLAST. Soil bacteria *Delftia acidovorans* was present in *F. velutina, F. quadrangulata, F. paxiana,* and *F. angustifolia* (Olm et al. 2017). *Cryptococcus neoformans*, a fungal aerobe that lives in plants and animals was present in *F. sieboldiana (Buchanan and Murphy 1998)*. These contaminants are mentioned here in case it is later discovered there are comorbidities with Ash pests or microbiome resistance. Further notes including sequences used in BLAST are available in S2: Assembly Inspection Notes. FluentDNA revealed that scaffolds containing organellar sequence were typically among the largest scaffolds and relatively easy to identify by hand, however these scaffold sizes are still too small to indicate whether they are connected to the nuclear genome.

3.3.2.3 Mitochondrial Genome Integration in Nuclear Chromosome 4

Fraxinus pennsylvanica chromosome 4 has a 50kbp region that lacks N gaps and has higher GC content (*Fraxinus*: 0.36 Mitochondria: 0.44)[2]. BLAST results matched to "Hesperelaea palmeri voucher E. Palmer 81 (MO) mitochondrion", an extinct member of the same Oleaceae family (Moran 1996). Visual inspection of the contig showed clear boundaries in kmer usage between mitochondrial and nuclear sequence, but no corresponding N gaps that would indicate an assembly error. FluentDNA was used to manually pick locations for PCR primers that spanned both styles of kmer usage. PCR tests indicated that the sequences were present in extracted DNA and thus the nuclear integration was a biological reality. Additional reads from actual mitochondria would still map to a nuclear integration during assembly, leading to an anomalously high coverage in the nuclear genome.

[2] In FluentDNA user trials using *F. pennsylvanica* assembly, the majority of naïve users were able to spot the outlier mitochondria sequence unaided within one minute.

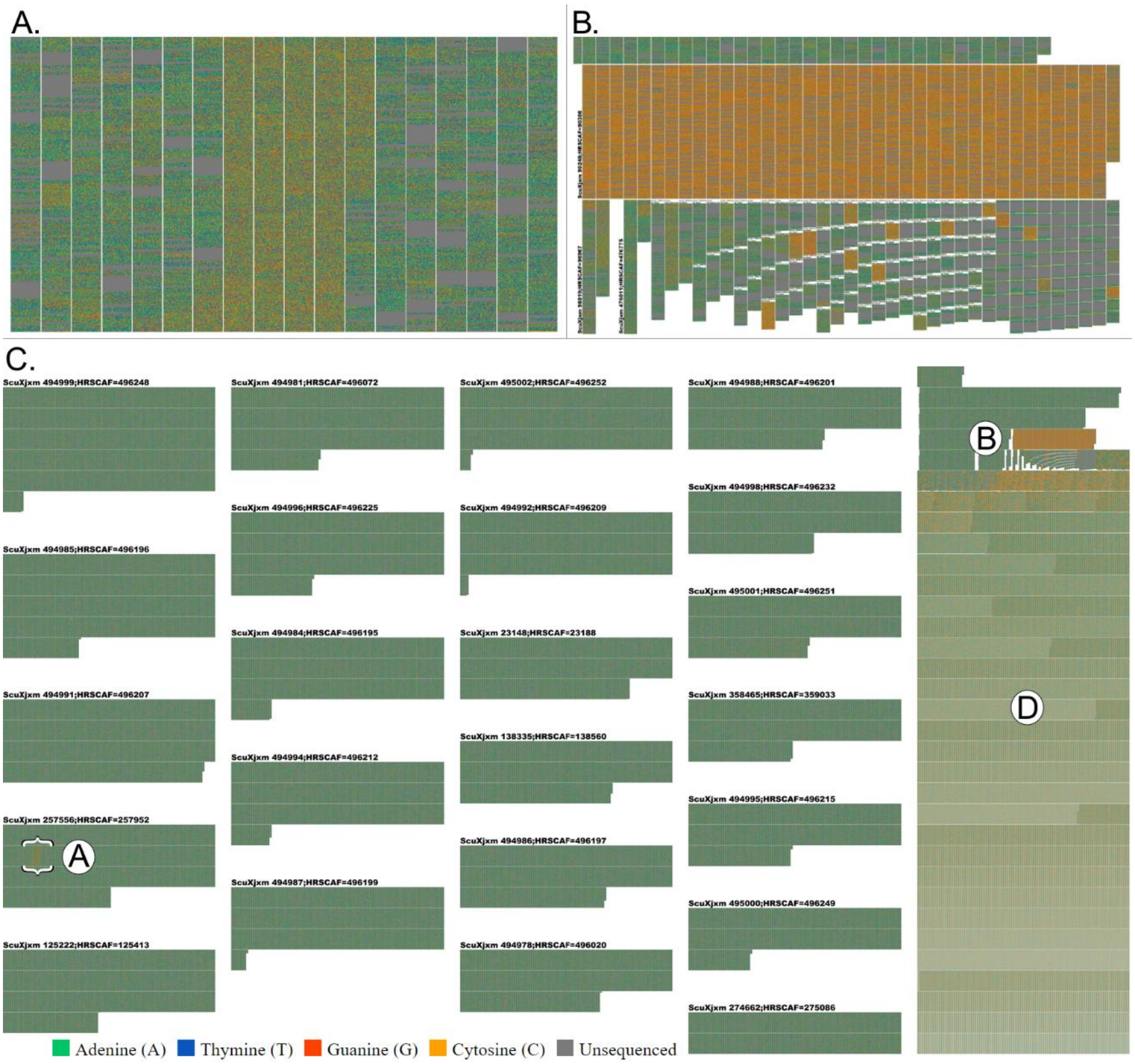


Figure 3.3: *Fraxinus pennsylvanica* Assembly: Visualization of whole *F. pennsylvanica* genome assembly using FluentDNA (Figure 2.2). The genome scaffolded using HiRise technology yielded chromosome-level scaffolds seen in this figure as 29 large rectangular bars of mixed color. A) A nuclear integration of the mitochondrial genome was found on *F. pennsylvanica* chromosome 4. It can be seen on the left as a slightly higher G/C content and lack of N gaps region due to the higher coverage. B) The genome of a newly discovered endophyte related to *Sphingomonas* can be seen as a 3.87Mbp scaffold with a high G/C content (red and gold). C) Whole *F. pennsylvanica* genome with A and B positions marked. D) Unscaffolded short reads are shown at the end (lower right). Of the genome, 23.5% is unassembled contigs, 76.5% is ostensibly assembled into chromosomes. This figure was generated with FluentDNA, a new visualization tool developed in Chapter 2. An interactive visualization is available at https://fluentdna.com/archive/Fraxinus%20pennsylvanica%20June%202017%20sorted/

3.3.3 *Fraxinus* Time Tree with Dates

We can make several notable observations from the calibrated time tree (figure 3.4) of *Fraxinus*. The first, most outstanding, feature is that *F. cuspidata* is an extreme outlier in genus *Fraxinus*. *Olea europea* split from *F. excelsior* 36 million years ago but *F. cuspidata* split only 34 million years ago. The rest of genus *Fraxinus* share 15 million years of history with *F. excelsior* before the next speciation event. Several of the species with questionable identification can be seen next to each other on the time tree. *F. caroliniana* clusters with *F. pennsylvanica*, while the three *F. angustifolia* individuals group together.

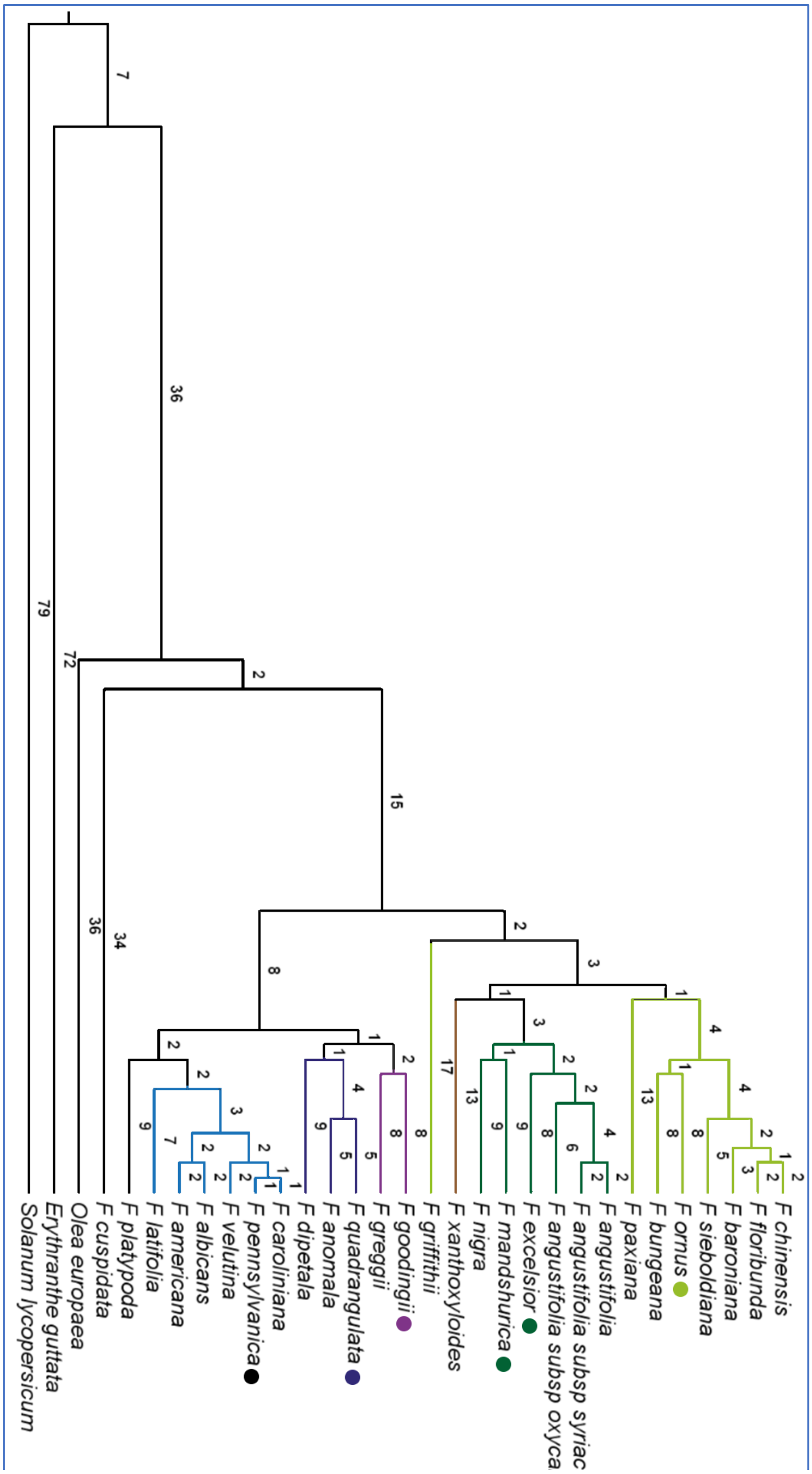


Figure 3.4: *Fraxinus* Time Tree with Dates: Speciation dates were inferred using RAxML and R8s with aligned sequence evidence. Branch lengths indicate millions of years between branch nodes. Branch colors indicate *Fraxinus* clades defined by Wallander (2012) and represented by clade exemplars (colored circles). Clades are:section Dipetalae (dark blue), section *Fraxinus (*dark green), section *Melioides* (light blue), section *Ornus* (light green), section *Pauciflorae* (purple), section *Sciadanthus* (brown). All outgroups and unplaced species are colored black.

3.3.4 Gene Family Results

The number of genes annotated in the original reference-based annotations varied from 99.2% - 99.55% of *F. excelsior* gene counts in Kelly et al. (2019). All genomes had approximately the same number of GeMoMa annotated genes, varying by 0.35%, despite varying genome sizes, morphological differences, and tens of millions of years of evolutionary divergence during a diploidization process that reduced genome sizes by nearly half (1.5.2). This is a higher level of annotation concurrence than is seen on the same individual under two different annotation runs with different settings (data not shown). One explanation is the high level of concurrence is an artifact of the method in essence requiring a one to one annotation of the non-reference genome with each *F. excelsior* gene, resulting in identical gene count.

Every gene has multiple isoforms, the gene count is based on taking only the longest isoform of each gene. However in Kelly et al. (2019) the genes are allowed to overlap, even in the same frame. If *F. excelsior* has fewer copies than *F. cuspidata* then true *F. cuspidata* genes will remain unannotated since each gene is only allowed one copy. If, however, *F. excelsior* has more copies than *F. cuspidata* then multiple overlapping annotations with slight variations will be mapped to the same *F. cuspidata* region. Either way, one annotation is generated for each *F. excelsior* gene. The only outcome which does not generate an identical annotation count is when an individual has lost an entire gene family and its associated pseudogene sequence which is present in *F. excelsior*.

More varied gene counts were acquired by allowing each *F. excelsior* gene to template multiple gene copies then excluding overlapping gene annotations (3.2.4). The new genome annotation performed on all 28 specimens here shows gene counts between 81.2% - 95.8% of *F. excelsior* gene count (Table 3.2). These results show 1) a more biologically realistic range of gene counts and 2) a reference bias in the annotation where the reference genome still has the highest gene count. The gene families which were annotated in all 28 species were the subset of families present in *F. excelsior*. This reference bias is unavoidable due to lack of availability of RNA-seq data for other specimens.

GeMoMa annotated an average of 34,419 genes per *Fraxinus* species (Table 3.2). OrthoFinder grouped these into 28,362 Orthogroups, meaning that for any given genome 75.88% of annotated genes are only present as a single copy. For each species, the average percentage of Orthogroups containing that species is 70.3%. Species specific orthogroups were 0.1% of all orthogroups (Table 3.3).

Table 3.3 **Orthogroup Results**

Number of genes	1,153,334
Number of genes in orthogroups	1,117,046
Number of unassigned genes	36,288
Percentage of genes in orthogroups	96.9%
Percentage of unassigned genes	3.1%
Number of orthogroups	28,362
Number of species-specific orthogroups	243
Number of genes in species-specific orthogroups	1,439
Percentage of genes in species-specific orthogroups	0.1%
Mean orthogroup size	39.4
Median orthogroup size	32
G50 (assigned genes)	60
G50 (all genes)	59
O50 (assigned genes)	6,257
O50 (all genes)	6,561
Number of orthogroups with all species present	6,609
Number of single-copy orthogroups	309
Date	2018-06-25

Statistics generated by OrthoFinder after clustering genes from 28 *Fraxinus* genomes plus 3 outgroup species (*Olea europaea, Erythranthe guttata,* and *Solanum lycopersicum*) into orthogroups.

3.3.5 Whole Genome Alignment shows significant shuffling and highlights technical hurdles

The calculated whole genome alignment covered 62.3% of the *F. pennsylvanica* genome and 52.6% of the *F. excelsior* genome. Only 54.9% of *F. excelsior* contigs have an alignment mapping, but those contigs account for 95.9% of the total genome length and 97.8% of its predicted genes. This indicates the unaligned scaffolds are mostly short and do not contain genes (Figure 3.3D). We may infer from this correlation that the intragenic regions are more diverged between the two species. Figure 3.5 shows an example alignment visualization.

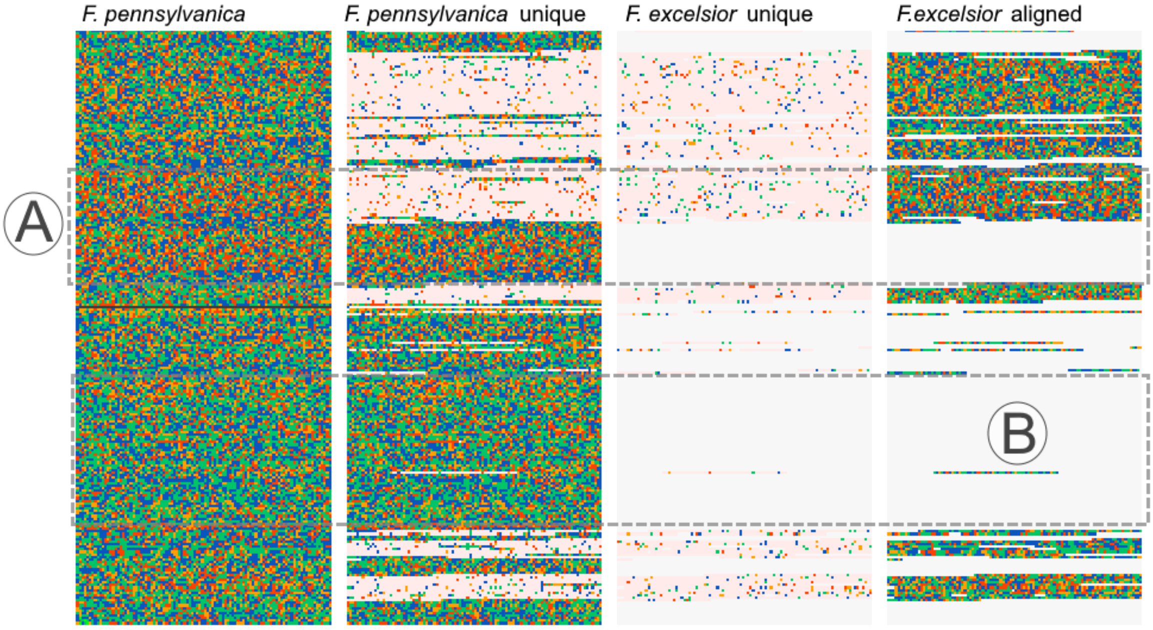


Figure 3.5: FluentDNA is used to inspect an alignment generated between *F. pennsylvanica* (left) and *F. excelsior* (right) visualized with FluentDNA. The center left column shows *F. pennsylvanica* specific sequence while the center right column shows *F. excelsior* specific sequence. A) Tandem duplication of 3,000 bp in *F. pennsylvanica*. The aligner finds a similar region in *F. excelsior* (right) but does not match it to the second copy as well. B) Region that did not align to *F. excelsior*. See Figure 2.3.6 for another example of this type of comparison.

3.4 Discussion

3.4.1 Genome Assembly

The inclusion of 800bp insert libraries improved all 13 genome assemblies that represented a wide range of assembly qualities. By scaffold N50, the worst and best assembly quality improved from 1,987 - 50,545 to 3,085 - 89,782. No amount of parameter changes were able to overcome the inherent lack of long mate pair libraries or low coverage in problematic areas in order to obtain chromosome-level assemblies. In contrast, the Chicago / HiRise method uses polymer physics plus extensive sequencing to reveal which sequences are located physically close to each other. This technique was capable of generating chromosome-level assemblies where no amount of *in silico* bioinformatics could. I will be recommending this approach in the future.

3.4.2 Visualization

FluentDNA assisted in the assembly quality control process by making the unknown known. While it was not as fast or concise as a readout listing the percentage of N's in the genome, it did provide nuance between two scenarios that would otherwise have identical statistics. Scenario A: N's are in large scaffolds after successfully joining many contigs (*F. mandshurica*), or Scenario B: N gaps are mainly from orphaned long mate pairs (LMP) that were never successfully chained into other contigs (*F. quadrangulata)*. An LMP comes with a built-in N gap insert size that doesn't carry any real information about the contiguity of the assembly or genome size of the organism if it's not placed. To the author's knowledge there are no tools that automatically annotate these two different outcomes. Once the use case is identified, it's relatively trivial to write a script to quantify the number of unplaced LMP scaffolds and remove them from the N content statistics. Visualization makes these invisible possibilities apparent (3.4.2.2, 6.1).

In most cases, the largest contigs were mitochondria, which are readily identifiable with BLAST. However, in the case of *F. pennsylvanica* it was possible to detect a true nuclear integration of the mitochondrial genome that is also likely to have aggregated mitochondrial reads, creating abnormally high coverage. FluentDNA was useful in picking out specific sequences to test in the context of GC changes and N gaps. This particular use case could be fulfilled by a dynamically zooming line graph of GC content.

3.4.2.1 Contamination

In almost every single genome, I was able to detect bacterial or fungal contaminants within two minutes of visualization due to the extreme differences in GC content between the genomes from different kingdoms. BlobTools offers a similar functionality by plotting contigs with GC content on one axis and read coverage on the other axis (Laetsch and Blaxter 2017). This fulfills one job that FluentDNA does but at higher sensitivity; however, FluentDNA can fulfill a broader range of tasks. The majority of user time was spent copying sequence snippets from FluentDNA into BLAST and waiting for results. Deep integration of sequence search and other assembly support tools into FluentDNA would enable users to rapidly interrogate and annotate their assemblies, particularly for identifying organelles, contaminants, and repeat families (Maaten and Hinton 2008; Crusoe et al. 2015; Laetsch and Blaxter 2017; Smit et al. 2015; Procter et al. 2021; Buchfink et al. 2015).

3.4.2.2 Advantages of Visualization Tools

Elhai (2011) discussed the importance of a capacity for surprise in biological research and pointed to bioinformatic black boxes as the primary problem area. He argues that, just as an explorer must maintain a deep awareness of their surroundings, a researcher must maintain a deep awareness of the inner workings of their wet lab and bioinformatic procedures. In this regard, abstractions can be misleading representations of the biological reality. FluentDNA and other visualizations bypass these black boxes by involving the researcher in every stage, staying close to the raw data, and taking advantage of the human capacity for creating explanations by combining biological knowledge with new observations.

A key attribute of using FluentDNA to explore new genome assemblies is that it maintains the possibility to show the researcher something unexpected. Scripts and rigid analysis tools can only show what is already anticipated and accounted for in the procedural recipe. Researchers working in wet labs or in the field use their senses in feedback loops to constantly inform and update what they are working on. A key weakness of automation is its tendency to blindly carry on after the situation has gone outside of the original intended task. It cannot adapt to information not already explicitly within its programming. Researchers, by necessity, are searching for new information and new connections. They are searching for surprise. Surprises are best recognized in rich, high bandwidth data.

3.4.4 Annotation & Gene Families

Within one species genome, the majority of genes do not cluster into a gene family with other paralogous copies. In this multispecies study, 96.9% of all genes clustered into an orthogroup with members of another species. Some of this may be due to the GeMoMa reference-based annotation method being more likely to find similar genes across species. Without the additional evidence from transcriptomes, we will be less likely to annotate novel genes specific to one *Fraxinus* species that lack homologs in *F. excelsior*. Since the *F. excelsior* gene count is the highest of any genome, caution will need to be taken in chapter 4 when interpreting a corresponding lower rate of gene loss inferred since the last common ancestor for all *Fraxinus*.

3.4.5 Moving Beyond Whole Genome Alignment Technical Challenges

Performing whole genome alignment was an important prerequisite to move our comparative genomics beyond gene families to whole genome assemblies. Unfortunately, a truly syntenic contiguous genome alignment has remained out of reach for technical reasons that have gone unaddressed for decades. There are a few factors that may have led to this neglected area of research between 2004 Chimpanzee genome alignment and 2018 update of UCSC alignment protocols. First, academic publishing has a known bias against publishing negative results, so it is possible the continual failure to produce satisfactory whole genome alignments has gone unpublished (Ioannidis et al. 2014). Second, intergenic regions are less conserved across species, more difficult to align, and also of less biological interest. A simple invocation of OrthoFinder or similar will give a researcher aligned protein families or cDNA alignments relevant to most research questions. Third, an automated pipeline for aligning genomes is most useful to a genomic center with many genomes such as EBI, NCBI or UCSC. Of those, UCSC has decided to focus on comparative genomics as a service rather than releasing software for researchers to run at their own institutions.

Researchers who decide to compute their own genome alignments will face significant technical challenges. The draft alignments of *F. pennsylvanica* and *F. excelsior* used in this thesis, for example, took Carey Metheringham several months of work writing scripts and sorting through old documentation (Metheringham 2018):

> “The task of aligning two genomes took significantly longer than anticipated. This was partly due to my unfamiliarity with the methods and difficulties in managing the time and memory requirements of LAST and LASTZ, and in part due to the lack of support available for alignment of partially assembled genomes. Alignment tools tend to focus either on mapping a large number of short reads to a reference or aligning two fully assembled genomes. In this case the *F. excelsior* genome was divided across 89,513 contigs and *F. pennsylvanica* across 243,852 scaffolds. While the LASTZ algorithm can perform alignments with multiple sequences as its query or target, using multiple target sequences produces an exponential increase in runtime. In order to bring run time under the cluster’s threshold, only *F. pennsylvanica* genes that had been placed on the linkage map were used as the target genome. This may have resulted in a loss of alignment coverage as parts of the *F. excelsior* genome would align to the unplaced *F. pennsylvanica* scaffolds, however alignment to the smaller unplaced scaffolds would provide limited additional information and was deemed not to be worth the computational cost. Many of the delays in the alignment process were due to difficulties in mismatched file formats and difficulties in conversion between formats. Converting the output of LASTZ (maf or axt files) into a LiftOver file was not a well-supported process and was once again hampered by the large number of files in use.”

As the rate of sequencing continues to outpace researchers' rate of understanding, comparative genomics will become ever more important. Bioinformatics depends on good alignment protocols and formats. As of 2018, these whole genome alignment methods were superseded inside UCSC by newer scripts and a tool that allows researchers to select genomes hosted on UCSC for alignment. While this tool makes LiftOver generation more user-friendly, the ultimate goal of this study has been to generate an alignment of 28 *Fraxinus* genomes that had not been uploaded to UCSC. LiftOver files are still inherently pair-based. In Chapter 5 of this thesis, I discuss a much better method of representing pangenome alignments using Graph Genomes.

3.5 Data Availability

All newly assembled *Fraxinus* genomes were deposited in EBI Nucleotide Archive which is available through the BioProject PRJEB20151 (https://www.ebi.ac.uk/ena/browser/view/PRJEB20151). These are placed alongside the first draft assemblies from Kelly et al. (2019) where improved scaffolding is shown in Table 3.2. The BioProject does not currently include the latest *F. pennsylvanica* Dovetail assembly, which will receive its own publication.

All other data associated with this chapter is available on Zenodo, here: https://zenodo.org/record/4302552

Chapter 4
Repeatability of Evolution: Analyzing Patterns of Gene Loss after an Ancient Whole Genome Duplication in Ash Trees

Author Contributions and Collaborations

Josiah Seaman wrote the entire text of this chapter. Endymion Cooper wrote the original concept for this project in 2014 as a Marie Curie grant. Elizabeth Sollars, Jasmin Zohren, Endymion Cooper, Laura Kelly, and Josiah Seaman all contributed to the assembly of *Fraxinus* genomes. Laura Kelly produced the species tree; GeMoMa and OMA runs used for the preliminary analysis of gene copy number, but not the final dataset. Josiah Seaman designed the visualization, improved 13 genomes, and carried out the gene count analysis. Meg Staton produced a *de novo* annotation of *Fraxinus pennsylvanica*. Josiah Seaman carried out the second round of annotation transfer with GeMoMa and OrthoFinder using the *F. excelsior* reference genome. This research utilized Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1.

Abstract

A core question in evolutionary biology is whether a particular event is due to chance or inevitability. Would the same outcome be repeated given the same starting conditions? Gene copy number change is a key factor in plant genome evolution and whole genome duplications (WGD) create a copy of every gene, most of which are lost over subsequent generations. In this study, I quantify the repeatability of gene copy loss using 28 new *Fraxinus* assemblies which share the same ancient WGD. Since many species share the same duplication, this is a powerful system in which to distinguish loss events that occur independently (i.e., convergently) in different lineages from loss events that occur once and are passed to all descendent lineages. I used patterns of gene presence/absence among the *Fraxinus* species to calculate rates of independent loss in different gene families, and compared these with a null model simulation where every gene family has an equal chance of loss.

I find evidence for convergent gene loss during the diploidization of *Fraxinus* through the concentration of losses in specific orthogroups while other orthogroups are retained in duplicate. Observed results have much wider distribution in the number of gene losses per family than predicted by the null model. My results show there were more subtrees (i.e., gene families) with zero losses (2.6x) and more subtrees with greater than five losses (4.8x) than under the null model. I discuss how repeated outcomes in evolution indicate some underlying cause, which can be either selection or physical constraints in neutral evolution (4.1.2.2). Protein function is found to be a weak predictor of gene duplicate retention, with developmental genes and transcription factors being overrepresented as retained duplicates. In contrast to some other studies, I find no evidence of different retention rates in one subgenome of the *Fraxinus* ancient allopolyploid over the other (no biased fractionation). I find evidence for convergent evolution of gene loss in *Fraxinus*. This study demonstrates a powerful method for detecting convergent evolution which could be applied to any clade which had undergone WGD and subsequent speciation.

4.1 Introduction

4.1.1 Distinguishing WGD in Species Trees

Whole genome duplications can be identified within a species tree using phylogenetic analysis. There are two possible approaches, one is to treat the tree topology as an unknown parameter and measure bootstrap values for a collection of tree topologies as was done in (Jiao et al. 2011; Julca et al. 2018). This method is particularly useful in the case of allopolyploidy, where the two parents can have different species trees combined in their descendant.

The second approach is to assume species tree topology and use the timing and position of the WGD as the unknown parameter. Maximum likelihood of observed gene counts in extant species is used to test the validity of a hypothesized WGD event. WGDgc is one such software simulation that uses gene copy birth/death rates to estimate the most likely timing of multiple WGD along a phylogenetic branch (Ta et al. 2013; Rabier et al. 2014). This method works best with multiple species with a wide range of divergence times before and after the WGD event. Underlying both methods is an attempt to identify consistent patterns of duplication across all gene trees that would indicate the presence of a WGD.

4.1.1.1 Verification Using Ks Values

Synonymous substitutions (Ks) can be used for independent cross-validation for the phylogenetic method of identifying WGD. Ideally, the two approaches to identify WGD (Ks and phylogeny) would agree (Ren et al. 2018 Figure S4).

An advantage in comparing techniques is that different techniques have complementary strengths and weaknesses at different time scales. The right technique will depend on the species and data available, but multiple methods is always better than one. The phylogenetic approach is less susceptible to the Ks saturation problems noted in Vanneste, Van de Peer, and Maere (2013) but required multiple surviving species. Shared synteny (blocks of homologous sequence in the same order in disparate organisms) is a third technique which can also provide evidence of higher ploidy levels but will lose coherence over time with high recombination rates and requires good quality assemblies (Bowers et al. 2003; Lyons, 2008; Lyons et al. 2008).

4.1.2 Superior Detection Power of Convergent Fractionation in *Fraxinus*

Chapter 1 proposes the Worldwide Ash Genomes (ashgenome.org/worldwide) be used as a test bed for a series of theories about paleopolyploidy already in the literature. For example, are gene losses primarily concentrated immediately after the WGD as predicted in Wolf and Koonin (2013) or do essential genes require a lag time to subfunctionalize as predicted by Robertson et al. (2017)? Ash trees provide a rare opportunity because there are 28 genomes sequenced of diploidized species that all share the same WGD. In other studies (Figure 1.5) there is a close to one-to-one relationship between the number of species sequenced and the number of WGDs discovered. Largely, this is due to the research incentive to sequence and study diverse model organisms while sequencing was very expensive. Now that sequencing is orders of magnitude cheaper, we will see more studies of closely related organisms. Figure 4.1 shows how the power to detect repeatability of evolution after a specific WGD (convergent fractionation) increases with the number of internal nodes in the species tree.

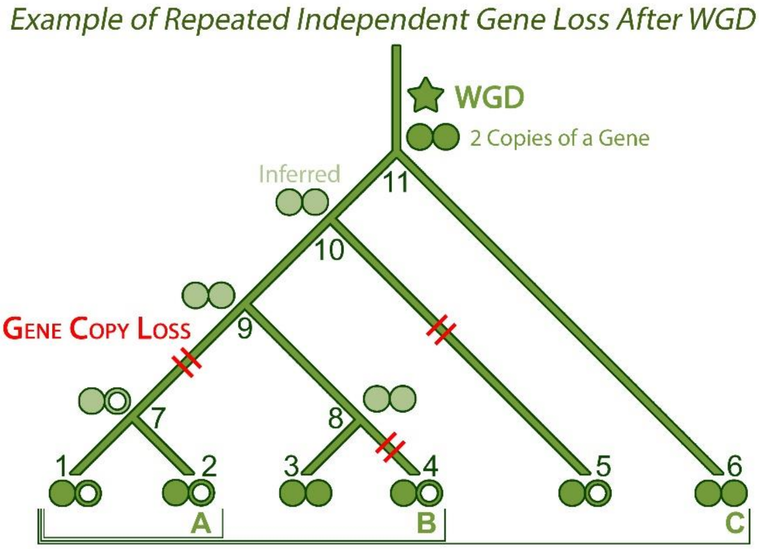


Figure 4.1: Example of Repeated Independent Gene Loss After WGD: Starting with a WGD at 2 copies (top), this example diagram shows how to infer the most likely series of events from present species gene copy numbers. Across the bottom, circles represent each species which has either one or two copies of the gene family remaining. From left to right: 1, 1, 2, 1, 1, 2 gene copies. Light green circles represent internal nodes from ancestral organisms where the gene copy is inferred rather than observed. A) If two species currently have a single copy of a gene family, then their immediate ancestor (node 7) is inferred to have one gene copy. If we infer that ancestor 9 had two copies, then a gene copy loss event is place in between the two ancestral nodes (marked in red). B) With four species there are sufficient internal nodes to detect repeated losses in two different branches. Starting with species 3 and 4, we can infer their ancestor 8 had two copies because species 3 still retains two copies. At this point, we've detected our first pair of convergent loss events (7-9 & 8-4). C) With six extant species at the tips, the species tree includes five internal nodes with inferred copy numbers and proportionately more detection power for independent changes. Species 5 and 6 add nodes 10 and 11 along with the inferred convergent loss 10-5.

All events and copy number of internal nodes can only be inferred using parsimony (the simplest explanation is usually best) and probabilistic modelling. For example, in the case of diploidization gene loss and gene gain do not have equal probabilities. Gene loss is many times more likely than gene duplication as is predicted in 1.4.4 and directly estimated in 4.3.2. This means that a longer chain of high probability events may have a higher combined probability than a single low probability event, despite the low probability event being more parsimonious. Figure 4.1 contains a direct example of this scenario. If instead, we inferred that nodes 8, 9, and 10 only had one copy, then a single duplication event could be placed at species 3. This would explain the existing copy number distribution with two events, instead of three. However, if the net probability is greater for three losses than a loss and a duplication, then the former scenario will be chosen as the most likely explanation. Evidence for convergent loss can also be backed by shared synteny (1.3.3), since novel duplications will not arise at the ancestral genome locus.

4.1.2.1 CAFE Simulation of Gene Copy Evolution

In a species tree with more than three species and each species with tens of thousands of gene families, there are many possible histories of gene duplication and loss that would all produce the same present day observed outcomes. CAFE is an algorithm that simulates the gene birth and death process that estimates the most likely rates of change in different parts of the species tree (De Bie et al. 2006). A major source of concern when measuring gene copy loss is that a gene may be present in the organism but simply not annotated, called a false negative annotation. The authors emphasize that their model accounts for noisy data by approximating the false negative annotation rate (Han et al. 2013). False negative rate is the percentage of all missing annotations which are due to errors as opposed to biological reality. If this number is estimated too low, then a poor annotation will lead to inferring the genome has lost significant gene content. If this number is too high, the tool would disregard true losses as noisy data, asserting few real losses had occurred. CAFE also includes detailed outputs of where gene copy changes likely occurred (4.3.2.1) that are normally only available in forward evolution simulations like MEGA5 (Tamura et al. 2011).

4.1.2.2 Neutral Evolution vs. Selection

The repeatability of evolution does not necessarily prove the action of selection in this study. Beyond technical artifacts, there are physical biological reasons why we might expect the same outcome to be repeated given the same starting conditions that have nothing to do with selection pressure mediated by the fitness of the organism's survival or the number of offspring it produces. For example, if a gene is near the telomere it is more likely to be lost because of the physics of polymerase, telomerase, and recombination. These factors bear equally on all sub-telomeric genes regardless of their fitness advantage. This set of dynamics is covered in neutral evolutionary theory and it highlights the large amounts of sequence which are not under selection pressure (Kimura 1979). Similar physical forces apply to the likelihood of nonallelic homologous recombination, silencing due to proximity to a transposon, or sequence robustness to mutation (Kidd et al. 2010; Ecco et al. 2016; Schulte et al. 2014).

Selection is one possible explanation for repeated outcomes in evolution. It may be the most likely explanation for convergent features in unrelated species, such as the independent development of echolocation in mammals: toothed whales, dolphins, and bats (Parker et al. 2013). However, in this study, where the physical constraints of the ancestor are identical in all organisms, neutral evolution is an equally viable explanation.

4.1.3 Null Model From Li Et al. 2016

A major advantage of stochastic simulations is that they can be used as a null model by removing input data or constraints, running the simulation, and comparing the resulting outcomes. The specification for stochastic simulations of gene birth/death which included WGD was introduced in Rabier, Ta, and Ané (2014) and implemented in the R package WGDgc (Ta et al. 2013). More recently, (Li et al. 2016) employed a null model following the same specifications and used the Wilcoxon rank sum to test the probability the null and observed data were from the same distribution.

CAFE 4.2 includes a null model which can be used to establish a baseline distribution for the amount of gene copy loss per gene family expected by chance (Han et al. 2013). While it does not support WGD, CAFE `genfamily` includes detailed simulation of a large number of gene families with different sizes and a complete history of change along the species tree topology. The simulation runs until it matches the observed distribution of gene family sizes in the actual data.

4.1.4 Parental Subgenomes and Biased Fractionation

In polyploids where a parental genome is still available, it is possible to detect which subgenome came from which parent and thus determine how much of each parental contribution may have been lost in fractionation (Edger et al. 2018). In cases such as Oleaceae where a parental population is no longer available, investigating biased fractionation is usually not possible.

4.1.5 Functional Enrichment Analysis

Many of the key explanations for why genes are lost or retained are tied to the functions of their proteins (1.4.1, 1.4.2, 1.4.3). Functional enrichment analysis seeks to establish that certain functional categories are overrepresented in the set of retained genes. For example, if transmembrane proteins made up 7% of an organism's proteome but 21% of the study set, then transmembrane proteins would be 3-fold enriched in the study set (Huang, Sherman, and Lempicki 2009). GOenrich is a Python library which uses the hypergeometric survival function to calculate the odds the same category would be picked multiple times (Rudolph et al. 2016). The calculation uses a background model based on the entire genome(s) provided and then calculates enrichment of terms for the subset of genes provided as a second list (Huntley et al. 2015).

4.1.5.1 Challenges of Non-model Functional Annotation

When working with non-model organisms there are methodological challenges in function annotation. For example, the pipelines used on *F. excelsior* and *Olea europaea* are based on finding similar sequences in well-annotated genomes like *Arabidopsis thaliana* and conjecturing they have the same function in a distantly related plant. In reality, even proteins with very similar sequences and protein folds can have different functions (Mindrebo et al. 2016). Many studies mitigate this shortcoming by focusing on high level functional categories that are less susceptible to change (Li et al. 2016; Maere et al. 2005; Rodgers-Melnick et al. 2012). The GO ontology structure is ideally suited to this hierarchical approach.

Finally, studies in one domain of life can limit themselves to appropriate terms for the organism. For example, GO Slim Plants is the subset of terms relevant to plants and would exclude animal or bacterial terms. Term relevancy is also accounted for by using the background model of the full genome to calibrate term frequencies for the organism.

4.1.6 Introduction Summary

In this chapter, I investigate the fates of genes which were duplicated in the last whole genome duplication (WGD) shared by *Olea* and different species of *Fraxinus*. This study focuses on quantifying the repeatability of evolution in *Fraxinus*. Through the analysis of 28 species we have greater power to detect repeated gene loss than any previous study (4.1.2). CAFE serves as the basis for phylogenetic inference and null models while I replicate the Wilcoxon ranksum test used in Li et al. (2016) and Han et al. (2013). I find the rate of gene loss has accelerated as predicted by the Lag Time model (4.3.2.2 & 1.4.5) (Schranz, Mohammadin, and Edger 2012; Dodsworth, Chase, and Leitch 2016; Robertson et al. 2017; Cheng et al. 2018; Clark and Donoghue 2017). Protein functions are shown to be weak predictors of gene retention (4.1.5) in non-model organisms (1.4; Li et al. 2016). Finally, I find no evidence of biased fractionation in parental subgenomes, which is consistent with reviews showing this is not a universal phenomenon (Table 1.1; Wendel et al. 2018).

4.2 Methods

The exomes of the worldwide Ash genome project (3.3) served as the input sequences for this study, plus *Olea europaea* (2.2.5.1). Since this study focuses on the Oleaceae WGD, it excludes *Solanum* and *Erythranthe*. As mentioned in 3.3.1. notable species, *F. apertisquamifera* and *F. uhdei* were later excluded as polyploids or hybrids. *F. pennsylvanica* is represented by the FRAX09 draft assembly, rather than the chromosome level assembly in 3.3.2, so FRAX09 is methodologically consistent with the other genomes in the results and in Chapter 2. The genes of interest in this chapter are the identified homeologs in 4.2.1 and compared against the background gene annotation set from GeMoMa section 3.3.4.

Data Availability

Scripts used in this study are all available in a single git repository:
https://github.research.its.qmul.ac.uk/btx142/DNA_Duplications

Figures and Supplemental Data Files are available on Google Drive:
https://drive.google.com/drive/folders/1hcvcYvN_tOGA-hyqH_pqJrUEiQa51KDz?usp=sharing

4.2.1 Homeolog Filtering Based on Reconciled Gene Trees

To ensure only the fate of homeologs was studied, the set of all Oleaceae genes from Chapter 2 was filtered down to a subset of homeologs. This chapter focuses solely on the subset of genes which were duplicated during the two WGDs at the Oleaceae root. Since Jasmine data was not available at the time (Figure 1.6) I was unable to phylogenetically separate the first and second Oleaceae WGD, however they clearly separate based on Ks.

Gene trees have duplication nodes any time a gene was duplicated in an ancestor (Figure 4.2A), leading to two subtrees containing one copy each per species of the two paralogs. For a gene family starting with one copy, which then went through two WGDs without loss, there would be three duplication nodes resulting in four gene copies. For the sake of clarity, I focus on the subset of homeologs with only one detectable duplication at the Oleaceae root. This criteria removes the effects that could be introduced from gene families of different sizes and significantly simplifies interpretation of the results. Starting from a single progenitor gene before two WGD, there are three scenarios that I consider here which can generate this outcome: 1) The older duplicate from the first WGD was lost and recent duplicates were retained. 2) Old duplicates are retained, one recent copy from each subtree is lost. 3) More copies were retained, but OrthoFinder did not assign them to the same orthogroup. The Ks results described in 4.3.1.2 indicate that scenarios 1 and 3 are unlikely due to the inclusion of outgroups in the rooting of orthogroups and the reference based annotation of GeMoMa (2.2.4). In effect, this study focuses on gene families which have a history of reducing to single copy number, making them ideal candidates for studying diploidization.

Gene trees from section 3.3.5 processed with OrthoFinder were used as input for the complete gene set to ensure only the fate of homeologs was studied; the set of all Oleaceae genes from chapter 2 was filtered down using gene tree criteria. OrthoFinder outputs the file Duplications.csv which lists, for each gene family, the duplication nodes in the reconciled gene tree and the corresponding species tree node found using the DLC search algorithm (Emms and Kelly 2019). The Oleaceae root where both WGDs occurred is labeled N2 in the species tree. The Homeologs_Analysis.ipynb script was written to process the file and collect the total number of mentions of N2 duplications per orthogroup. Orthogroups which contained only one duplication at N2 were selected for study (Figure 4.2B). These selected genes are hereafter referred to as "homeologs".

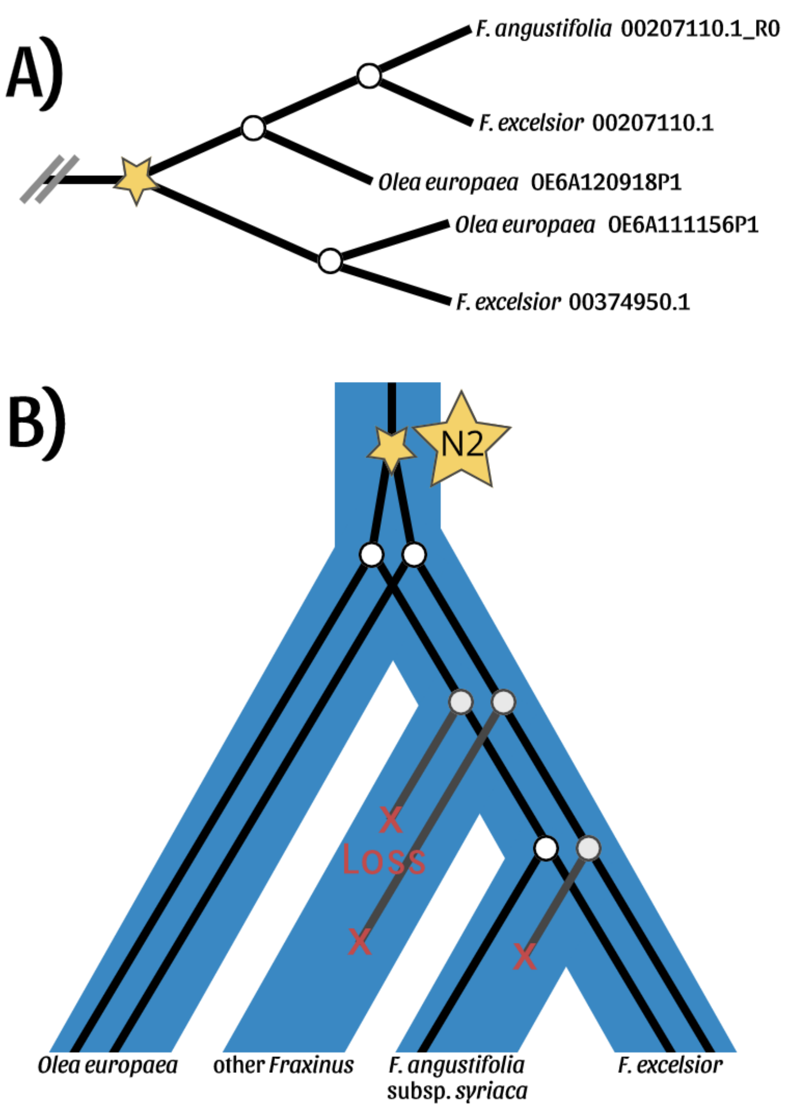

A)
F. angustifolia 00207110.1_R0
F. excelsior 00207110.1
Olea europaea OE6A120918P1
Olea europaea OE6A111156P1
F. excelsior 00374950.1
B)
N2
Loss
Olea europaea
other Fraxinus
F. angustifolia subsp. syriaca
F. excelsior

Figure 4.2: Example of reconciled gene tree with homeologs. A) The example gene tree (black lines) for orthogroup OG0006370 was specifically chosen for its small size. This subtree shows all detected gene copies arising from a single duplication event (star) distributed across three species. Gene trees cannot differentiate between the two WGD on the same Oleaceae branch, but we know from the position this subtree are homeologs. B) Gene tree reconciliation maps the gene tree topology onto the species tree, which can have a very different topology from the species tree (blue lines). Duplications (star) are branches in the gene tree with no corresponding speciation event (white nodes). Homeologs are identified as gene copies tracing back to an ancestral node that maps to the species tree position where we know the WGD occurred (N2 - star). In the simplest case, this creates two copies of the species tree topology in the gene tree, with one *F. excelsior* and one *Olea europaea* copy in each subtree. In total, two gene copies are retained in *F. excelsior* and *Olea europaea*. The scenario depicted here shows one copy retained in *F. angustifolia* subsp. *syriaca* (FRAX16) and one lost, while both copies are lost in all other *Fraxinus* species. Given its close relationship to *F. excelsior* it is also possible that it acquired its copy through hybridization.

4.2.1.2 Independent Validation of Homeologs with Ks and 4DTv

To verify the correct genes were selected using gene tree criteria, I used Ks values as an independent estimator of gene copy divergence time. The script KsPlots.ipynb was written to gather all the CDS sequences for all the identified homeologs and compare their aligned sequences against their homeolog pairs within the same genome. Comparisons were only made within each species and not between species, to avoid introducing additional factors such as a speciation peak. However, once the Ks values were calculated, the data were aggregated across all species under the assumption that all *Fraxinus* species have roughly the same rate of synonymous mutation accumulation.

The function count_synonymous_sites uses a standard codon table to count the number of synonymous substitutions that have occurred between the two sequences. The difference between Ks and 4DTv is that 4DTv only counts those positions which are four-fold degenerate codons (32 out of 64 possible codons) as opposed to wobble positions, and is intended to be less sensitive to mutational noise. I use non-calibrated Ks plots with the x-axis being a literal count of the ratio of synonymous changes per 100 opportunities for a synonymous change. Correct selection of homeologs would be indicated by a Ks peak at the second peak in Figure 1.6 on the WGD. A failure in the filtering steps outlined in 4.2.1 would be indicated by a lack of change in the shape of the Ks graph as opposed to the background Ks distribution.

4.2.2 CAFE for Phylogenetic Inference of Gene Loss Events

4.2.2.1 Preparing Orthogroups for CAFE

The contents of Orthogroups.GeneCount.csv was filtered to remove Orthogroups containing only one gene copy or over 100 copies as the designers of CAFE have found that very large gene families can inflate the estimation for rates of gene copy gain and loss ("CAFE Tutorial: Computational Analysis of Gene Family Evolution" 2016). I used the provided script cafetutorial_clade_and_size_filter.py to produce 2020_Feb08_homeolog_filtered_counts.txt which contains the full list of 7,836 homeolog subtrees analyzed below.

4.2.2.2 Estimating Birth and Death Rates With CAFE

CAFE 4.2 was used to estimate the birth (lambda) and death (mu) rates of gene copies for the entire species tree. The data file (filtered_OG_counts.txt) contains a table with a row for each Orthogroup and a column for each species with the number of copies in each table cell. The input time tree for CAFE was Species_tree_corrected_root_ultrametric_integers.tre. Using the estimated values for lambda and mu, CAFE then inferred the most likely gene count for each node in the input tree for each gene family using the Viterbi algorithm (Forney 1973; Han et al. 2013).

4.2.2.3 CAFE Used For WGD Fractionation

CAFE 4.2 does not support the simulation of whole genome duplication or an immediate copy loss parameter, unlike the R package WGDgc (Ta et al. 2013). However, CAFE was selected for its rich output and tracking of specific genes and time points. The solution was to hand-craft the data to use CAFE to answer the biological question of fractionation. The two homeolog subtrees created by the WGD for each locus were each submitted as their own "gene family." For example, OG0000069 was submitted to CAFE as 10000069 and 20000069 and copy numbers were tracked separately. Assignment to 1000* or 2000* is purely for tracking subtree pairs consistently and does not imply a particular parent subgenome. Consequently, the simulation was initiated with two ancestral copies at the time of the WGD, after which point the copy number could change in the simulation to account for extant patterns of variation.

The input data includes only subtrees of genes which are considered to have duplicated at the base of Oleaceae, it therefore excludes duplicated genes from Solanum (Solanaceae) and Erythranthe (Phrymaceae). While their inclusion in Chapter 2 influences the scope of orthogroups, their evolution rates and copy numbers are not present in these results. Finally, rate estimates did not include the WGD event itself. The simulation is a study of only the decay phase of the Boom and Decay Model (Figure 1.3; Wolf and Koonin 2013) because the assumed short term genome instability immediately following WGD leaves no evidence behind for study ~60 million years after the event.

4.2.2.3.1 CAFE Simulation Inputs

The species counts for each of the split homeolog subtrees was used as the input for a CAFE simulation of maximum likelihood ancestral copy number inference (Han et al. 2013). First, an estimate of global annotation error rate was calculated using the caferror.py script included with CAFE. This script performs a grid search for the error parameter which maximizes the probability of observing the data. The starting error estimate input was 0.08 based on species overlap statistics. Per species error rates were not estimated.

CAFE was then used to run a search for the most likely birth (lambda) and death (mu) rates using the obtained error rate (lambdamu -s). Separate rates were used for birth and death since this study specifically focuses on the preferential loss of homeologs after a large duplication. Finally, the included Python script cafetutorial_report_analysis.py was used to summarize results.

```
load -i 2020_Feb08_homeolog_filtered_counts.txt -t 8 -l
onerate/Feb008_homeologs_onerate_0.1_error.log
tree
((((((((((FRAX30:2.0,FRAX32:2.0):1.0,FRAX28:3.0):2.0,FRAX12:5.0):4.0,(
FRAX07:8.0,FRAX29:8.0):1.0):4.0,FRAX08:13.0):1.0,(((((FRAX01:2.0,FRAX1
6:2.0):4.0,FRAX15:6.0):2.0,FRAX00:8.0):2.0,(FRAX06:9.0,FRAX23:9.0):1.0
):3.0,FRAX25:13.0):1.0):3.0,FRAX21:17.0):2.0,(((FRAX19:8.0,FRAX20:8.0)
:2.0,((FRAX11:5.0,FRAX27:5.0):4.0,FRAX04:9.0):1.0):1.0,(((((FRAX03:1.0
,FRAX09:1.0):1.0,FRAX13:2.0):2.0,(FRAX26:2.0,FRAX14:2.0):2.0):3.0,FRAX
05:7.0):2.0,FRAX33:9.0):2.0):8.0):15.0,FRAX31:34.0):2.0,Oeuropea:36.0)
errormodel -model cafe_errormodel_0.1.txt -all
lambdamu -s
report onerate/Feb008_homeologs_onerate
```

4.2.2.3.2 Convergent Losses in Homeolog Lines

Using the inferred copy numbers the amount of convergent gene loss was quantified: specifically, how many times has the same gene family lost a copy in independent lineages. For each homeolog subtree, the species tree was traversed. Each time a direct child node had a copy number lower than its parent node, a “loss event” was recorded in a table for that homeolog subtree (Figure 4.1). The total number of independent loss events for each homeolog subtree was used to measure the degree of polyphyly in the pattern of gene loss since all homeologs started at the same copy number and had the same number of opportunities for loss in the species tree. The process was repeated for gene copy gains. The code can be found in the notebook TreeCountAnalysis_Oleaceae_Homeologs.ipynb in the method populate_history_histogram.

4.2.2.4 Rate of Change Analysis

To test whether the rate of change has been constant over time, the number of changes inferred with the time elapsed was compared. The timeline of gain and loss events was aggregated into one-million-year time interval bins according to the phylogenetic branch where the event occurred. Each branch contributed their total number of events normalized by their branch length to each interval. With more extant species there are more possible events that can happen, so the total number of events per time interval is normalized by the number of extant branches in that time interval. If many branches are present at a particular time point, they add granularity to the results, but do not artificially inflate the rate of change. The code can be found in the notebook TreeCountAnalysis_Oleaceae_Homeologs.ipynb in the method populate_change_histogram.

4.2.3 Null Model Simulation

A null model was generated in order to compare actual results with the results expected from gene birth and death rates along the same species tree topology. The CAFE function `genfamily` was used to simulate a stochastic birth-death process over the actual species tree. The simulation was restricted to contain the exact same number of homeologs with the same count per extant species and the same root size distributions estimated from ancestral copy number inference in the previous step. This restriction ensured the only variable between the null model and observed data was the distribution of events in the internal nodes of the species tree. Parameters used for `genfamily` match the maximum likelihood estimates from the previous steps: Birth: 0.00093971443748; Death: 0.00947308078407; errormodel with copy number transition probabilities: +1= 5%, +0= 90%, -1= 5%. The null model uses the following script:

```
load -i 2020_Feb08_homeolog_filtered_counts.txt -t 8 -l
simulate_onerate/simulate_onerate.log
tree
(((((((((FRAX30:2.0,FRAX32:2.0):1.0,FRAX28:3.0):2.0,FRAX12:5.0):4.0,(
FRAX07:8.0,FRAX29:8.0):1.0):4.0,FRAX08:13.0):1.0,(((((FRAX01:2.0,FRAX1
6:2.0):4.0,FRAX15:6.0):2.0,FRAX00:8.0):2.0,(FRAX06:9.0,FRAX23:9.0):1.0
):3.0,FRAX25:13.0):1.0):3.0,FRAX21:17.0):2.0,(((FRAX19:8.0,FRAX20:8.0)
:2.0,((FRAX11:5.0,FRAX27:5.0):4.0,FRAX04:9.0):1.0):1.0,(((((FRAX03:1.0
,FRAX09:1.0):1.0,FRAX13:2.0):2.0,(FRAX26:2.0,FRAX14:2.0):2.0):3.0,FRAX
05:7.0):2.0,FRAX33:9.0):2.0):8.0):15.0,FRAX31:34.0):2.0,Oeuropea:36.0)
errormodel -model cafe_errormodel_0.1.txt -all
lambdamu -l 0.00093971443748 -m 0.00947308078407
genfamily simulate_onerate/simulation -t 1
load -i simulate_onerate/simulation_1.tab -t 8 -l
simulate_onerate/simulation_viterbi.log
report simulate_onerate/simulate_onerate
```

4.2.3.1 Wilcoxon Rank Sum for Comparing Null Model Metrics

To quantify the probability that the simulated and actual data came from the same distribution, the two sets of observations were compared using the Wilcoxon Rank Sum test (Wilcoxon et al. 1970). This test was also used against null distributions of rate of change in gene copy number, repeated events in one homeolog subtree, and biased fractionation. It is a non-parametric test which counts the number of times one observation in set A is larger than an observation in set B for all pairwise combinations (Wild and Seber, 1999). This test is used to check the significance of the difference between null simulation and actual data for 4.2.2.4 Rate of Change, 4.2.2.3.4 Number of Repeated Losses per Homeolog.

4.2.3.2 Species Overlap Probabilities

The significance of overlap in loss of gene families between distantly related species is used to control for loss caused by errors in annotation. To test whether losses in one species could predict losses in other species, a set intersection of the set of genes lost in one species and the set of genes lost in another species for all pairwise combinations of species was performed. The probability of an overlap in two randomly distributed independent samples is the probability of observing *k* successes in *n* draws where there are no replacements is defined by the hypergeometric distribution (Rice, 2006). Scripts for these comparisons can be found in TreeCountAnalysis_Oleaceae_Homeologs.ipynb starting with simple_overlap_prob.

4.2.5 Functional Enrichment Analysis

Function categories were assigned by taking the existing functional annotations from *F. excelsior* present in each orthogroup (3.3.4) and transferring those functions to all other member genes of the same orthogroup. This is a low precision inclusive approach that is well suited to test claims in the literature that resistance to fractionation is associated with high level categories such as developmental genes and transcription factors (Li et al. 2016). This analysis focuses primarily on high-level functional categories, which are likely to be stable over great phylogenetic distances.

Gene function annotation of all *Fraxinus* was based on the annotation produced in (Sollars et al. 2017) available in the file Fraxinus_excelsior_38873_TGAC_v2.gff3.functional_annotation.tsv. For this study, a script for annotation transfer was written and is available in FunctionsForFamilies.ipynb. Final results are in GO_Enrichment.ipynb. Each *F. excelsior* gene has a set of GO terms associated with it. Each *F. excelsior* gene has one orthogroup clustered by OrthoFinder (3.3.5). The union of all *F. excelsior* GO terms for genes contained inside each orthogroup are assigned to the orthogroup in a lookup table. All orthogroups with no GO terms are removed from this analysis.

Goenrich 1.10.1 (https://github.com/jdrudolph/goenrich) was used to carry out the enrichment analysis. The ontology used was go-basic.obo available at http://geneontology.org/ontology/go-basic.obo. The background model was built from all orthogroups that have GO terms. Note that this treats large or small gene families as one sample each, eliminating the effects of large gene families. Similarly, selection criteria for a query set is based on the entire orthogroup, not individual genes. Enrichment of the query set also treats large or small families as one sample each. Goenrich analyses the query set against the background and outputs an HTML table sorted by p-value. Internally, enrichment probabilities are calculated with the same hypergeometric survival function (4.2.3.2).

Given the dependence between gene functions and orthogroup selection, accurately reporting the magnitude of overlap significance is problematic, though the significance ranking is stable. A 10% overlap in a set of 10 orthogroups is not significant, however a 10% overlap in a set of 1,000 genes is highly significant. If the calculation is done using gene counts, the top results will be $p < 10^{-80}$. When calculated by orthogroup, the order is the same but all results are $p > 0.001$. These conservative probabilities are the values reported in the final results.

4.3 Results

As the number of species in the study increases, the number of internal nodes increases geometrically and with it the detection power for repeatability of evolution. For this research, a pilot study was originally carried out using only the 6 exemplar *Fraxinus* species and *Olea europaea* in a species tree with only 5 internal nodes. The results were too coarse to draw any meaningful conclusions (data not shown). In this study, data was pulled from a species tree with 28 species and 26 internal nodes which share the same WGD. I was able to detect up to 20 independent events per gene family. My results have enough granularity to support statistical analysis such as rate of change (4.3.2.3), separate out categories of gene families (4.3.4), and contrast with the results of a null model birth/death process simulation (4.3.3).

4.3.1 Identifying Homeolog Nodes Inside of Gene Trees

There were 4,076 orthogroups which contained one duplication in the Oleaceae root branch, in total 44.4% of the possible orthogroups were used in the study (Figure 4.3). Only genes in the subtrees under the selected duplication nodes (not the whole orthogroup) were included: 198,955 genes in total.

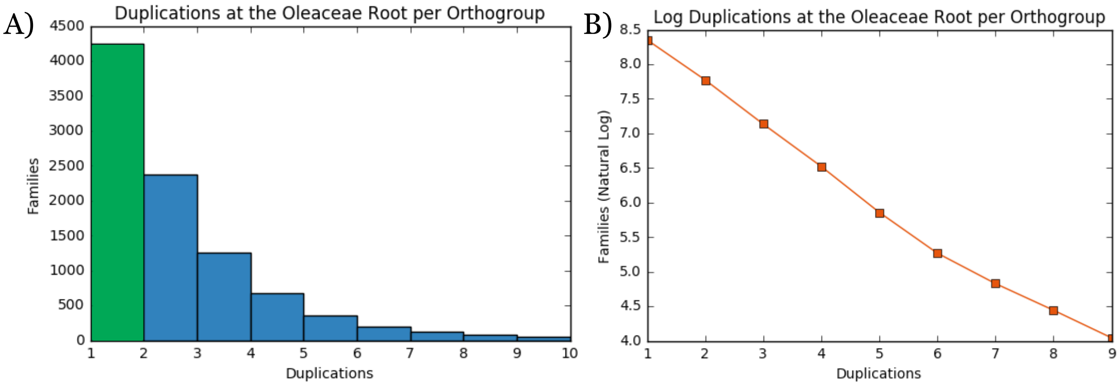


Figure 4.3: Number of Gene Families with Multiple Duplication Events at Oleaceae Root: A) Only 4,076 orthogroups were selected (green) out of the 9,356 orthogroups with a duplication event at the appropriate node. These orthogroups only had a single duplication event. The remaining 66% of orthogroups have more than one duplication event and are thus more complex to interpret. B) Same orthogroup counts as A graphed on a log scale y-axis. The slope of this line indicates each bar in (A) is 54% the size of the previous bar. The first 3 duplications can be explained by two WGD (see 4.1.1), while further duplications require small scale duplications (SSD) and large gene families to explain. The odds of duplication is remarkably consistent out to 10 duplications or 1,024x amplification, possibly hinting at a deeper biological phenomena. The consistency of the ratio of n+1 duplications is indicated by the straightness of the line in the log scale plot. Standard deviation of the ratios is 6.2% from the mean of 54%.

4.3.1.2 Homeolog Filtering Identifies Older Oleaceae WGD

It was previously known that Oleaceae had two WGD, at 36 Mya and 60 Mya respectively (Figure 1.5). It was unknown which WGD would contribute the majority of homeologs after filtering based on the criteria of only a single Oleaceae duplication event (4.2.1). Ks values show that the majority of homeologs after filtering were duplicated during the 60 Mya WGD (Figure 4.4 & 4.5). Manual checks with *F. excelsior* gene names confirm that 36 Mya diverged homeologs were almost all filtered out leaving only the 60 Mya set. The filtered study set only includes 18.6% of the total *Fraxinus* genes so the majority outcome for the study set is not the most likely scenario for all *Fraxinus* genes. In the majority of study genes, the gene started as a single copy at the Oleaceae root over 60 Mya and was duplicated in the 60 Mya WGD. After this, there is no surviving evidence of the second 36 Mya WGD. I would infer that immediately after the WGD there were four copies. After this, any combination of two gene losses that left one homeolog from each of the 60 Mya subtrees will allow the homeolog gene tree to pass the filtering criteria be used for the study. Initial genome instability of neopolyploids (1.2.2.5) is a possible explanation for this undetectable loss. Aimed at studying diploidization, the target genes consist of orthogroups with a tendency to reduce to single copy number but only after being retained in duplicate for at least 24 million years.

CAFE simulations (4.3.2, 4.3.3) only operate on a branching phylogenetic tree, and thus encompass the most recent 36 million years starting at the divergence of *Olea* and *Fraxinus*. Meaning there are 24 million years between the 60 Mya WGD that created the study homeolog pairs and the start of the species branching necessary for simulation. Going back farther than this is not possible without additional species. However, this does not pose a major problem for interpretation since the root copy number for each gene family is inferred by observed gene counts in filtered Orthofinder output (2020_Feb08_homeolog_filtered_counts.txt). Counts may be less than two, accounting for the history that transpired in the 24 million years not included in the simulation. The study here focuses on the events inside genus *Fraxinus*.

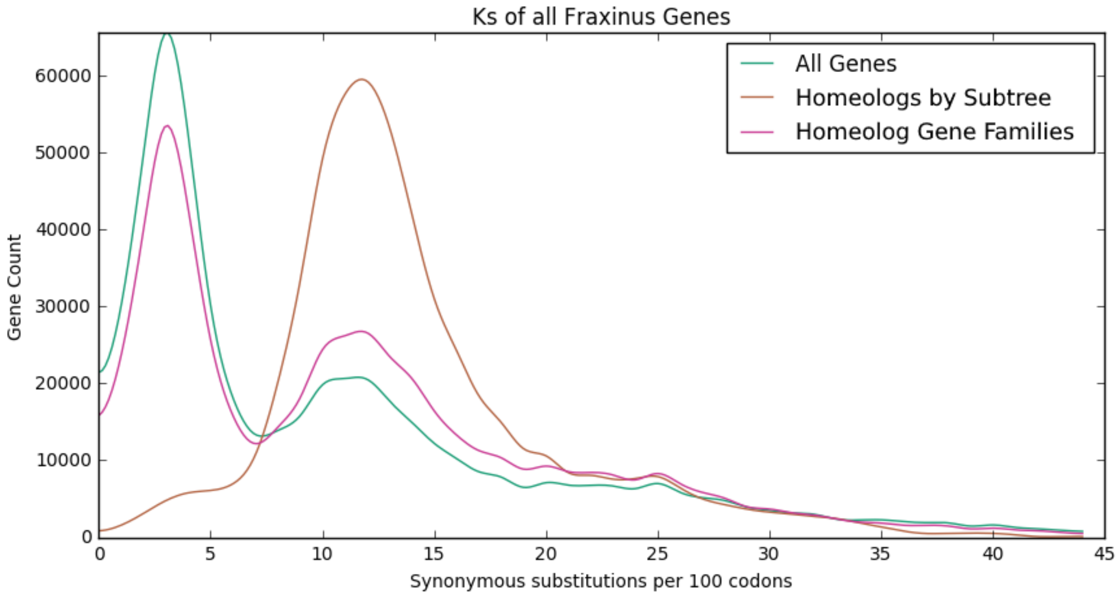


Figure 4.4: Ks of all *Fraxinus* Genes: Synonymous substitutions per site (Ks) were used to verify the divergence time of gene copies matched the second peak, or older Oleaceae WGD ~60 million years ago (Mya). Effective filtering for homeologs should eliminate the first peak and amplify the second peak. *Fraxinus* homologs across all *Fraxinus* species were used to measure Ks. The first peak on the left corresponds to the first WGD at 21 Mya. Using all genes within a gene family containing homeologs (pink line) makes only a small difference in the filtering. Filtering by gene tree using only the genes descendent from a WGD node (4.2.1) completely eliminates the first peak and leaves only the second peak intact. These results validate that the gene tree filtering criteria used in the rest of study are effective in filtering homeologs with a consistent age from other kinds of duplications. There is a third WGD centered at 25 substitutions which corresponds to the ~135 Mya "beta" WGD in eudicots (Yu et al. 2017) seen in Figure 1.5. *Note:* To ensure the distributions were of comparable size, random samples were taken from "All Genes" set to match the size of the set of homeologs.

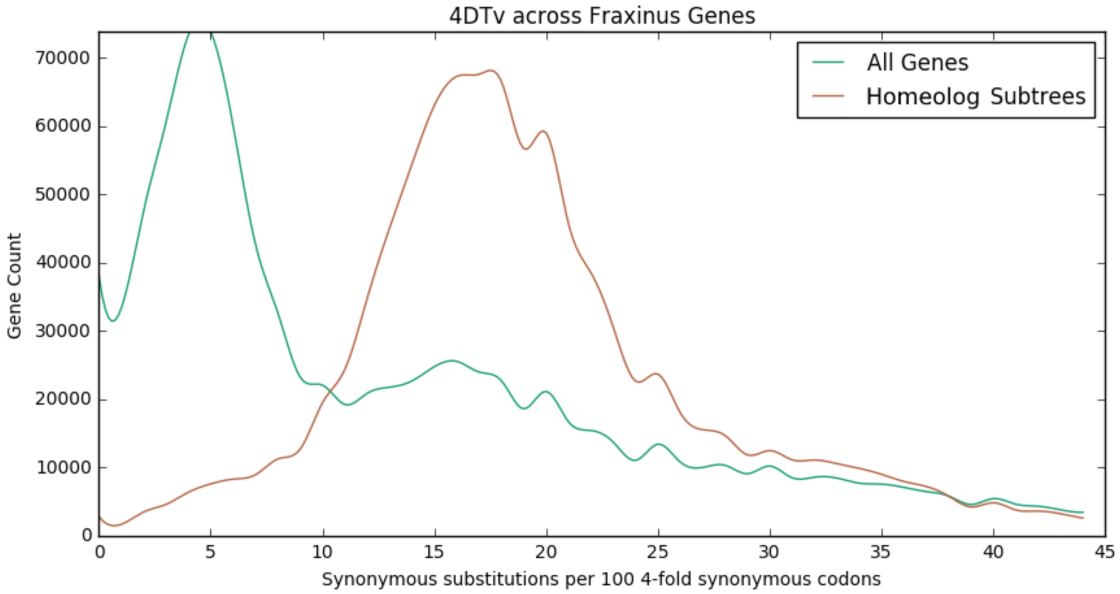


Figure 4.5: 4DTv Across *Fraxinus* Genes: 4DTv plots also show a selection for the more ancient WGD peak between 10-25 substitutions per 100 four-fold degenerate codons. Four fold synonymous substitution (4DTv) is a special case of Ks that focuses exclusively on the 24 codons that code the same amino acid in all four values of the wobble position. 4DTv saturates faster and can provide better time resolution for more recent evolutionary events.

4.3.2 CAFE Copy Number Simulation Results

4.3.2.1 Ancestral Copy Number Inference

CAFE estimated the average rate of gene copy birth or death (respectively) per homeolog subtree per million-year interval at Lambda (Birth) = 0.00093971443748 and Mu (Death) = 0.00947308078407 with a log likelihood score of -139824. Individual gain and loss at each species tree node in the simulation are shown in Figure 4.6. After CAFE accounted for estimated annotation error, birth went down by a factor of 3.2x and death went down by 1.9x, further widening the gap between birth and death rates.

The CAFE errormodel (4.2.2) estimated the annotation error rate at 10%. Meaning the 5% chance of false positive annotation and a 5% chance of false negative annotation results (+1: 5%, +0: 90%, -1: 5%). From this, it was estimated each species has roughly 407 (8,152 * 5%) species-specific losses due to annotation errors.

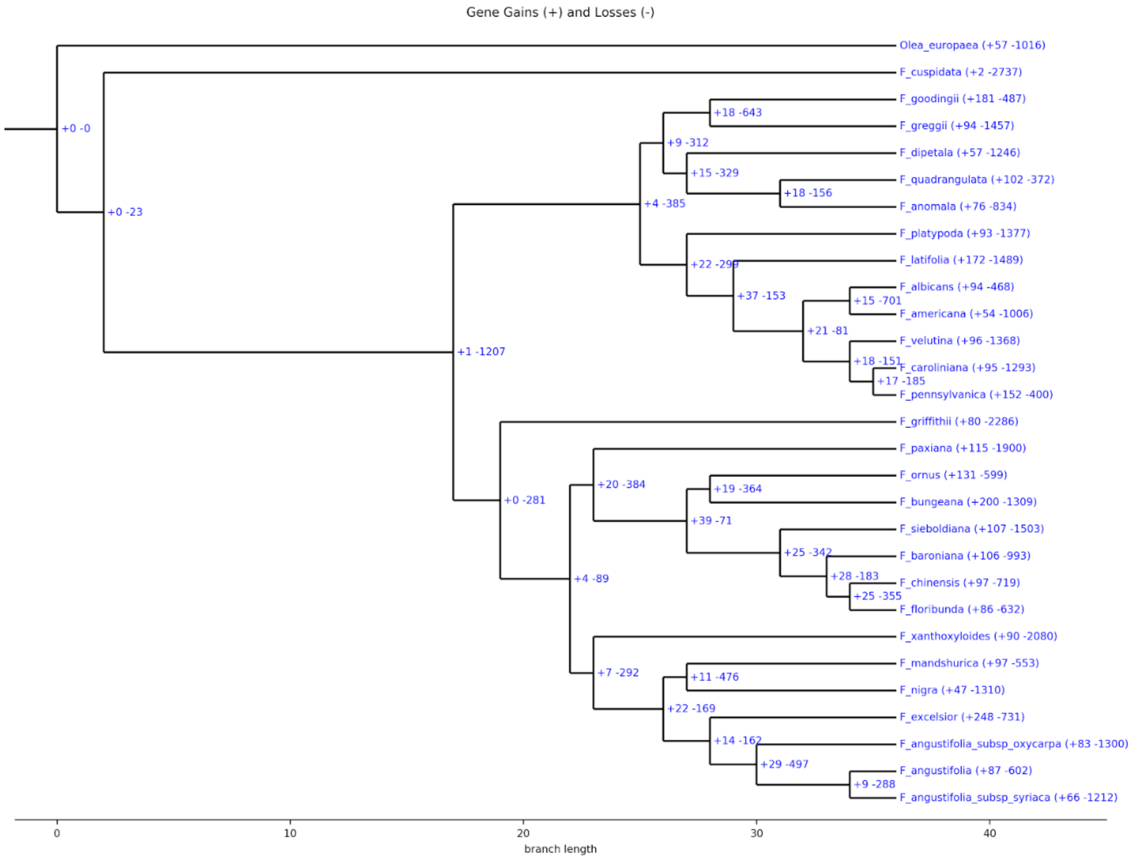


Figure 4.6 Tree of Inferred Copy Number Changes: All gains and losses of homeolog pairs at each node using CAFE maximum likelihood ancestral copy number inference. Branch lengths are given on the x-axis in millions of years since the WGD. Out of an average of 6,338 surviving study homeologs per individual , one can estimate the proportion of additional ancestral copies that are now lost by totaling the changes along any path from root to tip. Starting from the Oleaceae root (left) we see very few gene gains or losses in the internal nodes shared by large groups of species compared to the tips. Losses in each species are either unique or polyphyletic. In particular, the Oleaceae branch containing the two WGD (far left) that gave rise to both *Olea* and *Fraxinus* was not assigned any gene losses because there was no shared fractionation between the two genera (Figure 4.9). For example, the total estimated gene history for *F. excelsior* from tip to root would be 248-731 +14-162 +22-169 +7-292 +4-89 +0-281 +1-1,207 +0-23 = -2635 change in gene copy number. Meaning that diploidization has removed 2,635 total copies in the selected gene families between the common *Fraxinus* ancestor and *F. excelsior*. This is a typical outcome with *F. quadrangulata* at 2,332 net losses and *F. cuspidata* at 2,735 net losses. I note that *F. excelsior* (the annotation reference) is not an outlier in gene copy gains, but does have more gene counts (Table 3.2). This may be because *F. excelsior* is better assembled and has fewer paralogs collapsed together into a single gene model, but this had a negligible effect on family level aggregated counts. For *F. excelsior*, diploidization has removed 27.3% of the original gene content in the study homeologs.

4.3.2.2 Fractionation Follows Repeatable Patterns Across All Genes

Repeatability of evolution can be expressed in terms of the number of independent species show the same loss in the same gene family. For each gene family, each species only starts with two copies, so the maximum number of losses is two. However, with 28 species the maximum number of detectable species that have lost a gene family is 26, leaving two species with a shared gene family in order to establish its existence. Out of 26 possible losses across 28 species, the most common number of losses per gene family was two with a maximum observed of 20 losses in a single gene family (Figure 4.8). Given two sister species that share the same loss, CAFE will assign the loss to the ancestor node. Meaning the maximum number of losses detectable is less than 26 and 20 may be the theoretical maximum given the species tree topology (4.1.2). The shape of the distribution of losses did not match a normal distribution and was strikingly linear past two losses.

4.3.2.3 Rate of Change has Accelerated

In Figure 4.7 A&B, the rate of change over time was plotted from the WGD to present. For the x-axis of this graph, the WGD is assumed to immediately precede the split between *Olea* and *Fraxinus* (Unver et al. 2017). The simulation is started after the second WGD due to technical limitations of CAFE. Simulated data is relatively flat or decreasing in time. In both graphs, the first ~15 million years of the tree show little variation in rate because rates can only be calculated between nodes of the tree. For example, only one estimate can be obtained for the period from 2-17 Mya. Using the Wilcoxon Rank Sum test, the total distributions between simulated and actual data are not significantly different (p > 0.477). However, excluding the uniform first 20 million years, there is a significant difference in the last 16 million years (p-value = 4.6931e-05).

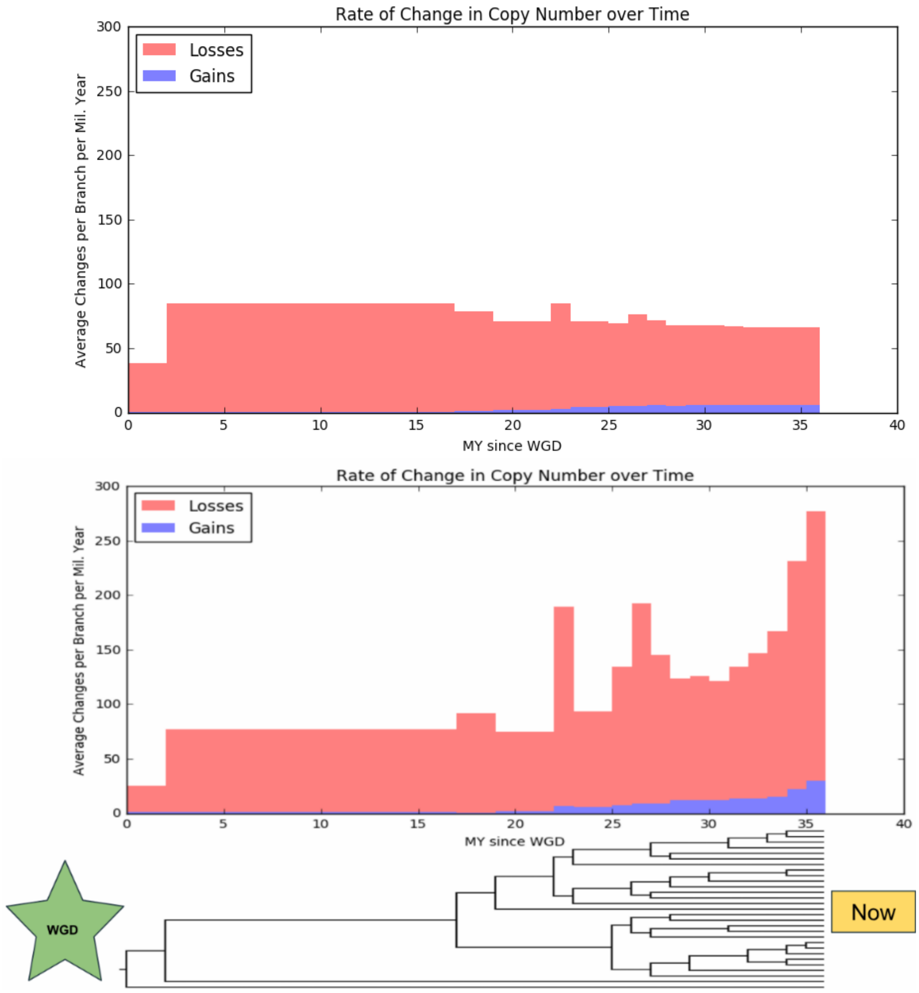


Figure 4.7: Rate of Change Over Time: The rate of change in simulated data is predominated by gene loss in red. Simulations show a much flatter distribution of rates (top panel). Whereas the real data (middle panel) is most consistent with an increasing rate of change in recent times (right) and intermittent spikes of losses at key speciation points. These spikes may be artefactual. Care has been taken to normalize for the number of extant species branches at every time point (bottom panel) on the shared x-axis between timeline and species tree. Changes in the height of the histograms correspond to the start of new species branches, but they do not artificially force higher values, as shown in the simulated data. This upward trend supports the Lag Time Hypothesis, the idea that genes retained after WGD must first diverge in sequence before they are free to be lost entirely (Schranz et al. 2012; Dodsworth et al. 2016).

4.3.3 Repeated Loss is Greater Than Expected By Chance

The number of independent gene loss events inferred in the actual data approached the maximum theoretical number of events that could be observed given the species tree topology (4.1.2). Null model simulations do not have any subtrees with more than 11 loss events (Figure 4.8). Subtrees with greater than ten independent loss events are only possible in cases where there are no losses in the ancient, shared branches and losses only occur in many recent, polyphyletic losses (Figure 4.7).

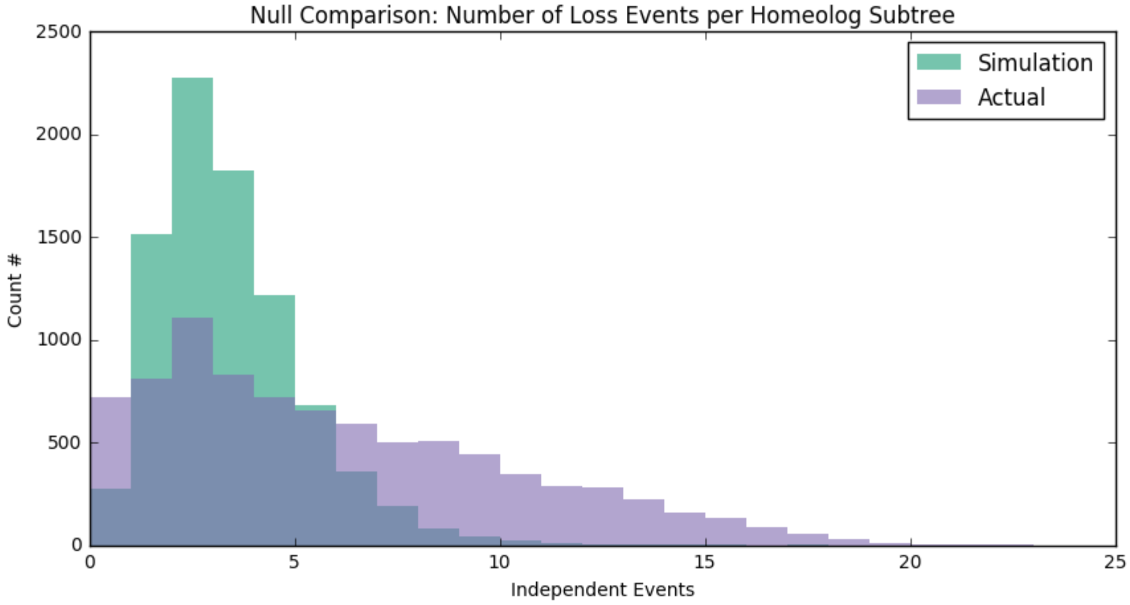


Figure 4.8: Null Comparison: Number of Loss Events per Homeolog Subtree: The same loss counting method was applied to both actual results and simulated birth / death process using the same species tree, rates, and starting distribution. The difference in the two histograms indicates actual observations of losses had a much higher variance. There were 2.6x more subtrees with zero losses and 4.8x more subtrees with greater than five losses than under the null model. The difference between actual and simulated distributions using the Wilcoxon Rank Sum Test was Z-score = 38.34 corresponding to a probability too small to calculate (p-value ≈ 0.0) (Wilcoxon, Katti, and Wilcox 1970).

4.3.3.1 Gene Set Overlap Between Species

Next, I tested whether or not the same genes were repeatedly lost in disparate species by looking at the intersection of sets of genes. Randomly distributed errors have a low probability of overlapping by chance, unless an unspecified attribute of the gene makes them more likely to cause annotation errors. Convergent evolution is detected by gene copy losses that are shared across species that cannot be attributed directly to the ancestral state. The significance of overlap between sets of genes lost between polyphyletic species was tested using the hypergeometric survival function (see Methods: Species Overlap Probabilities).

Out of 406 species pairs, 74.6% had an overlap significance less than $p < 0.001$ and 51.5% had $p < 0.000000001$ (Figure 4.9). The simulated null model had significant overlaps in less than 13.5% of pairs at $p < 0.001$ and 0 pairs at $p < 0.000000001$.

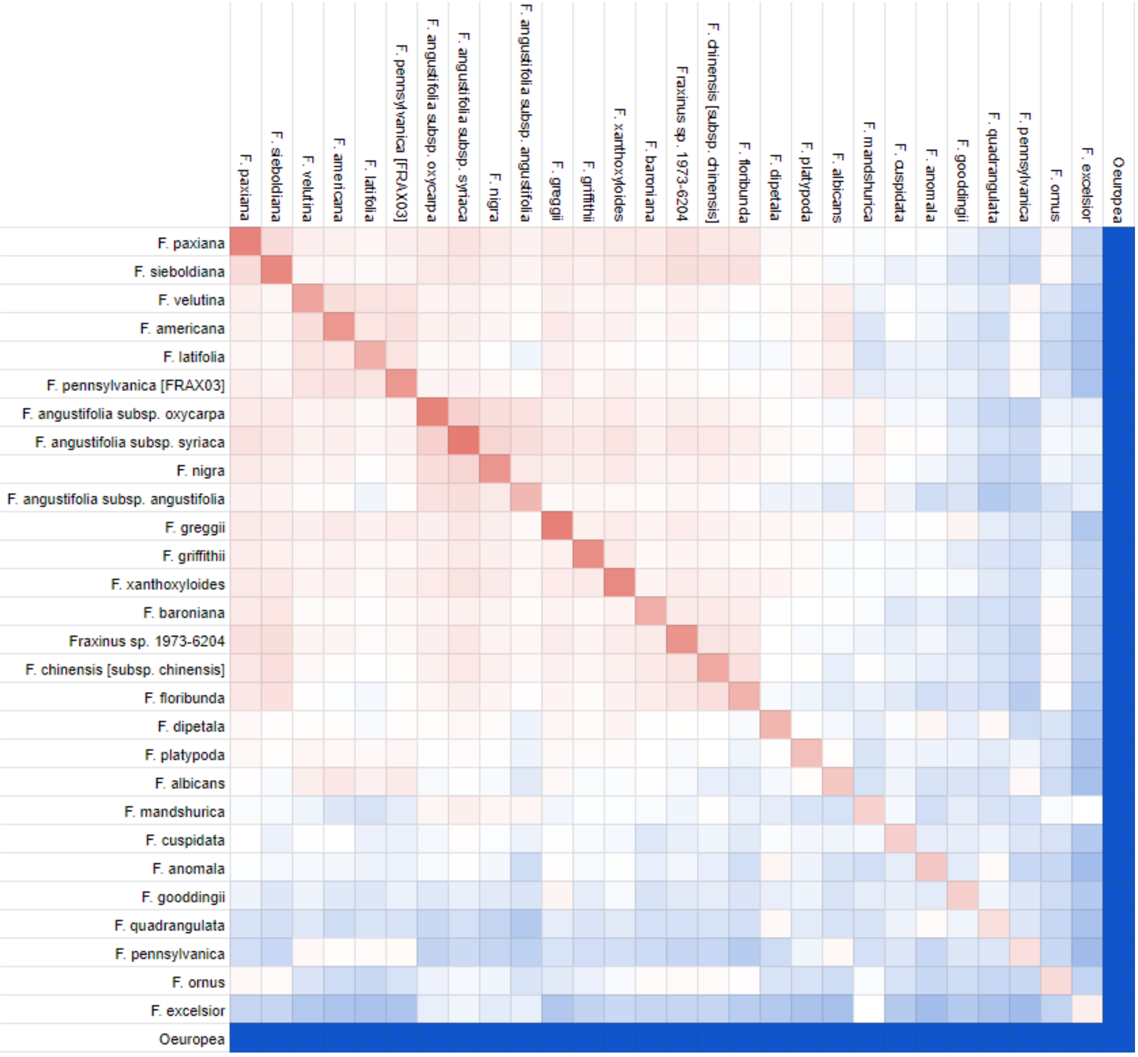


Figure 4.9: Number Of Gene Families With Loss Of Both Homeolog Copies Shared Between Species: Loss events were compared for agreement across all pairs of species. Due to the size of this table, it is presented here as a heatmap. The raw numbers are available as S4.1_Homeolog_Loss_Species_Overlaps.csv. Red indicates a higher than average level of overlap in losses. Blue indicates lower overlap, more genes are conserved. Rows and columns were sorted to maximize clusters. The dominant trend is that well-assembled genomes *F. gooddingii*, *F. quadrangulata*, *F. pennsylvanica*, *F. ornus*, and *F. excelsior* retain more genes and thus share fewer losses with every clade. Poor assemblies such as *F. greggii* and *F. xanthoxyloides* had more overlap with all genome losses. Clade structure is still visible, for example *F. angustifolia* and *F. nigra* as well as *F. velutina, F. americana, F. latifolia*, and *F. pennsylvanica* [FRAX03] cluster together, but the two *F. pennsylvanica* assemblies did not cluster together. Real biological signals are still visible; for example, *F. anomala* is closely related to *F. dipetala* and *F. pennsylvanica* still matches to its clade. *Olea europaea* has an additional WGD and higher gene copy numbers which meant its gene families rarely met the double loss criteria for overlap in this figure.

4.3.4 Functions of Homeologs

Functional enrichment results are shown starting from least specific to most specific filtering. The set of orthogroups containing one or more Oleaceae root duplications (57.4% of background) was notably enriched for transcription factors. Notable terms include: sequence-specific DNA binding (124/189), DNA-binding transcription factor activity (177/279), transcription regulatory region DNA binding (11/13), and transcription factor complex (193/311). There was also a notable enrichment in reproduction terms: fertilization (9/10) and sexual reproduction (32/46). The statistical significance in all terms was poor ($0.01 < p < 0.08$). It is not possible to obtain a 2x deviation from expected when the query includes over half of the background.

The inclusive set of orthogroups which contain homeologs with a single N2 duplication node per my definition in 4.2.1 is 25.3% of all annotated genes. The inclusive set is the same as “Homeolog Gene Families” in Figure 4.4 and should not be confused with the homeologs themselves. This set had the least significant term enrichment. It was notably enriched for gibberellin-related terms: response to gibberellin (20/49), cellular response to gibberellin stimulus (4/4), with no other significant results.

Next, I focused exclusively on homeolog genes and normalized p-values by genes instead of gene families (13.8% of background). Significant terms are too many to list here. However, the 90th percentile includes morphogenesis of anther, organ, and branching structures which are notable for their developmental consequences. Gibberellin, COP9, and stress-response signaling may be notable for response to new environments. The methodological difference is that the number of genes in the homeologous subtree affects the term enrichment as opposed to the background level. Without this effect, the results would be identical to the above inclusive set. Using a smaller percentage of the background and a much larger N allows for vastly smaller p-values. Full results can be found in S4.2 Homeolog Functional Enrichment.html.

The set of orthogroups which contain 6 or more duplication nodes (3.8% of background) at the Oleaceae root is necessarily filtered for high copy number within Oleaceae. Sorted by p-value, top results are all related to ribosomes until defense response (33/397), response to bacterium (18/204), response to fungus (14/139), and polysaccharide binding (6/36). These gene families were likely in multiple copies before the two Oleaceae WGDs.

4.4 Discussion

4.4.1 Homeolog Identification

Phylogenetic filtering plus requiring only a single Oleaceae duplication successfully selected homeologs arising from the older 60 Mya WGD (Figure 4.4). Timing of fractionation rates are best calibrated when all of the study homeologs arose at the same time. The selected homeologs only represent 18% of the average gene content available and more could be selected if I was to include gene trees with multiple Oleaceae duplications (Figure 4.3A). For example, if I allow two and three duplications in Figure 4.3A to open up homeologs that were affected by both WGD, then I would add another 3,609 to the existing 4,076 orthogroups, nearly doubling the number of orthogroups. Not all orthogroups are of equal size, and so this may more than double the number of study genes given that they have more duplication nodes in each gene tree.

Another option to expand the study would be to include future sequence data from Jasmine. Inclusion of Jasmine would allow me to phylogenetically separate homeologs from the two WGDs (Figure 1.6) and contrast the diploidization rates of the two sets of homeologs independently of each other. Contrasting the effects of homeolog age on diploidization would be particularly interesting for investigating Lag Time hypotheses since the homeologs in question would coexist inside the same organism and thus share a complete history where the only experimental variable is the age of a gene copy divergence. Contrasting in this way would also help to separate out the effects of species bottlenecks from divergence age. For example, in Figure 4.6 why are there only 27 losses in 2 million years in the branch immediately following *F. cuspidata's* speciation? One possibility which could be tested by this setup is that this branch is still to soon after the WGD.

4.4.2 Repeatability of Homeolog Loss

The results suggest the repeatability of gene loss is related to a linear scalar attribute of the gene family, rather than a separate category of gene families. Close inspection of Figure 4.8 shows there is no indication of a second peak that would indicate a second category of genes. Most observed losses were polyphyletic and towards the tips of the species tree (Figure 4.6). By itself, this would be evidence of incomplete annotation, however losses correlate strongly across diverse species, showing either convergent evolution or an intrinsic property of the gene that makes annotation less reliable, e.g., short or repetitive sequence. In either case, the polyphyletic pattern itself is highly significant and the species tree does not explain the cross species overlap of gene losses.

Current estimates of annotation error rates from CAFE can account for between one third and one quarter of the observed variation in gene copy number (4.3.2). If the error rate were 3x higher, the majority of results could be explained away as artefactual. Even the correlations and highly statistically significant patterns could be explained in terms of the deeply shared genetic structure within the genus which makes all downstream observations and events deeply correlated.

To further reject this null hypothesis, I could establish that lack of annotations were caused by true lack of necessary sequences. Every study is of limited scope. Given more resources, the next steps would be to define clear criteria for the depth and synteny expected for true positive alignment in non-pseudogenized genes that were not annotated and establish if there are reads that map in the species ostensibly missing those genes. It would also be helpful to analyze how overlapping genes affected the results. The Extractor tool in 2.2.4 was used to remove overlapping genes unlike Kelly et al. (2020) which returned an average of 4,530 less gene models under nearly identical settings and inputs.

4.4.2.1 Support For the Lag Time Model

Fraxinus data shows an increase in homeolog loss in the last 6 million years, which matches what is predicted by the Lag Time Model from literature (Robertson et al. 2017; Cheng et al. 2018; Clark and Donoghue 2017). In Figure 4.7, the first 30 million years after the WGD show a lower rate of loss of homeologs. The Lag Time Model predicts significant evolutionary time is necessary for dosage-sensitive genes to differentiate enough through subfunctionalization or other means so one copy can be lost without significant detriment, specifically Robertson et al. (2017) reports a 50 million year time frame for Salmonid lines to fully diploidize. By 54 million years after the first WGD (30 + 24 My between WGD) there are already ten extant species, so subsequent loss of these previously dosage-sensitive genes would necessarily be polyphyletic. An alternative explanation is that the Lag Time Model was developed to explain recurrent methodological artefacts (4.4.3.1).

4.4.3 Cause of Repeated Loss

The null model assumes every homeolog has the same probability of loss and repetition is by chance. The results irrefutably show that gene copy losses are concentrated in a subset of gene families. Actual observations are most consistent with a continuous trait, enriched in a subset of gene families, which makes gene copy loss more probable within that gene family.

However, this does not necessarily mean that natural selection is the only explanation for the repeatability of gene loss. A wide variety of factors, both biological and technical, will be specific to gene families and capable of explaining a higher tendency towards loss or retention.

4.4.3.1 Some Gene Loss Overlap May Be Explained by Assembly Quality

Gene set overlap may be explained by a mix of genome assembly quality and species tree topology. Notice *F. cuspidata* which by the species tree in Figure 4.6 is the most divergent species after *Olea*, yet has a similar trend of overlap in the heat map as *F. excelsior*. This level of agreement can best be explained by its rank of 8th highest assembly quality in Table 3.2. The hypothetical scalar property of gene families mentioned in 4.4.2. could simply be the gene family's sensitivity to assembly errors. Families which are more sensitive to poor assembly are underrepresented in low quality genomes and highly correlated, regardless of the genome's position in the species tree. In genomes with high assembly quality, a low degree of overlap means that losses among homeologs are not correlated and are mostly independent from each other. Therefore, quality issues appear to only be affecting the genomes with N50 < 7 kbp (Table 3.2). Doubling N50 in these genomes would likely bring scaffold size above the range where annotation errors are an issue in the future.

4.4.4 Parental Subgenome Dominance

In neo-polyploids one parental genome is frequently observed showing preferential expression and retention over the other subgenome (Buggs et al. 2012; Cheng et al. 2018; Yoo, Szadkowski, and Wendel 2013). Hypothetically, the presence of one copy makes the loss of the other copy more likely because it is invisible to selection or even overdosed (Emery et al. 2018).

In Garsmeur et al. (2014) the authors discuss two distinct classes of paleopolyploidy determined by either biased or unbiased fractionation. Based on the lack of parental dominance, *Fraxinus* appears to be a Class II unbiased fractionation lineage, along with banana, poplar, and soybean. They propose the difference is due to ancient allopolyploids versus autopolyploids, but there are now known counterexamples. Oleaceae and *Cucurbita* are ancient allopolyploids and show unbiased fractionation (Sun et al. 2017; Julca et al. 2018). For *Olea europaea* the picture is complicated because there are three WGDs, each of which could be allo- or auto-polyploids.

It is conceivable there is biased fractionation of the genome which is not detectable without assigning each homeolog subtree to a parental genome and then pooling the results for each subgenome. If the "shoulder" in Figure 4.10 was all phased to the same subgenome it would match more closely with the preferential retention observed elsewhere (Emery et al. 2018). Phasing could conceivably be accomplished using the *F. pennsylvanica* chromosome assembly. However, subgenome assignment has challenges of its own and is beyond the scope of this chapter (Edger et al. 2018).

4.4.5 Functions of Homeologs

GO term enrichment was carried out because the majority of the theories and models in the literature center around the interaction between gene function and gene copy number under selection. It is possible to show repeatability of evolution without any strong connection to function if it is driven by other factors such as position on the chromosome. Still, it can be noted where the GO term enrichment parallels findings from other studies.

The broadest set of Oleaceae homeologs was enriched for transcription factors, though not specifically developmental genes. Thomas, Pedersen, and Freeling (2006) report *Arabidopsis* homeologs are enriched for ribosomes and transcription factors. Rensing (2014) and Edger and Pires (2009) tie morphogenic transcription factors to plant development and use this category as a proxy for dosage sensitivity. The inclusion of ribosomes, which are already in high copy number in the absence of WGD, raises the question of whether these are simply filtering for categories which are generally high copy number to begin with. Other functional enrichment trends were either not present or not observable in *Fraxinus*. Unlike *Arabidopsis* and *Populus*, the only enriched signal transduction was gibberellin signaling (Rodgers-Melnick et al. 2012; Thomas, Pedersen, and Freeling 2006).

4.5 Conclusions

In this study, I used numerical simulations to measure the differences between chance distributions and convergent evolution in *Fraxinus*. A subset of gene families were found to have lost a disproportionate number of copies while other homeologs were retained over a 60 million year time frame. These losses were polyphyletic, recent, and significantly overlapped across species, but did not show evidence of biased fractionation.

I have demonstrated it is feasible to use 28+ species to analyze a single WGD for repeatability of evolution. Alignment and annotation of large numbers of genomes entails a host of technical challenges. It is never possible to completely eliminate the possibility that correlated results are due to both similarities in biology as well as shared methods. The next chapter will examine Graph Genomes as a way to capture the full genetic diversity of an entire species in a single alignment data structure.

Chapter 5
A Scalable Approach to Pangenome Visualization

Author's Contribution

Collaboration Notes: This chapter was written solely by Josiah Seaman while working in collaboration with the 1001 Genomes Project. Portions of this chapter are planned for inclusion in an upcoming publication of Pantograph.

The Pantograph Team includes: Josiah Seaman, Simon Heumos, Andrea Guarracino, Artem Tarasov, Bonface Munyoki, Christian Kubica, Christine Seaman, Dmytro Trybushnyi, Eloi Durant, Hannah Sewell, Jack Tierney, Jacob Windsor, Jerven Bolleman, Jörg Hagmann, Katherine Innamorati, Njagi Mwaniki, Robert Fornof, Mark Seaman, Michael R. Crusoe, Stacie Seaman, Thomas Townsley, Torsten Pook, Toshiyuki T. Yokoyama, Travis Clark, and Erik Garrison.

Josiah Seaman wrote the entire text of this chapter. Pantograph was a highly collaborative project with JS serving as the main originator of visualization concepts, the project manager, and lead developer regarding JavaScript. Pantograph arose from hackathons held in Fukuoka, Japan in September 2019 and Tübingen, Germany in November 2019, at which the concepts were developed and preliminary implementations made. The pangenome visualization schematic designs were developed by Josiah Seaman in discussion with other team members, and Artem Tarasov wrote Component Segmentation code. Torsten Pook integrated HaploBlocker and developed chromosome recombination breakpoints; Simon Heumos worked on the RDF database, programmed the browser, and evaluated the graph sorting algorithm; Jerven Bolleman developed an RDF representation of graph genomes and built data access tools; Erik Garrison had previously developed odgi including binning and pangenome matrix concepts and extended it to support this project, as well as being a consultant for the project's direction. Pantograph development was started in earnest at a virtual Hackathon April 2020 with daily meetings led by Josiah Seaman and Simon Heumos. Josiah Seaman, Bonface Munyoki, Mark Seaman, and Stacie Seaman did project management. Christine Seaman did product testing. Pantograph was presented at ISMB 2020 by Josiah Seaman, Simon Heumos, Toshiyuki Yokoyama, Torsten Pook, Jerven Bolleman, Andrea Guarracino, and Thomas Townsley. Christian Kubica and Sebastian Vorbrugg generated the *A. thaliana* pangenome as part of the 1001 Genomes project.

COVID-19 Foreword

Pantograph as a project started in 2018 as a way to unlock the next level of population genetics for researchers for organisms including bacteria, ash trees and humans. Graph Genomes discussed here are a new way of capturing sequence data designed to fix problems systemic to the technology we've been using for the past 30 years. For example, reference bias means genetic analysis is more accurate for Europeans than for Africans (Liverpool 2019). Pantograph was being developed by a team of ten scientists before COVID-19 emerged as a global pandemic.

We quickly realized that Pantograph could be extremely relevant to the current pandemic because the success or failure of our efforts to fight the disease rely upon the sequence diversity of the virus itself. Tests for infection rely on knowing the exact sequence being tested. A rearrangement in the order of genes, even if the content is the same, will return a false negative test if the rearrangement changes the order of the target sequence. Second, the vaccine targeting the Spike protein on the outside of the virus relies on a lack of genetic diversity in the Spike protein sequence ("NIH Clinical Trial of Investigational Vaccine for COVID-19 Begins" 2020). If there are any strains with a mutant protein, the vaccine could be rendered ineffective and the virus would continue to spread. SARS-CoV-2 is an RNA virus that will likely infect billions of people, giving it a much higher mutation capability than has been previously dealt with in pandemics. For example, the common cold is impossible to eradicate precisely because of the number of people infected and thus the high number of mutations which exist around the globe (Kistler et al. 2007). Current sequencing techniques may be under-representing the full sequence diversity of the virus because they are reference based. Eliminating reference bias and enabling species genetic diversity on thousands of individuals is the core goal of using a graph genome.

Thus, COVID-19 became the central focus of Pantograph development. Pantograph is a very small piece in a worldwide effort to eradicate this disease. It is by no means the most important and the disease will likely attenuate without any involvement on our part. However, given the scale of the pandemic, even a tiny improvement or speedup can result in thousands of lives saved. That's a difference which is worth investing our resources in. Pantograph will continue to be useful in a wide range of disease applications after the current crisis is averted, so we are never caught unprepared again.

Abstract

A single reference genome does not capture the genetic variation within a population. Reference bias is caused whenever genome representations in current use cannot capture variants in sequence absent from the reference. While we have effective multiple sequence alignment browsers for genes, we now need the ability to visualize the full genomes of hundreds of individuals. These limitations are holding back progress in understanding the genetic basis of important traits. Genome graphs are a recent solution to summarizing all variation within large sets of individuals, but we lack scalable approaches to their visualization. Structural variants are associated with many diseases, for example, cancers, mental illnesses, immune disorders, and most recently COVID-19. In order to make sequencing data actionable for clinicians, structural variants need to be put in the larger context of all known genetic variation.

Here we present the design of Pantograph, which promises to be the first visualization tool with the scalability to render graph genomes from thousands of individuals over gigabase genome sizes with the ability to show both whole chromosome features as well as zoom into nucleotide sequence variation. No tool to date has satisfied all these criteria. Scalability is accomplished through graph sorting and binning adjacent sequences to create shared syntenic blocks called Components. Non-linear structural variants, called Links, are treated as a single feature which can be shared by many individuals. By separating the pangenome into syntenic Components connected through Links we have created a browser capable of displaying a graph at variable levels of complexity; making complicated alignments comprehensible to the researcher. Here, this tool is applied to *A. thaliana* and SARS-CoV-2 and lays the foundation for a *Fraxinus* graph genome.

5.1 Introduction

5.1.1 Graph Genomes as a replacement for linear reference genomes

Graph Genomes are a major disruptive technology now emerging in bioinformatics that may hold the key to solving deeply rooted problems in handling genetic variation (Figure 5.1). Sequencing costs have been dropping exponentially for decades, making bioinformatic analyses the primary cost bottleneck and barrier to new discoveries. Single linear reference genomes are simple, easy to understand, easy to implement, and have been used in bioinformatics for over twenty years. However, they have intrinsic limitations when representing genetic variation. New sequencing technology has enabled researchers to deeply sequence the variation within model organism populations, revealing many alternative loci and novel insertions not present in the original reference genome (Telentia et al. 2016). Simultaneously, there is increasing evidence of function in non-protein coding regions of the human genome and increasing evidence of function in the 3D interactions of chromatin folding (ENCODE Project Consortium 2012; Rosenbloom et al. 2013; Pennisi 2012; Lieberman-Aiden et al. 2009). Structural rearrangements in functionally relevant sequences can cause long range and unexpected phenotypic consequences such as cancer (Gryder et al. 2017).

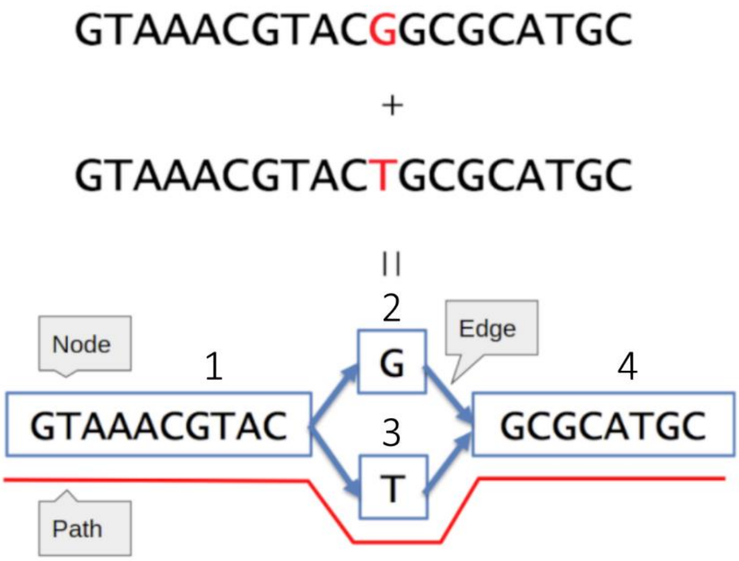


Figure 5.1 Simple Graph: Graph Genomes store sequence in nodes identified by numeric IDs. A genome (or chromosome) is a Path through a series of node IDs. For example, Genome A could have path 1,2,4 and Genome B could have the path 1,3,4. Node 1 and 4 would be conserved in both genomes, whereas Nodes 2 and 3 would be alternative alleles found in the same position relative to Node 1. Nodes are bidirectional so a Path can visit the minus strand 2- to represent an inversion and nodes can be visited in any order to represent any rearrangement. Nodes can be large enough to contain genes or as small as a single nucleotide to store SNPs. **Image source**: Masahiro Kasahara for the Pantograph ISMB poster, used with permission.

Until now, technology has been directed towards short reads which bias results towards the discovery of Single Nucleotide Polymorphisms (SNPs) and short indels. Researchers have therefore focused on disease etiologies that are possible to understand through the data available. To move to complex epistatic interactions involving large scale structural rearrangements and chromosome conformation changes, data arising from long reads and phased assemblies must first be made accessible to researchers. Variant Call Format (VCF) is the current most common format for storing the genetic variation within a species. Variants are called relative to a reference genome coordinate frame and a major versus minor allele. The format is not capable of storing certain kinds of variation, such as an insertion within an insertion (Yokoyama et al. 2019). The ideal format should also have a single representation to allow equality checks. VCF has four different ways to represent an inversion depending on which tool it came from. It does not specify which unit of a tandem repeat was lost in the case of copy number variation.

Graph Genomes can represent complex structural variants among many genomes in a self-consistent format (Yokoyama and Kasahara 2020). Graph Genomes do not contain the same reference bias problems inherent to assigning a major and minor allele at every position and using a single reference coordinate frame. For example, a study in humans found that reference guided tools overestimated the differences between two Africans when using a European reference genome (Liverpool 2019). By way of analogy, this is equivalent to translating from Japanese to English then to Mandarin Chinese, instead of translating directly from Japanese to Mandarin. Graph Genomes allow a more direct translation of findings, coordinates, and shared variation (The Computational Pan-Genomics Consortium 2018, Paten et al. 2017). We predict the next generation of reference genomes will be Graph Genomes that more fully contain the knowledge acquired about the total genetic diversity of the species under study (Ballouz, Dobin, and Gillis 2019), (X. Yang et al. 2019). This change is already under way in humans and *A. thaliana*.

While Graph Genomes require new tools, they also make many tasks much easier. With a reference Graph Genome, it is possible to do fast and accurate haplotype mapping of complex sequences such as the MHC region of the human genome, which is clinically relevant for immune response and precision medicine. The significant divergence of MHC region has impeded accurate genotyping thus far. Graph reference increases the mapping rate for these regions (Garrison et al. 2018, Dilthey et al. 2015, Dilthey et al. 2016). This allows allele-specific expression levels to be inferred from RNA-seq data. Graph Genomes can also enable alternative-splicing-aware alignment, comparative genomics, and increased accuracy of reference guided assembly.

These examples show the rich potential of a Graph Genome based approach. Some pipelines are already being explored, others are still under development (Llamas et al. 2019, Garrison et al. unpublished). Therefore, the continuous development of graph tools is needed.

Graph Genomes as defined in Figure 5.1 deviate from generic graph data structures in some crucial aspects. Graphs have existed as a field of mathematics for over a two centuries (Biggs et al. 1986). A graph is typically composed of a set of Nodes interconnected by Edges, which can be directional or unidirectional. Both Nodes and Edges can be decorated with other attributes. Some tools from graph theory are usefully applied to Graph Genomes such as the minium number of cuts to make a graph acyclic. However, Paths and the large scale biological linearity of Graph Genomes deviate from the generic concept of graphs in important ways.

To keep the biological context in mind, users will want a visualization that maintains a pseudo-linear format (akin to a braid), whereas most graph approaches make no assumption about network topology by default. Custom created visualizations for a specific task consistently out-perform generic tools (O'Donoghue et al. 2018, page 281). Visualization tools help to not only understand genome variants, but also improve and debug Graph Genome construction.

5.1.2 Visualization Tools

Clear visualizations for browsers are necessary to facilitate human reasoning about the complex interactions between structural variants and biological questions. As the scale of input data increases, our visualization techniques must also be able to scale to thousands of eukaryote genomes. A review of currently available Graph Genome visualization tools shows many desirable features and a wide range of techniques, however, none of them can scale to thousands of individuals with gigabase genomes while maintaining the nonlinear nature of a graph. The scalability of an algorithm can be calculated mathematically based on what is being rendered. In this chapter, I focus exclusively on tools that influenced the design of Pantograph and are thus comparable in approach. For a complete overview of the field see Eizenga and Novak (2020) and Yokoyama and Kasahara (2020).

Previous work on Graph Genome visualization follows three main areas: 1) force directed layouts (Figure 5.2), 2) Sequence Tubemaps (Figure 5.3) and 3) rectangular matrices (Figure 5.4).

Force Directed Layout visualizations (Figure 5.2) use a physics model to layout graphs in a natural way. Its strength is aesthetic appeal and clearly communicating components and topological complexity but annotation and navigation is more challenging than a linear layout. Bandage and Jason Chin's Graphviz both utilize layout methodologies that predate graph genomes (Wick et al. 2015; Chin 2019; Ahmed A. et al. n.d.). A major issue with these methods is their runtime scalability. Force directed layout has quadratic or even cubic costs with respect to graph size. Whilst heuristics exist to make force directed layout practical for large graphs (Barsky et al. 2008), they are in general not effective for graph genomes because in these graphs, their edges do not fully capitulate the spatial layout that allows them to be most effectively interpreted. An ideal Graph Genome layout will be predominantly linear while allowing for the possibility of non-linear rearrangements across large distances in the pangenome. Equally difficult, inversions require a clear sense of Edge directionality be communicated which is much easier to maintain when the visualization has a strict left to right orientation to reference off of.

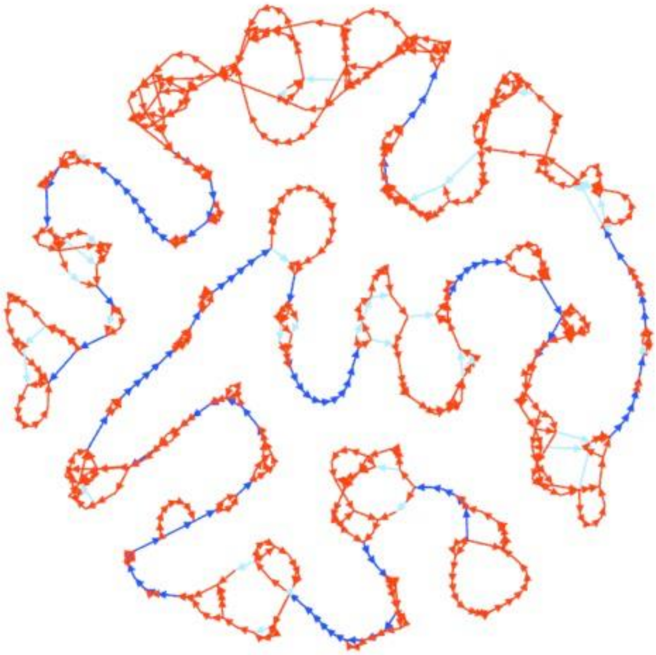


Figure 5.2: Force Directed Layouts: Jason Chin's Graphviz utilizes preexisting layout methodologies to visualize graph genomes (Chin 2019). He divides a variant graph into two categories: conserved regions are marked in blue and share a single set of nodes between all individuals. Variable regions are marked in orange and can contain alternate alleles for different individuals. This keeps the overall linear structure while allowing for local non-linearity. Deletions are shown as faint shortcut edges for some individuals.

Sequence Tubemap uses nodes containing sequence and visualizes paths as colored bands which visit nodes (Beyer et al. 2019). Tubemaps are very clear at communicating rearrangements and users can follow a single path with very little training (Figure 5.3A). Sequence Tubemap is a novel enough approach it merits special discussion here. MoMI-G also integrates its own modified implementation of Sequence Tubemap to visualize genomic rearrangements (Yokoyama et al. 2019). Sequence Tubemaps will no doubt go on to influence projects in the future.

Figure 5.3B shows scalability challenges in real world data. For nonlinear ordering, paths must be rendered above all nodes in between, creating undesirable pileups around high traffic nodes. This approach suffers from scalability issues for rendering thousands of individuals. It uses nodes for both linear SNP differences and non-linear rearrangements, meaning that as the number of individuals increases, fragmentation and vertical stacking also increase linearly. Sequence Tubemap could be made performant, that is, with increased speed, by only rendering Edges; rendering which variants exist in the population without rendering which individual has each variant. Examples of where Sequence Tubemap scalability breaks down can be seen in Supplemental 4.

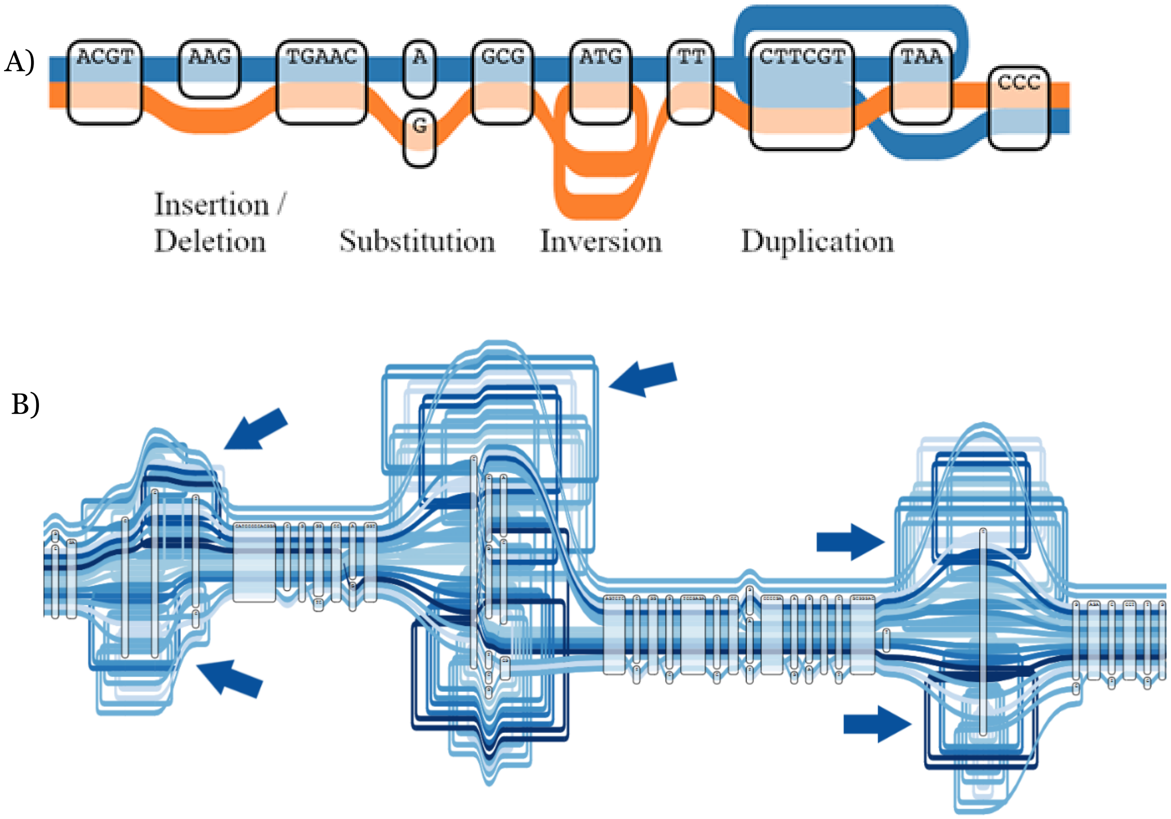


Figure 5.3: Sequence Tubemap: A) Example visualization given in Sequence Tubemap publication (Beyer et al. 2019). Each colored path (blue or orange) represents one chromosome which visits Nodes (outlined) containing sequence. The Tubemap analogy is particularly appropriate for communicating traversal direction in inverted segments, however the visual clutter for inversion is high when the original and inverted start positions are far apart in the visualization. In simple examples Tubemap is both compact and easy to understand, but when applied to complex examples, scalability challenges become apparent. B) Human HLA regions in 12 individuals. Tangles (arrows) are due to the need to retread all nodes laid out in between nodes relevant to the path, even when that path does not interact with any of the intervening nodes.

Odgi uses a rectangular Matrix View (Figure 5.4) (https://github.com/vgteam/odgi/). The Matrix View's strength is in clarity for showing presence/absence visualization. It works at the single nucleotide level and can extend to the full chromosome level (5.3.7), which gives it many of the scalability properties necessary to render thousands of paths. Its one shortcoming is that odgi bin has not been designed as an interactive visualization. A single image is simply too small to capture all of the rearrangements in a pangenome while still allowing it to be decipherable. Nonlinear links are nearly impossible to follow by panning over a large image. Interactivity allows for selective filtering, search, and jumping to relevant features in a complex pangenome. If odgi bin had interactive navigation of rearrangements it would satisfy most of the criteria necessary for a scalable Graph Genome browser.

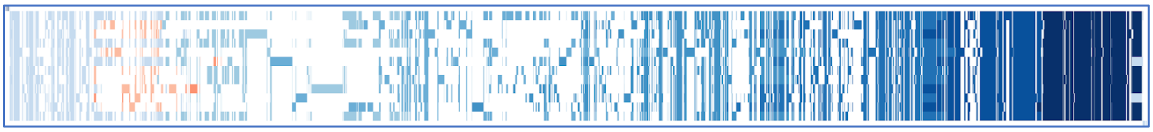


Figure 5.4: odgi Matrix: 12 *Arabidopsis thaliana* individuals on chromosome 1 from the 1001 Genomes project (5.1.3). Each row is one individual. X-axis roughly corresponds to position along the chromosome due to pangenome sorting (5.3.4). Cells indicate the presence or absence of a Bin in an individual (5.3.7). Bins are shaded by Path position which shows that all individuals largely agree in overall ordering across the chromosome. This would be expected for organisms still capable of sexual recombination. Red bins indicate regions which are inverted in eight individuals on the upper chromosome arm. One can also see many private insertions in the middle where unique sequence variation close to the centromere is not shared between other individuals.

5.1.2.1 The Density Problem

As graphs become larger, their topology can become more complex, because the number of potential connections (edges) grows with the square of the number of nodes. The number of possible paths through those edges grows even faster, and in the case of fully connected graphs, increases factorially with the number of nodes. However, in graph genomes, the number of variants (nodes) which can be realized is practically limited by biology, and the number of paths is at most the number of observed organisms involved. Though it is worth noting that draft genomes introduce many more paths, graph genomes are in general sparse, and thus when visualized as matrices, they are predominantly white-space (Henry, Fekete, and McGuffin 2007), this is demonstrated in Figure 5.5.

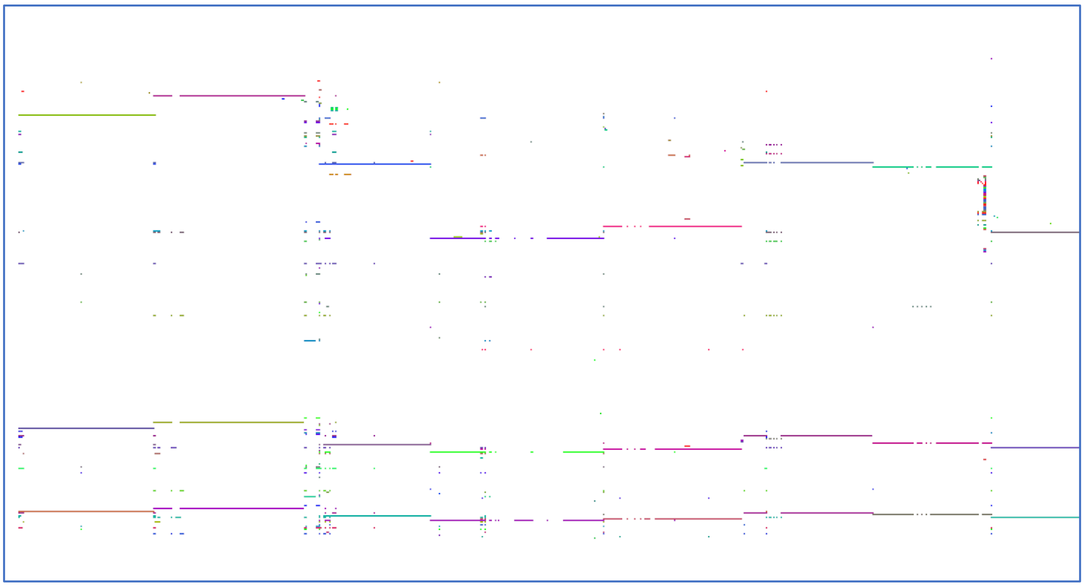


Figure 5.5: Density Problem: Erik Garrison's odgi Matrix using human genome data. Used with permission (Garrison, 2019, Private Communication).

Sequence Tubemap suffer from a similar issue when path stacking pushes the layout of sequence nodes further and further apart. As the number of paths or copy number increase, the percentage area of the image which communicates sequence approaches zero. Force directed graphs have the opposite challenge caused by undesirable line crossing of unconnected paths which is unavoidable in two-dimensional space. This leads to large datasets becoming indecipherable “hairballs”, a problem fundamentally rooted in their structureless nature.

Here we introduce the design and implementation of Pantograph, a new Graph Genome browser made to address the scalability challenges of showing the complete genetic diversity of one thousand individuals of a species simultaneously. Pantograph combines the clear presence/absence matrix of odgi bin with a hybrid of Sequence Tubemap connectors that maintain browsable information about non-linear rearrangements, called Links. Collectively, this layout paradigm that the software tool Pantograph uses to visualize data is called Pangenome Schematics for its ability to enumerate all the genetic variants which exist in the population. Care has been taken in the design to reuse the same rendered element in every individual who inherited the same variant to reduce visual clutter. Pedigree and lineage relationships can also be exploited to compress and further optimize visualization of the pangenome (see 5.3.8).

5.1.3 1001 Genomes Project

Pantograph is being used in the 1001 Genomes project to quality check assemblies, alignments and the Graph Genome construction process in a similar manner to how FluentDNA was employed in 3.3.2. The 1001 Genomes project seeks to sequence and assemble 10 individuals from 10 different ecotypes in 10 regions around the world plus the reference TAIR10 assembly, totaling 1001 genomes.

The pilot project graph has all five nuclear chromosomes from 12 individuals assembled together with links between them. This graph is then used to diagnose challenges in the assembly of a larger graph involving 24 individuals, which in turn form the foundation for a larger resequencing effort in the future. In the Results, I show how Pantograph can be used to investigate large scale features in chromosome graph assemblies and address scalability problems as they arise. This same browser can be used to zoom in and investigate nucleotide level variation in *Arabidopsis thaliana* or as demonstrated in the SARS-CoV-2 pangenome.

5.2 Methods

5.2.1 Quantifying Performance Characteristics

The scalability of an algorithm can be calculated mathematically based on the method of rendering. Computer scientists use Big-O notation to describe how an algorithm will scale given arbitrarily large inputs (Knuth 1970). This analysis can be carried out as a mathematical proof on a design without the need for a running program or actual computer because the key aspects defining the scalability of an approach are in the design, not the implementation. This section briefly describes the method for mathematically evaluating the scalability of a visualization. All aspects of the design of Pantograph are covered in 5.3.

The criteria for evaluating the scalability of a visualization are:

1. The number of elements rendered (minimize)
2. Layout time for the visualization (minimize)
3. Number of processed individuals demonstrated with non-linear rearrangements in eukaryotes (maximize)

For criteria 1 and 2, designs are judged based on a theoretical dataset containing N nodes, P paths, S SNPs, and R structural rearrangements. The Big-O notation is the equation which tells either the number of rendered elements or layout time in arbitrary units. Number of rendered elements (1) is a proxy for both visual complexity and rendering time independent of rendering platform. Layout time (2) includes all computation required to setup the visualization, and is usually a one-time precompute.

Finally, Demo Individuals (3) are based on concrete examples from this study or found in the literature. Examples must include structural variation beyond simple insertions and deletions. These examples are by no means hard maximums and are most likely to change.

5.2.2 *Arabidopsis thaliana* graph preparation

The pilot graph was constructed with the following 12 representative individuals: TAIR10, AT9784, AT7328, AT6906, AT1741, AT6909, AT7213, AT5784, AT6911, AT7186, AT6981, AT9518. Only sequence from the five nuclear chromosomes was used, but no pre-filtering was done for mapping to a reference genome. Graph sorting was Path guided stochastic gradient descent (SGD) (Recht et al. 2011) on 32 cores with 1TB of RAM available. We found experimentally that 128 GB of RAM was not enough to sort the graph. The command used was:

```
odgi sort -Y -i sebastian.Athaliana.all.50000.gfa.odgi -o
Athaliana.all.50000.gfa.odgi.sorted -t 32
```

To check the extent of evidence for ancient whole genome duplications in *Arabidopsis thaliana* the TAIR10 v26 assembly was used on CoGe using their SynMap2 tool (Lyons and Freeling 2008, Haug-Baltzell et al. 2017). CDS was based on Arabidopsis_thaliana.TAIR10.26.gff3 (Lamesch et al. 2012). Note that dot plots are frequently only computed over CDS rather than whole genome content in order to exclude repetitive elements.

5.3 Pantograph Design

Our method for visualizing pangenomes combines approaches used in odgi bin Matrix and Sequence Tubemap and builds upon them. Pantograph is implemented as a pipeline of software modules, each step is explained further in its own numbered section (Figure 5.6). The approach is as follows:

1. Segment the pangenome into shared syntenic blocks by identified the minimum set of structural rearrangements (5.3.1)
2. Use a rectangular Multiple Sequence Alignment (MSA) to represent those syntenic blocks present in the population that are not broken by segregating structural rearrangements, using columns rather than colors to represent SNPs (5.3.2).
3. Use colorful Links representing structural rearrangements to join these MSA blocks (Components) (5.3.3).
4. Minimize the number of Links needed using gradient descent to sort the pangenome (5.3.4).
5. Use color within the MSA to represent copy number variation and inversions (5.3.5).
6. Annotated features with its own path and coloring (5.3.6).

7. Enable zooming from nucleotides, to gene regions, to whole chromosomes by binning of sequence content (5.3.7).
8. Cluster related individuals by sorting the rows into haplotypes (5.3.8).
9. Reduce graph complexity by unrolling copies of distributed repeats (5.3.9).
10. Scale the width of variants by their frequency in the population (5.3.10).

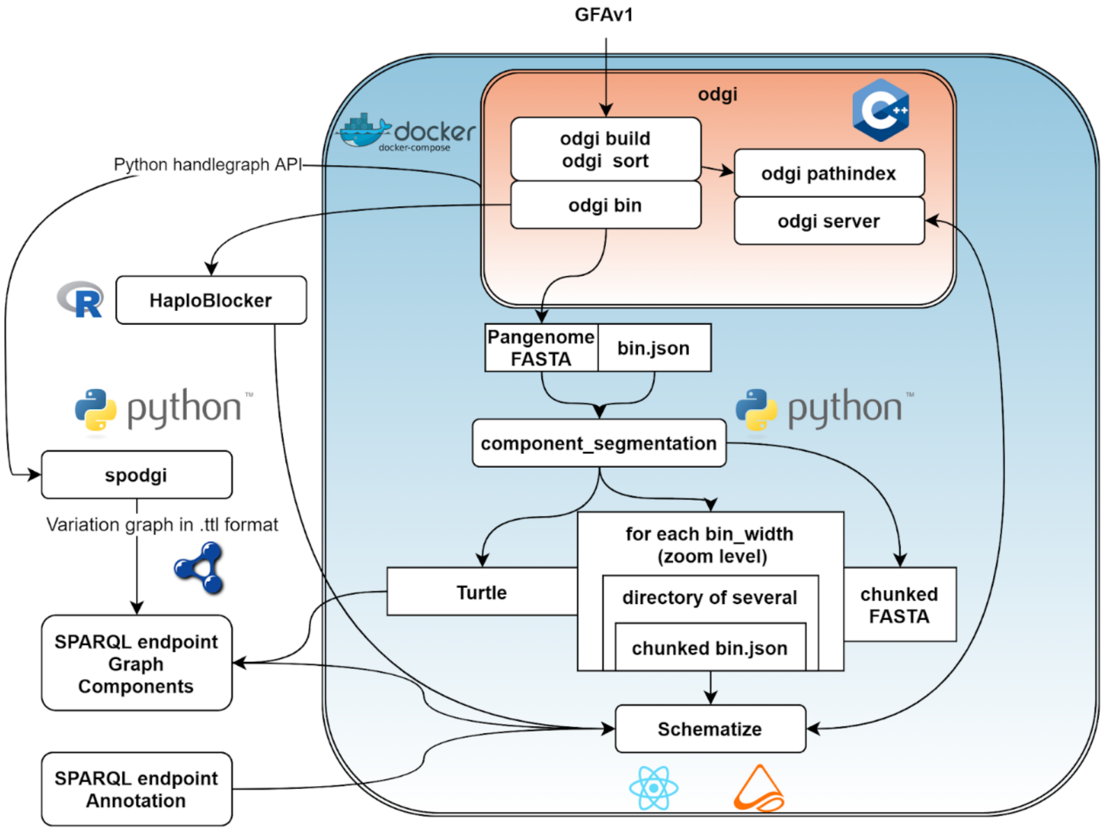


Figure 5.6: Modules of the Pantograph Data Pipeline: Pantograph takes GFA Graph Genome files as input and provides a React JavaScript browser as output (github.com/graph-genome/pipeline). GFA is a common, human readable format for storing Graph Genome files. First, odgi is used to read the graph into memory, sort the nodes, and bin them into a JSON file and a pangenome FASTA. The bin size is set by the user; large bins create small files with more information lost. These bins are segmented into collinear Components by Component Segmentation (github.com/graph-genome/component_segmentation). HaploBlocker optionally identifies recombination breakpoints and sorts rows by haplotypes (github.com/tpook92/HaploBlocker). Components are read into Schematize and displayed as React components (github.com/graph-genome/Schematize). The ability to jump to a particular nucleotide index in an individual is enabled by the Path Index Server which uses ‘Odgi server’ microservice built on a compressed path index file. Spodgi makes outputs of odgi bin and Component Segmentation available through a SPARQL endpoint by storing the results in RDF (.ttl format) for use in other tools. **Image Source:** Designed by Simon Heumos for Pantograph ISMB presentation (Seaman et al. 2020).

5.3.1 Component Segmentation

Component Segmentation is the key conceptual step underlying the Pantograph pangenome schematic visualization paradigm. It breaks the pangenome into sections based on the presence of major structural rearrangements. Each section is called a Component. The algorithm seeks to minimize the number of Components better because they represent rearrangements within the population. Too many rearrangements create too many Components, which makes the visualization less clear to understand.

Therefore, efforts are taken to merge similar structural variants into a single shared feature. Sorting (5.3.4) maximizes the contiguity of Components and places likely mergeable variants next to each other. Next, the pangenome coordinates of all rearrangements are listed as possible dividers. For each divider, Component Segmentation checks the occupancy until the next closest rearrangement. If no individuals have any intervening sequence, then the two rearrangements can be merged into a single divider without loss of information. Scanning for a viable merge is done for the next and previous divider for each divider. Whenever a merge is successful, the new divider is rescanned. After one pass through the pangenome, every divider will be constrained, and no further merges are possible.

The reference implementation of Component Segmentation is written in Python and is available at https://github.com/graph-genome/component_segmentation. This implementation has received extensive optimization work, resulting in over 110x reduction in runtime since optimization work started (https://github.com/graph-genome/component_segmentation/issues/20). The only obvious way to optimize it further would be to integrate the logic directly into odgi in order to give it direct access to the graph object model in memory, removing the need for file I/O.

5.3.2 Pangenome Matrix Depiction of Single Nucleotide Variants

In a multiple sequence alignment (MSA) browser such as Jalview, SNPs are represented by the presence of different nucleotides/amino acids in the same column (Waterhouse et al. 2009). In a Pangenome Matrix, each variant has its own column and is shown as either present or absent in each individual (Figure 5.7). In this binary representation of the multiple sequences, SNPs, insertions, and deletions are all represented in the same way.

However, the matrix itself is not an efficient means of depicting nonlinear structural rearrangements. Therefore, the matrix is broken up into blocks, called Components, where the ends of each Component correspond to the break-points for inversions and translocations in the genome (5.3.1).

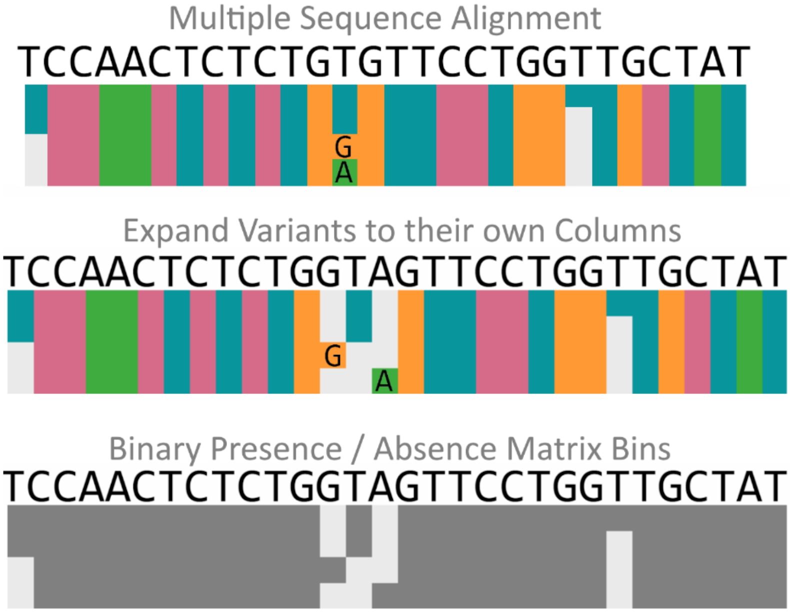

Multiple Sequence Alignment
TCCAACTCTCTGTGTTCCTGGTTGCTAT
G
A
Expand Variants to their own Columns
TCCAACTCTCTGGTAGTTCCTGGTTGCTAT
G
A
Binary Presence / Absence Matrix Bins
TCCAACTCTCTGGTAGTTCCTGGTTGCTAT

Figure 5.7: Pangenome Matrix: Top: Example of SNPs using color to communicate nucleotide information in an MSA. Middle: Trivariate site in the center is expanded to three columns. The color coding becomes redundant since all information is encoded in position plus the pangenome sequence. Bottom: Color is freed to communicate other pieces of information like inversion or copy number.

5.3.3 Links to Join Components

Pantograph uses colored arrows to show Links, structural variations present in the pangenome. These join different components in different ways for each individual. A Link is an alternative Edge representing a structural rearrangement that shows the order of matrix components in each individual (Figure 5.8A). Graph Genomes use edges for every kind of variant. Our Component Segmentation software identifies the rare few edges (<1%) that are nonlinear rearrangements and groups individuals sharing the same structural rearrangement (5.3.1 Component Segmentation). If rearrangements are rare, one Component can contain thousands of Graph Genome Nodes.

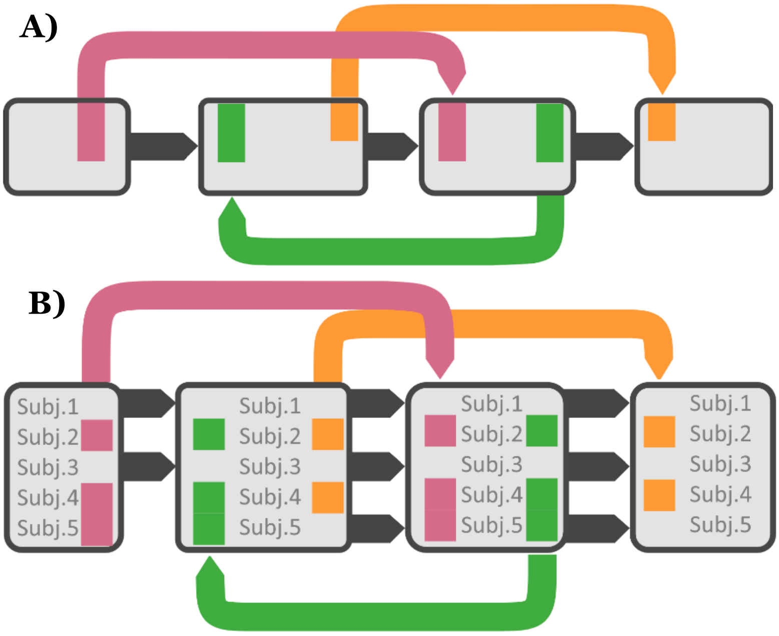


Figure 5.8: Link Columns and Structure Variant Only View: A: Each box is one Component. Adjacent Components are connected by black Links. Alternative Links are structural variants shown in various colors. The allele frequency is shown as a bar for each Link column based on the number of individuals that share the structural variant. B: For more detail, a Link Column underneath each Link can show which experimental subjects contain each structural variant with a colored cell. Individual rows with no rearrangements have a black adjacent Link to the next Component. On the far left, we can see Subjects 1 and 3 do not contain the pink structural rearangement. On the far right, we can see Subjects 2 and 4 end with a structural rearrangement indicated by the yellow Link.

In order to show clearly which Link applies to which individual, we introduce Arrival and Departure columns at the beginning and end of each Component respectively (Figure 5.8B). Links are drawn with an arrow point on the Arrivals side, and presence/absence of that particular Link is shown in the Link Column below for each individual. There can be more than one Departure or Arrival Column at the edges of one component, to show multiple structural rearrangements at the same pangenome position.

To follow a particular individual, the user can hover over an individual to highlight the entire path. Each Link can be used to jump the genome browser to the other side of the Link when the corresponding Component is outside the viewport.

5.3.3.1 Integrating Structure and Matrix Views

Pantograph integrates the concepts of a Pangenome matrix and the structure diagram (Figure 5.8) into a single view called Pangenome Schematics (Figure 5.9). These schematics allow the user to read every nucleotide from every aligned individual in a pangenome and supports all types of structural variants. Examples follow with equivalent Multiple Sequence Alignments (MSA) for reference.

```
TCCAA   ---CTCTCTGTGGTTCC   GGTT   GCTAT*
TCCAA   GGTCTCTCTGTGGTTCC   ----   GCTAT*
TCCAA   ---CTCTCTGGGGTTCC   GG-T   GCTAT*
TCCAA   GGTCTCTCTGAGGTTCC   ----   GCTAT*
TCCAA   GGTCTCTCTGTGGTTCC   GG-T   GCTAT*
```

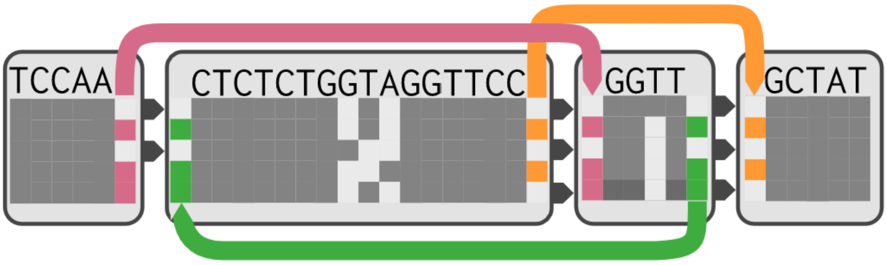


Figure 5.9: Pantograph Schematic Layout for Graph Genomes: Top: The five aligned sequences that were used to generate the Graph with color coded components. Bottom: Schematic showing all information available inside of a Graph Genome: SNPs, indels, structural rearrangements, and copy number variation. The last row of GGTT is colored more darkly because of the two traversals of the same Component. Every component reads from left to right; only follow each Link once.

Components are always separated by sets of Link Columns and black Adjacent Connectors. The grey bounding boxes for each Component can be removed as unnecessary chart junk to give a more simplified visual style (Figure 5.10). The simple style was chosen for the Pantograph 1.0 release because it scales better with large pangenomes.

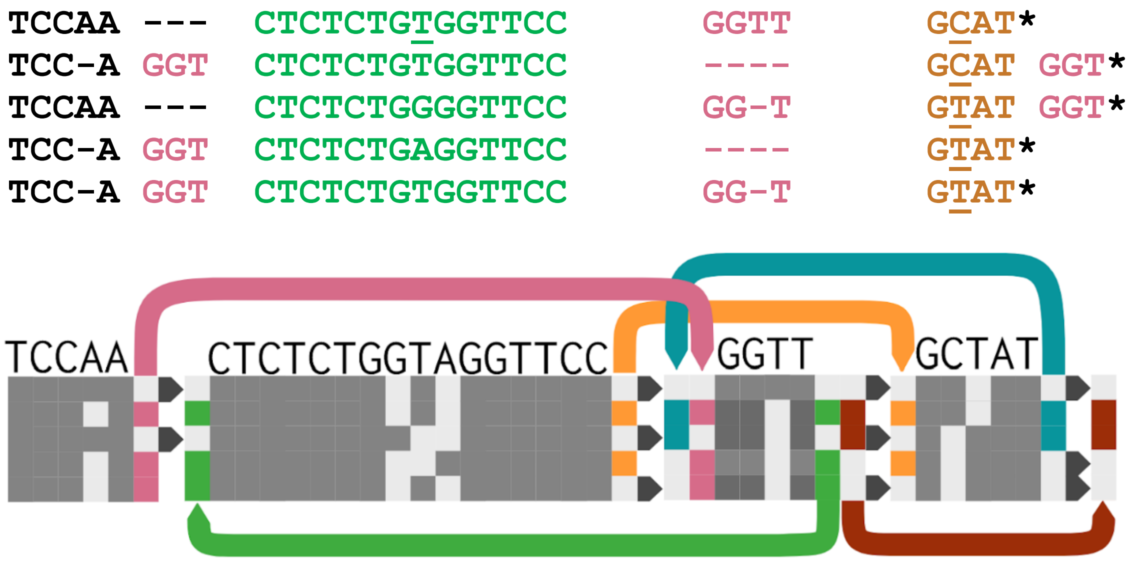

TCCAA --- CTCTCTGTGGTTCC GGTT GCAT*
TCC-A GGT CTCTCTGTGGTTCC ---- GCAT GGT*
TCCAA --- CTCTCTGGGGTTCC GG-T GTAT GGT*
TCC-A GGT CTCTCTGAGGTTCC ---- GTAT*
TCC-A GGT CTCTCTGTGGTTCC GG-T GTAT*
TCCAA
CTCTCTGGTAGGTTCC
GGTT
GCTAT

Figure 5.10: Schematic with Distributed Repeats: Top: Five example sequences in a multiple sequence alignment. Similar sequences have been colored by hand to indicate rearrangements not shown in a multiple sequence alignment. Bottom: Pangenome Schematic showing all the same variation but including the transpositions and duplications. SNPs, duplications, and rearrangements are shown from five individuals.

5.3.4 Sorting the Graph

Graph sorting imposes a single linear coordinate system on the whole graph by determining an order to list Nodes. This is extremely useful for implementation and navigation. The global coordinates do not exactly match a reference genome or any other genome. However, the Graph Genome is already mostly linear. From an evolutionary standpoint, we start with a single individual with a linear genome. Deletions and SNPs add more columns to the matrix, but it is still collinear (syntenic). Inversions and translocations introduce the first truly non-linear variation.

Finding an ideal sort is a difficult problem addressed by the 'odgi sort' subcommand. The goal is to place all syntenic variant Nodes next to each other, then the few rare Links will bridge long distances across the pangenome to describe the unique ordering of each individual genome. A bad sort results in too many Links and chromosomes scrambled together. To address the challenges posed by bad sorting outcomes, a series of sorting algorithms are used in a pipeline to refine the number of Links spanning across the graph genome.

First a topological sort is applied to the graph with a chunksize specified to partition sections to be sorted at each phase (Garrison 2019). Sorting then moves from the previous sorted partition to the next unsorted chunk in the graph. Second, there are two rounds of Stochastic Gradient Descent (SGD) meant to smooth out local problems in the sort by minimizing an energy function when moving a single pair of nodes at a time (Zheng, Pawar, and Goodman 2019). Sorting has diminishing returns, but additional steps can be set by the user.

5.3.5 Inversion

Inversions are shown as an entire component row colored in red for specific individuals. Inversions still use the same Arrival and Departure columns but their sequence is interpreted as the reverse complement of the listed pangenome sequence. The visualization does not pick which strand to present a set of nodes, but graph construction should place the majority of individuals on the plus strand.

```
AGTA -----------TAG-CTACG-TAGCAT T-AG AATTA*
ACTA T-ATT T-AG ATGCTA-CGTAG-CTA*
AGTA -----------TAG-CTACG-TAGCAT TTAG AATTA*
ACTA T-ATT T-AG ATGCTA-CGTAG-CTA*
AGTA -----------TA-ACTAC-ATAGCAT TTAG AATTA*
```

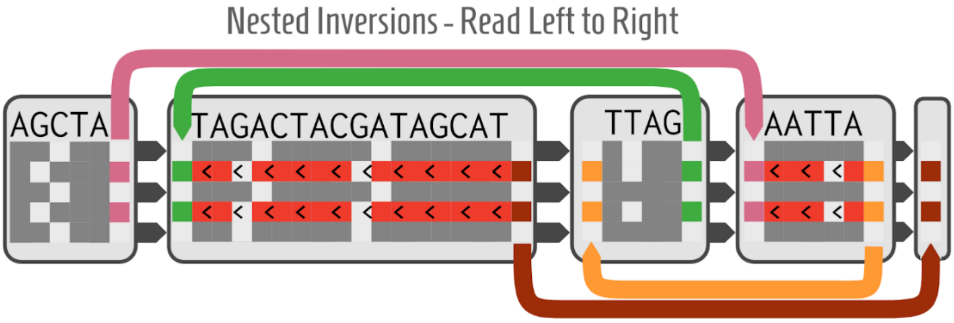


Figure 5.11: Inversions: Top: A best effort MSA of the same five sequences. MSAs are not ideal for showing inversions (underlined) and it's impossible to use a pure linear ordering to place the flipped pink and yellow sequences in inverted positions in this nested inversion. Bottom: Inversions are red rows in the inverted individuals that run the length of a Component. Insertions can be inverted although the upstream is still on the left and downstream on the right, the sequence content is interpreted as the reverse complement. The node AATTA would be TAATT in the inverted individuals. In this example, we show an inversion (TTAG) nested inside a larger inversion (center 3 Components) in order to show how Pantograph handles complex topology in many individuals.

5.3.6 Annotation

Annotation can be carried out as either Path objects in a new row of the matrix, or a rich graphic rendered by JBrowse2 and placed in context of the pangenome (Cain and Buels 2020). JBrowse2 integration is a future planned feature for Pantograph 1.2. The advantage of using JBrowse2 is the immediate ability to render any kind of annotation. We do not wish to reinvent the wheel in displaying different types of annotation: repeats, wiggles, ChIP, etc.

Since all annotations are paths, they are stretched to the same pangenome coordinate system, allowing easy comparison. Overlapping annotation Paths take up more than one row of the matrix, similar to the layout of Sequence Tubemap. More complex annotation types will only be present in a synchronized JBrowse2 panel.

Pantograph contains a script which takes a GFF annotation and GFA Graph Genome as input and outputs a new GFA with the annotation integrated in. The program loads the graph into memory using odgi and then locates the beginning of each annotation using the Path position (5.3.11.4). Annotations typically start in the middle of Nodes, which are the split into two Nodes to mark the start coordinate of the newly created annotation Path. The same procedure is repeated for the end position and any gaps in coverage created by introns. The Path is given a name matching the annotation name, such as a gene name. Metadata.json tags these paths as genes so that they may have special coloring and be sorted to the top of the display. All Paths in Pantograph have mouseover text that shows the Path positions, name, and type of the path under the cursor which allows researchers to quickly see the cDNA position and compare between individuals down to single nucleotide variances.

5.3.7 Binning Nucleotides for Scale

In order to be able to visualize the largest variation across a gigabase genome, some noise reduction is required. Binning makes data quantities viable for browsing, while zooming in is handled by decreasing the bin width for greater and greater detail. Given a sorted graph, the sequences of each Node are laid out in order into a Pangenome Sequence. This sequence is meaningless without the graph since alternative variants are listed in order, rather than taking a single consensus at each position. The pangenome sequence will be longer than any individual sequence since it expands to contain variants and private insertions.

From this pangenome sequence, odgi creates Bins of fixed width (e.g. 100bp) by binning together sets of bin width nucleotides at a time. Crucially, this means that multiple nodes end up in one bin and one node can be split across two bins. Bins allow us to describe generalization about a genomic region across all individuals. A bin can be present or absent in an individual, making large deletions visible. Bins have a coverage value and an inversion value that is a floating point, because they are the sum of many individual nucleotides that may be present, absent, or inverted. This can all be visualized in a rectangular grid for a quick overview of a large number of individuals across sequence space of any size (Figure 5.4). Notably, this approach differs slightly from gene presence/absence matrices like PanTetris (Hennig, Bernhardt, and Nieselt 2015) in that it does not require a gene annotation to function. With an annotation, the two methods become comparable. Binning allows us to zoom out in a consistent and performant way.

5.3.8 Haplotyping and Break Points

Additional scalability and insight can be gained by sorting similar individuals into groups called Haplotypes. Pantograph integrates a pre-existing tool called Haploblocker because its approach to SNP data already used a graph simplification algorithm which works on graph genomes with only minor modification (Pook et al. 2019). We first quantize the floating-point coverage values based on user defined threshold in order to create discrete features for analysis. The default thresholds are Absent: $c < 0.1$; Low: $0.1 <= c < 0.8$; Normal: $0.8 < c <= 1.2$; High: $c >1.2$. Rows are then clustered together based on sharing small features and a multi-step process attempts to stitch together larger and larger features into haplotype blocks. Absent areas are given less weight because an absence is less informative than shared sequence. Eventually the growing boundary of a haplotype block reaches a point where adding more sequence would reduce the local similarity score and the block stops growing.

Using coverage thresholds as the primary feature addresses a wide range of biological phenomena relevant to real haplotypes. SNPs, insertions and deletions will all cause drops in the coverage of pangenome loci where they occur. Clamping the coverage to thresholds allows some variation robustness while the user can still make the test more sensitive by adding additional levels if desired. Discrete features also save compute time.

For visualization, the more interesting challenge is finding the ideal order to sort the rows by haplotype. Different loci have different optimal row sort orders and break points make the constraints local and solvable. Haplotype rows lose coherence along the chromosome (Figure 5.12), meaning no single row sort order is globally optimal. This loss of coherence means, in a formal sense, that the set of individuals which share a correlated sequence similarity through inheritance at one locus are not the same set of individuals at a distant locus. The common cause for this mixed ancestry is sexual recombination when chromosomes crossover in eukaryotes, splicing together two ancestral chromosomes from different parents. These crossing over points are not randomly distributed.

Our solution was to introduce break points where the rows could be rearranged and a new sort order introduced. The position of these break points is determined based on maximizing *r*:

$$r = f_{break} * (1 - f_{continues}),$$

Where f_{break} is the frequency of individuals in a block ending at that locus and $f_{continues}$ is the frequency of individuals in a block spanning that locus. These are not redundant measures because haplotype blocks can overlap. A smoothing function is applied to *r* and then a target number of break points are chosen by iteratively picking the point with the highest score, ensuring they are separated by at least $\frac{1}{8 \times n_{break}}$ (Nadaraya 1964; Watson 1964). The end result is fairly distributed break points allowing the rows to be resorted along probable recombination hotspots to make haplotypes clearly visible in the data.

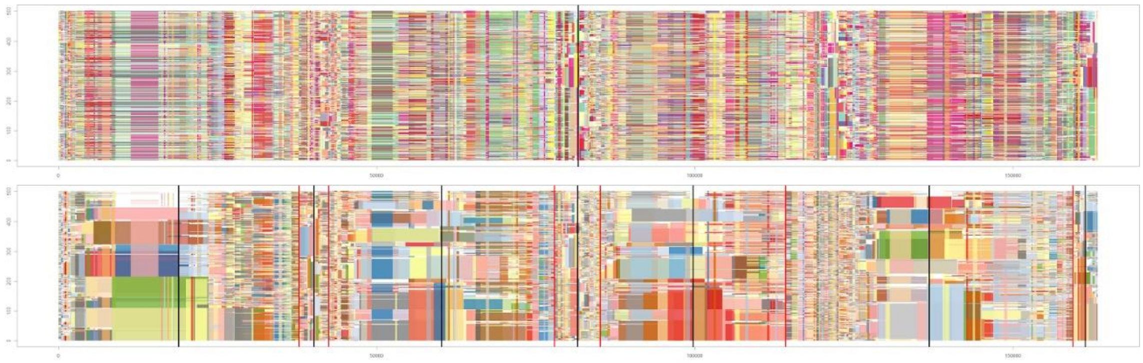
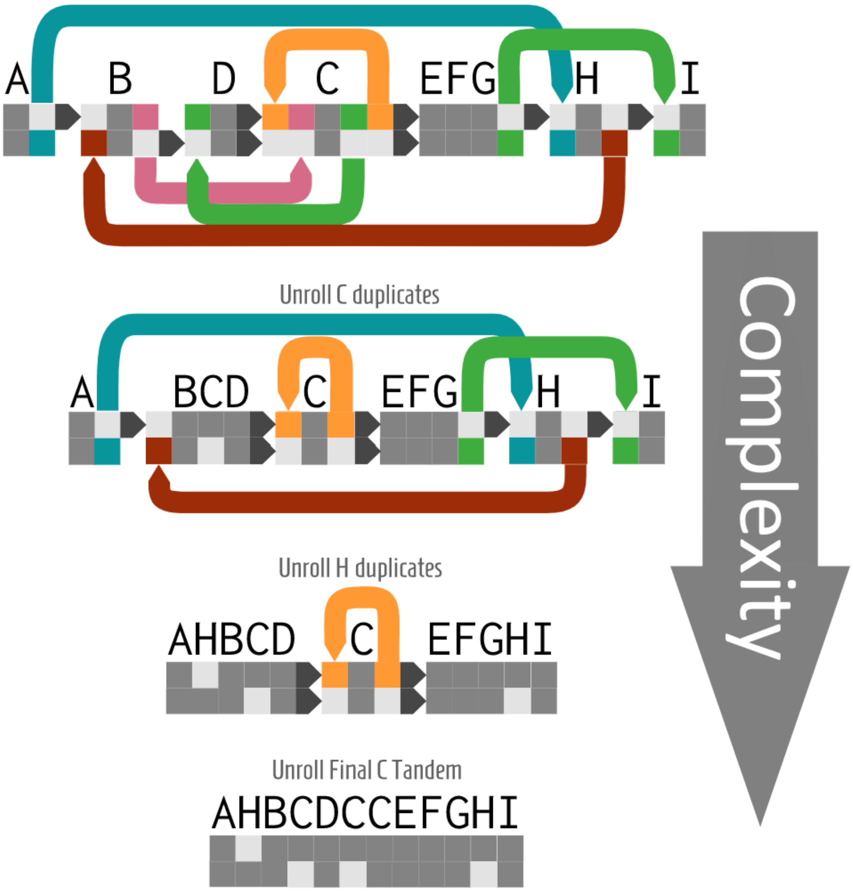


Figure 5.12: Row Sort Demonstrating Break Points: Top: Haploblocks of 501 Maize lines which were characterized using SNP data show very poor coherence when sorted by their middle coordinate. Bottom: Introducing seven break points where rows are reordered to maximize vertical neighbor matching haploblocks reveals distinct blocks for the majority of the dataset. Break points were developed for Pantograph row sorting to allow visual scaling to much larger datasets while still showing coherent patterns. **Image Source:** Torsten Pook for "Pantograph: A Scalable Method for Visualizing Diverse Pangenomes" (in preparation); co-author Josiah Seaman.

5.3.9 Unrolling Distributed Repeats

In software compilers, unrolling is an operation where an instruction is replaced by the output of that instruction. For example, "It's a small world" (repeat five times) would be replaced with "It's a small world, It's a small world, It's a small world, It's a small world, It's a small world". This is our main tool for decreasing topological complexity in the displayed graph. The main source of Link complexity originates from cases where a single component is inserted into many different places. We can see each of these pairs of Links as an instruction "insert X here" which we can replace with a copy of X, with appropriate coverage for each row.



```
A-BCDCCEFGHI
AHB-D-CEFG-I
```

Figure 5.13: Unrolling Distributed Repeats: The same end sequence can be simplified topologically by unrolling one component into copies at each site where the component occurs. The Links are eliminated, and the information migrated to the Matrix while the only information lost is that Copy #1 and Copy #2 are related. Notice at top H has coverage in both rows, however at the bottom image, the two copies of H each only have one cell filled. The top figure has seven Components, whereas the bottom only has one. The last step unrolls “C” into “CC” tandem repeat.

Unrolling is a type of zoom that reduces topological complexity (Figure 5.13). It is triggered by a threshold for the size of insert that justifies a Link. Any smaller insert will be unrolled. The size threshold for unrolling will be increased as the user zooms out, keeping the number of components on the screen at a relatively stable number even as the amount of sequence brought in increases by orders of magnitude. Since Links define the boundaries between Components, unrolling eliminates Links and joins Components. Large Components can be visually compressed to hundreds of megabases because they can be treated like an image and squished along the x-axis (this is called coverage binning).

Each level of unrolling is precalculated as a separate pangenome coordinate frame (5.3.11.2). The Path Position Server must keep a separate index for each unrolling level and annotations must be mapped onto each level separately (5.3.11.4). This change in pangenome coordinates will cause some jittering and warping as a user unrolls a graph. User affordances such as animations should make this transition as intuitive as possible.

5.3.10 X-scaling: Proportionate Representation of Variants

Rare variants are a major source of the Density Problem because a 10 bp private insertion in 1 individual out of 1,000 will create 10 * 999 = 9990 pixels of white space (Figure 5.5). Low density can be fixed with a dynamically scaled x-axis. The width of each sequence column is proportional to the percentage of the population which has that variant—i.e., the coverage of that column. X-scaling means that the sum of two exclusive columns carrying an allele at 50% frequency will be 50% + 50% = 100% the width of one conserved column. A 10 bp private insertion in 1,000 individuals will be 1/1,000th its expected width of 1/100th of one highly conserved column.

Fractional columns will require vector graphics. When zoomed in, one bin column is multiple screen pixels. Depending on zoom, some minor alleles could be less than one screen pixel. A minimum size will ensure that minor alleles are never completely hidden. However, this choice depends on the dataset and the chance of sequencing noise. The more common an allele is in the population, the more prominently it appears to the user. Regardless of their appearance, minor alleles will still affect membership in haplotypes which are visible to the user at megabase scale (5.3.8).

5.3.11 Pantograph Infrastructure

5.3.11.1 Data Representation

Internally, all coordinates are features are stored as lists of pangenome coordinates. Pangenome coordinates are particular to a ZoomLevel (5.3.11.3) since the pangenome will be different sizes at different levels. All data from the original odgi matrix is broken up into a series of smaller matrices and stored in Pantograph Components. Each Component has two lists of ordered Link Columns. Arrivals are incoming Paths that start traversal at the beginning of the Component. Departures are outgoing Paths that link to other components. Link coordinates are in pangenome coordinates, according to bin size, and in JSON these are stored twice for fast lookup: once as a departure, and once as an arrival. Link Columns are ordered starting closest to the Component matrix and new columns are appended on the outsides (see Supplemental 5 for more details).

5.3.11.2 Precalculation

Many of the performance improvements of Pantograph are based on storing precalculated layouts for fast rendering. This architecture was based on lessons learned from FluentDNA. We were able to scale up to real-time rendering of every nucleotide of gigabases of genome by precalculating the color averages of pixels laid out and stored in advance. In this tool, Component Segmentation and sorting are the main precalculation steps. Either RDF storage or JSON storage of Components are ready for Schematize to render on demand. Unlike FluentDNA, the stored information is not an image and so schematize still requires some computation to layout the vertical stacking of links and to calculate the exact XY coordinates of elements based on user settings.

There is a direct trade-off between the flexibility of user settings and the amount of storage space used by precalculations. If a user setting would require a separate copy of the data to be stored, then each possible setting is a multiplier for storage requirements. This situation can become untenable if there are multiple interacting user settings which create an exponential increase in the number of combinations of possible parameters which need to be stored. Currently the only example of this is the interaction between bin size and the level of unrolling. If there are 10 different bin sizes available then Pantograph must store 10 copies of the pangenome. If there are also 10 different levels of unrolling available at every bin size then 100 copies are now needed for storage. The practical solution is to tie the level of unrolling with the level of binning. There is no point in storing data which can never be understood in a visualization. Binning and unrolling are both types of zoom. We recommend only two unrolling levels, Simple and Complex, available at each bin size. With two unrolling levels, only 20 precalculation copies are necessary instead of 100.

A more advanced architecture has even better storage performance. If all fetches for precalculated data are formed as SPARQL queries to an RDF triple store (5.3.11.3), then calculations can be done only after they are needed the first time. New calculations can be stored in the database so similar queries in the future would be much faster. This optimization is ideal for datasets which have a few hotspots of interest and would otherwise require large amounts of storage. Currently, this on demand computation would require Component Segmentation to be ported and integrated into odgi (see 5.3.11.1).

5.3.11.3 Semantic Variation Graphs using RDF

Pantograph uses an RDF triple store (see below) for storing the graph genome along with all metadata, and precalculated analysis. RDF is a semantic web technology which makes it easily compatible with a host of other features. Key among those is “A System to Link Knowledge Graphs and Genome Graphs” proposed in Moustafa et al (undated preprint):

> “Knowledge graphs can be represented as directed graphs where nodes represent any kind of assertions (accession number, date of collection, isolation source, etc.). Different types of edges can express different relationships between those assertions. “Genome Edges” represent genomes and include all the assertions that are true for that genome. “Categorical Edges” allow grouping of nodes by category, for example grouping all accession numbers in single edge. “Relationship Edges” describe known relationships between assertions, for example if a set of SNPs is correlated to a disease. Finally “Query Edges” are constructed on the fly and represent a subset of assertions that the user is interested in, for example “all the isolation dates this year”.
>
> One can therefore construct a genome graph using the genomes represented by edges that intersect any number of categorical, relationship, and/or query edges.”

RDF is a general purpose format/protocol that consists of triples: subject, relation and object. This is used in life sciences to store various kinds of annotations. All data necessary to run Pantograph can be stored in an RDF triple store. Both odgi and Component Segmentation output RDF triples in Turtle output (TTL) using the pangenome ontology defined by the Pantograph team (Figure 5.14). A JSON format was developed as a temporary measure for transferring data between Odgi, Component Segmentation, and Schematize. However, to make our data accessible to the wider community we plan to exclusively use RDF for future development.

One concern to address was estimating the scalability of storage and speed of RDF for gigabase genomes. RDF is already used in the Swiss Institute of Bioinformatics for large datasets. We compared the storage size of a human chromosome variation graph in odgi with the same data exported in spodgi and found the RDF was roughly twice as large (data not shown). In most target applications, doubling the storage requirement was not prohibitive. Similarly, storing the entire precomputed ZoomLevel pyramid (Figure 5.14) required four times the space of the original compact variation graph in odgi. These storage requirements are geometrically identical to FluentDNA's DeepZoom stack (OpenSeadragon n.d.) with the same trade-off: more storage requirements for faster retrieval time and performance.

The standardization of pangenome data format is necessary for interoperability. A query syntax similar to SQL called SPARQL allows federated queries from all RDF triple stores using the same ontologies. Pantograph output can be stored on the SPARQL endpoint, which can be queried from Schematize. This means we do not have to publicly serve a pre-built JSON format. Instead, other services can access the same inputs that Schematize uses through SPARQL queries. Our ultimate goal is enabling a federalized query for retrieving subgraphs and annotations from one query, which means that we can completely relate genome variation graphs (vg ontology github.com/vgteam/vg/tree/master/ontology) and annotation knowledge graphs (geneontology.org). Bridging these two worlds will likely require a further extension of the vg ontology with relations such as "has_annotation". The complete vg ontology including Pantograph extensions is available at http://biohackathon.org/resource/vg. Pantograph was specifically built using this version: https://github.com/vgteam/vg/commit/606acd28fe29f98b06a3f829622c06b55894e35a.

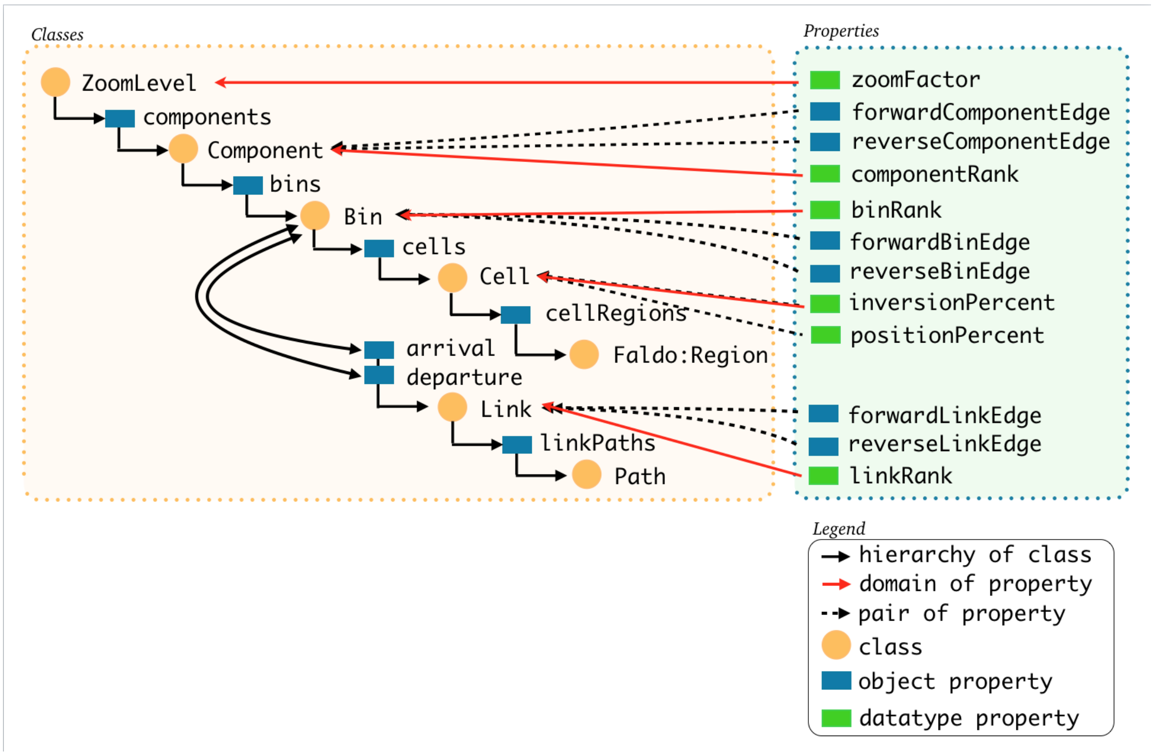


Figure 5.14: Pangenome Ontology: This diagram shows the hierarchical nature of the semantic graph genome ontology which can be found at http://biohackathon.org/resource/vg (Supplemental 5). Each ZoomLevel contains a complete copy of the pangenome at that level of detail with its own set of bins containing inversion and position information. Paths are traversed via forward and reverse link edges connecting Components. Every cell has Faldo region coordinates which map back to universal standard nucleotide coordinates for reference genomes found in the Path Position Server (5.3.11.4). **Image Source:** Yokoyama et al. (2020), used with permission.

5.3.11.4 Path Position Server

Pantograph uses a single set of pangenome coordinates which are global but do not correspond to any assembly. The reference genome and all other genome annotations use their own coordinates which disagree. The Odgi Path Position Server transforms all of these coordinates to pangenome coordinates, allowing complete navigation and comparison. Queries contain the path name and the position along the path, which starts from one rather than zero. The server references a file containing the “succinct variation graph index” for the closest node in the path and adds the remainder. For more information about this file format see Garrison et al. (2018). odgi commands for generating and using the server are documented at https://pangenome.github.io/odgi/odgi_docs.html.

5.4 Results

5.4.1 Pantograph Scalability Properties

All software designs were evaluated based on the scalability criteria laid out in Methods (5.2). Table 5.1 contains the Big-O performance equation while the legend details a rationale for why each tool received the score it did per column. Overall, we find that Pantograph scales very favorably in terms of number of elements rendered compared to other tools. It achieves this through additional precalculated layout work. odgi bin also scales to a large number of individuals in tests, however it generates static images whereas Pantograph is a fully feature interactive genome browser.

Table 5.1 Visualization Scalability Metrics

	Elements ⇓	Layout Time ⇓	Demo Individuals ⇑
Bandage	NP	N^2	1
Sequence Tubemap	NP + Rnp	1	12
Odgi Bin view	R	N	1000
Pantograph	log(S)log(R)	N + R + P	167

Each metric is scored based on an equation given N nodes, P paths, S SNPs, and R structural rearrangements. S and R increase as P increases, since there are more total variants in the population. Visualizations that allow shared variation use log(S) to denote the probability that variation is shared between multiple individuals.

Elements: Bandage renders one segment for each Node or contiguous set of Nodes without branching. Since Bandage is not currently applied to variation graphs, it is not clear how this will scale in the future with multiple Paths sharing Nodes. Sequence Tubemap draws a new vector entity for every Node/Path pair (NP). Every rearrangement causes the retraversal of every Node/Path pair between the start and end of the rearrangement (np). While this number is not as large as NP, Sequence Tubemap contains no sorting to minimize the size of np. Odgi Bin uses a square matrix and so does not actually render more elements to show an increasing number of SNPs. Moreover, binning actually hides SNPs. However, each Rearrangement requires an arc drawn and there's little effort to merge arcs shared between individuals. Pantograph performs somewhat better by merging nearby Rearrangements, but has the added complexity of rendering separate Components proportionate to the number of Links or log(R).

Layout Time: Bandage's force directed simulations require tensions between all nearby node pairs (N^2). Sequence Tubemap makes no real attempt to sort the pangenome beyond very quickly flattening the Node list. Odgi Bin and Pantograph use the same sort, which is run for a time linear to the number of Nodes in the graph (N). Pantograph also requires Segmentation which has loops for Nodes, Paths, and Rearrangements, though never multiplicative.

Demo Individuals: Bandage's original publication only lists examples with one genome (Wick et al. 2015). Sequence Tubemap examples online don't include enough structural variation to be comparable. Figure 5.3 contains 12 HLA Paths. odgi was used on the 1000 Human Genomes Project dataset in a tool comparison in Eizenga et al. (2020) though not specifically for visualization. Pantograph release 1.0 on graphgenome.org featured 167 SARS-CoV-2 individuals with an admittedly small amount of structural variation.

5.4.2 Haplotyping SARS Pangenome Overlap

SARS-CoV-2 was our initial target application for public release of Pantograph 1.0 in July 2020. However, as a recent viral subpopulation it does not yet have many structural rearrangements. A more interesting view is the haplotype view which allows us to compare across potential sources of SARS-CoV-2 (Figure 5.15). SARS1, SARS-bat HKU and SARS-CoV-2 share the majority of their genome content in terms of pangenome locus. However, at a bin size of 1w, only 5% of haplotype blocks are shared between SARS1 and SARS-CoV-2. To investigate this further, we modified the Haploblocker algorithm to only create blocks shared by the two SARS families and only 14.16% of the pangenome was covered, indicating substantial differences in the variants between the two viral lines. The reason for this difference is biologically apparent, since each population comes from a separate recent radiation after a presumed genetic bottleneck that would have fixed a set of variants.

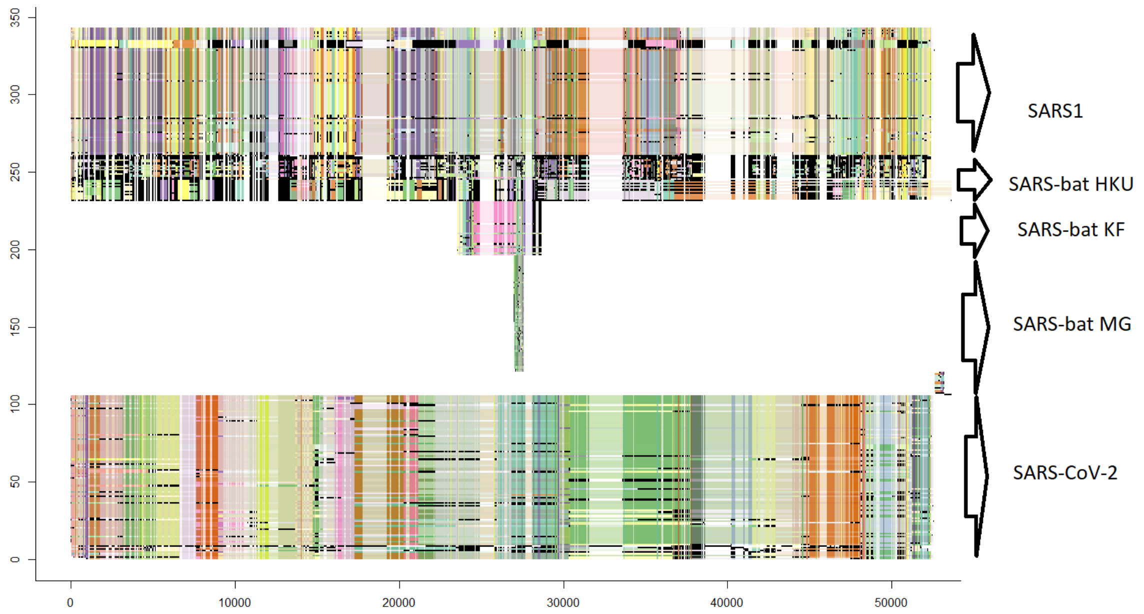


Figure 5.15: Aligned Graph Genome Haplotypes for SARS lines: This Graph Genome was constructed from sequences of 343 SARS individuals from SARS1, SARS-CoV-2 and bat variants (SARS-bat HKU / KF / MG). Colored blocks are haplotypes which may be overlapping. Lighter colors indicate no sequence present in the individual at that pangenome locus. Black indicates sequence is present but not assigned to a haplotype, particularly common in the lines where we have fewer individuals sampled. Each SARS line appears to have just one major haplotype, unlike typical populations with multiple haplotypes present. There are exceptions for example at position 50,000 nearly half of the SARS-CoV-2 individuals have a minor haplotype. From the large degree of overlap in SARS-bat HKU we can infer that there are bats carrying a very similar virus and from KF and MG we can infer that some small portions of sequence are likewise shared with other viruses that presumably have their own unique sequences not shown in this graph. Ideally, we would have been able to observe a SARS bat sample with overlapping haplotypes with SARS-CoV-2. An in-depth analysis of the topic is beyond the scope of this thesis. **Image Source:** Torsten Pook for “Pantograph: A Scalable Method for Visualizing Diverse Pangenomes” (in preparation, co-authored by Josiah Seaman).

5.4.3 *Arabidopsis thaliana* Pangenome

To truly test Pantograph on a complex eukaryote genome, we visualized twelve *Arabidopsis thaliana* individuals from the pilot of the 1001 Genomes Project. Unlike SARS-CoV-2, the majority of visual columns of the pangenome schematic are Link Columns showing complex topology rather than Matrix cells with genome content. This ratio makes optimizing the sort a top priority to simplify the visualization (5.3.4). Manual inspection of problem areas shows that bins contain a wide array of discontinuous ranges. Components often have Link Columns pointing to themselves to facilitate internal rearrangements. Similarly, the majority of Links point to a Component that immediately points back to the current Component. Unrolling (5.3.9) is designed to address these problems but is not yet implemented. *Arabidopsis thaliana* results indicate implementing unrolling and addressing self-referencing Components will be mission critical to keep the number of Component manageable.

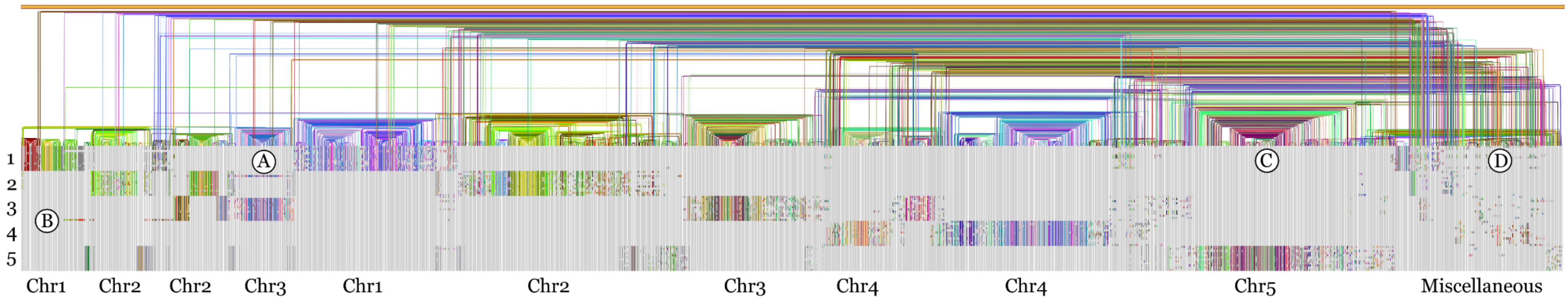
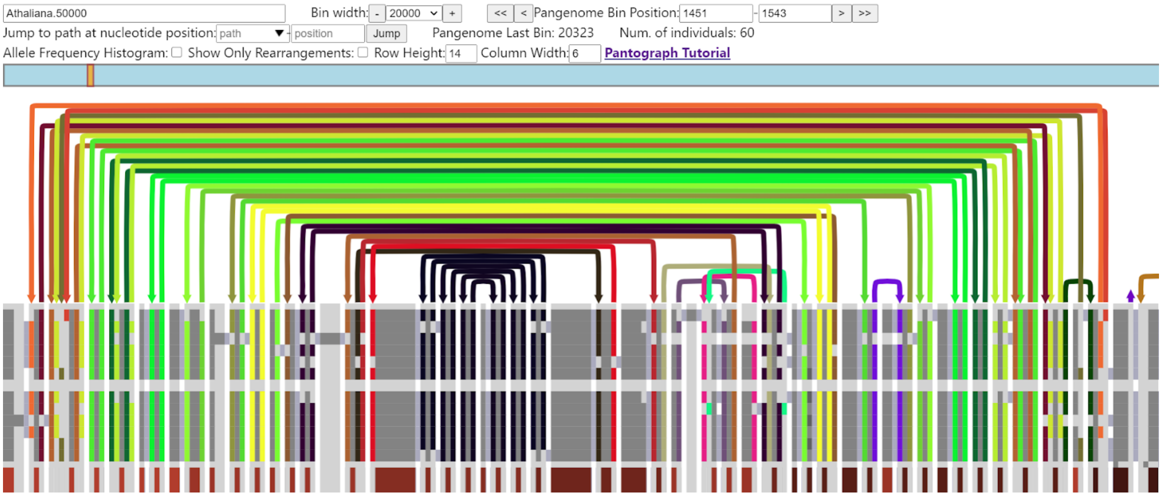


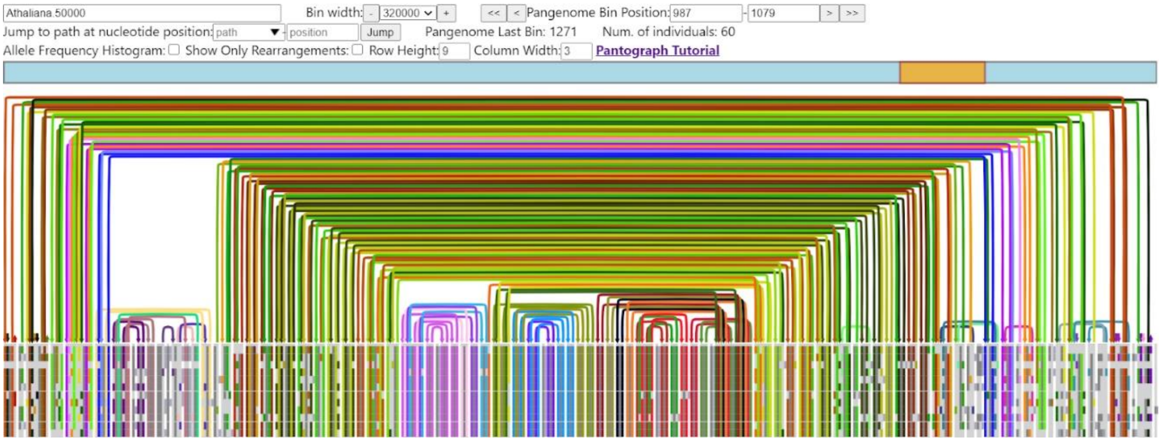
Figure 5.16: Complete *Arabidopsis thaliana* pangenome with 12 individuals visualized in Pantograph. This image was generated by Pantograph on a 3875 px wide monitor at Bin size 640,000 bp. Frames take 3 seconds to render on a computer with 3Ghz and 16GB of RAM, making interactive navigation slow but possible across the whole pangenome in a browser. The majority of the pangenome can be seen covered by large chiasmi, at least two per chromosome. Chromosomes are sets of rows, numbered on the left. Rearrangements between chromosomes will appear as exceptions to the block structure separating chromosomes on the x-axis. A) For example, AT7328_chr2 appears out of place, covering loci associated with chromosome 3. This could indicate a structural rearrangement in AT7328 or an assembly error. B) Similarly, AT6906_Chr3 is visible inverted in red in loci normally associated with chromosome 1. C) Figure 5.18 can be seen on chromosome 5 on the right side. Except for this structure, the rest of chromosome 5 is scattered throughout the bottom row of the pangenome with apparently few links connecting them. D) On the far right, the sorting has placed Nodes which were difficult to cluster by chromosome. These are centromeric repeats that maps to multiple chromosomes as well as shared telomeric sequences. Links from this unsorted miscellaneous section stretch across the top of the figure to link with the ends of chromosomes. **Source:** This dataset can be browsed interactively at http://arabidopsis.graphgenome.org

The most notable structural feature starting at the beginning of chromosome 1 and covering the majority of all chromosomes are a set of symmetrical concentric nested Components linked in pairs such that they alternate left and right around a focal point Component (Figure 5.17). These symmetrical structures can be found nested inside one another following a tree structure (Figure 5.18). For brevity, I refer to this symmetrical structure as a chiasmus (plural chiasmi), a literary term meaning a motif whose beginning is repeated at the end in inverted order. I used Pantograph to determine what was causing the chiasmi.



Athaliana.50000 Bin width: 20000 Pangenome Bin Position: 1451 - 1543
Jump to path at nucleotide position: path - position Pangenome Last Bin: 20323 Num. of individuals: 60
Allele Frequency Histogram: □ Show Only Rearrangements: □ Row Height: 14 Column Width: 6 Pantograph Tutorial

Figure 5.17: Example Chiasmus on Chromosome 5 of A. thaliana: The chiasmus pictured here is a symmetrical pattern of interlinked Components centered around a focal point. In the center, a smaller section with entirely black colored links makes it easier to see the alternating left-right link direction characteristic of a chiasmus. Each grey column of Matrix contains 20,000bp of sequence, the Bin width at top, and contains multiple genes. To check that this chiasmus was not simply an artifact of the sorting process, I have edited in a visualization of the last chromosome position for each Bin along the bottom row in red. Light red is the 5' end progressing to dark red at the 3' end, color scale is relative to the view window. This smooth spectrum shows that the sort is a real feature along the chromosome. It even highlights the exceptions to the spectrum (purple and forest green links) also do not match the larger structure.



Athaliana.50000 Bin width: 320000 Pangenome Bin Position: 987 - 1079
Jump to path at nucleotide position: path - position Pangenome Last Bin: 1271 Num. of individuals: 60
Allele Frequency Histogram: □ Show Only Rearrangements: □ Row Height: 9 Column Width: 3 Pantograph Tutorial

Figure 5.18: Nested Chiasmus: This nested symmetrical structure covers over 30 megabases of pangenome sequence. The chiasmus can be represented by the parenthesis sequence in Vienna notation: (x(()()())) where x is the region with no long range Links. This entire figure can be seen as a small feature in Figure 5.16B.

A few key observations of chiasmi in the *Arabidopsis thaliana* pangenome are as follows. They cover nearly the entire pangenome with one or two major chiasmi per chromosome arm (Figure 5.18). They do not appear to span centromeres, telomeres, or interconnect multiple chromosomes. For example, if these were caused by sequence similarities by homeologs on different chromsosomes, we would expect to see Links spanning telomere boundaries to interconnect multiple chromosomes together. The same is true of real biological rearrangements that would place sequence in different individuals on different chromosomes as well as technical artifacts arising from interchromosomal alignments. Instead, they appear to be localized to one chromosome arm.

They can be nested up to the maximum resolution of the current visualization of 10,000 bp (Figure 5.17). Bins near the middle of a chiasmus have more discontinuous ranges than Bins which are not part of a chiasmus. These observations led to a hypothesis of the chiasmus discussed in the next section.

The results from the 1001 Genomes pilot project made it clear that addressing the accumulation of Links was necessary for the scalability of Pantograph to thousands of megabase genomes. Binning did not fully address structural scalability because large Bins had more discontinuous ranges with a higher probability of a Link extending outside the Bin. Self-loops, where Components have a Link to themselves need to be suppressed to keep the number of Components down. The undesirable visual complexity of Figure 5.16 is due in large part to unnecessary segmentation from internal reordering of Bins. The Links are being added to facilitate the ordering of elements that are invisible to the user. Given the high number of ranges inside a chiasmus, this reordering behavior iteratively subdivides them into individual Bins. From this new information, Component Segmentation Issue #50 (https://github.com/graph-genome/component_segmentation/issues/50) was created to address Component self-loops and implementing unrolling will fix repeat content causing complications.

5.5 Discussion

5.5.1 *Arabidopsis thaliana* Pangenome Symmetry

The visualization of *Arabidopsis thaliana* pangenome demonstrated it was practical to browse a eukaryote Graph Genome in a browser and use it to diagnose assembly challenges. I uncovered several expected chromosome rearrangements (Figure 5.16 A and C) but also a mystery in the form of ubiquitous symmetrical nested structures called chiasmi. From all the observations, I would hypothesize that chiasmi are an emergent phenomenon from the combination of sorting algorithm, visualization, and the underlying chromosome structure. Chiasmi do not appear everywhere or in all datasets using the same settings (e.g. SARS-CoV-2), so they are likely symptomatic of large similar sequences which can be aligned but which still pass through the smaller repeat filters (10 bp).

There are two classes of repeated sequences with an intertwined history: transposons and ancient whole genome duplicates (1.2.2.6). Transposons are often found concentrated to one specific chromosome arm and chiasmi also do not cross chromosome boundaries. Transposons are likely to be the primary cause of many discontinuous ranges in the same Bin. Between the base angiosperm ancestor and *Arabidopsis thaliana* there have been three rounds of WGD which have left extensive evidence of homeologs in *A. thaliana's* small diploidized genome (Figure 5.19). Based on a dot plot of the TAIR10, 87.5% of the CDS has at least one extra copy present in the genome (Lyons and Freeling 2008; Haug-Baltzell et al. 2017). It is possible a chiasmus is an artefact of the sorting algorithm struggling to reconcile nodes shared because of matching sequences left over from an ancient WGD. It is also possible it is a degenerate case of sorting that is only apparent given some types of input data.

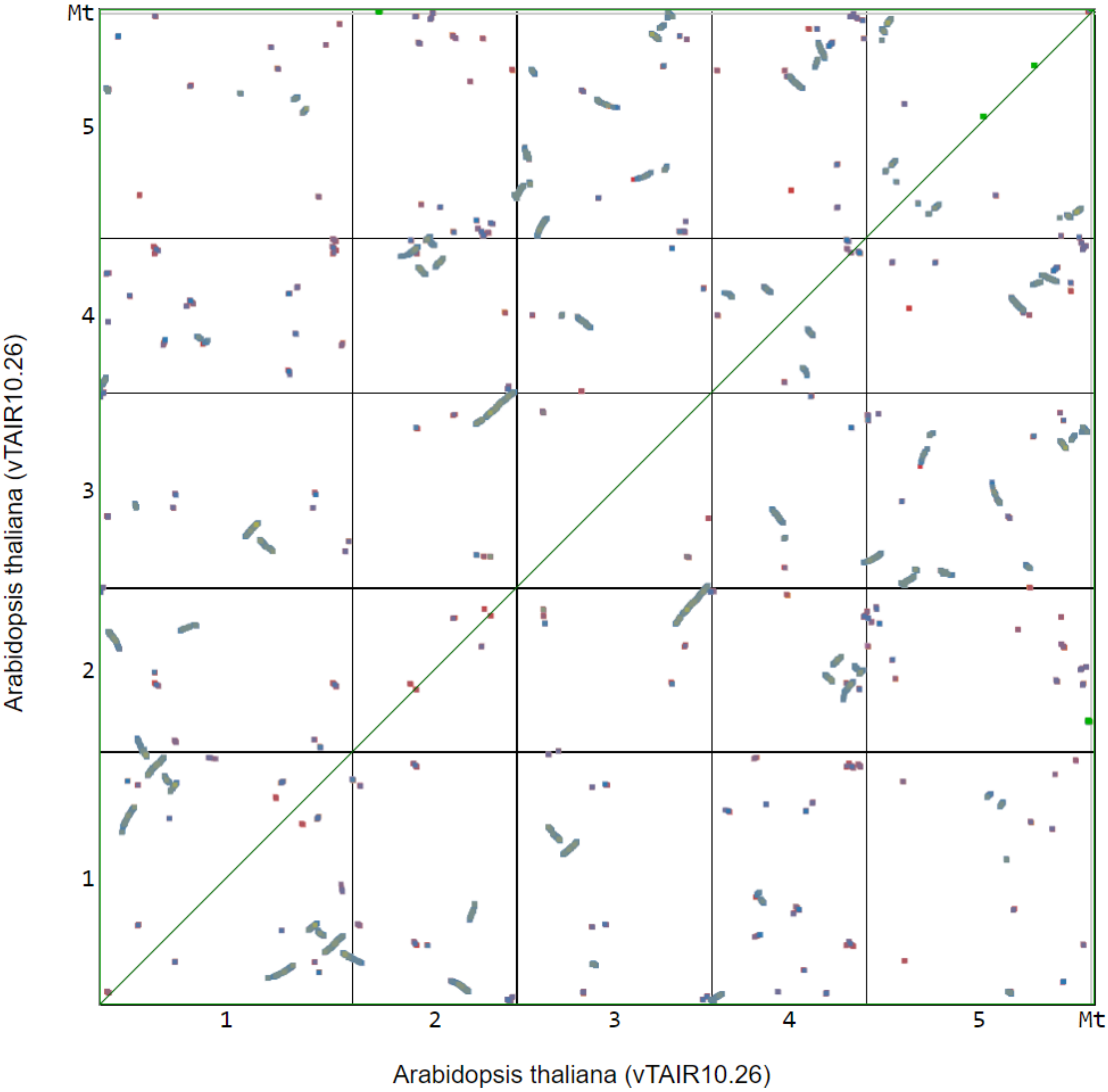


Figure 5.19: Dot Plot showing *Arabidopsis thaliana* homeologs: The x and y axis both represent position along the TAIR10 *A. thaliana* genome as rendered in this CoGe dot plot (Lyons and Freeling 2008). Dots indicate a sequence match between distant regions in the genome, filtered for superfluous repeats, these are indicators of homeologs left from ancient WGD. Color indicates level of sequence identity: orange is more ancient, blue and green are more recent. To determine what percentage of the exome is homeologs, one can visually extend columns up and down from every point on this dot plot. From this method I obtain an estimate that *87.5%* of the exome has at least one similar sequence match . This dot plot can be explored interactively at https://genomevolution.org/r/rz0o.

5.5.2 Genome Browsers in the Age of Graphs

Most users interact with genome visualizations in the context of genome browsers like UCSC (Rosenbloom et al. 2013), JBrowse (Buels et al. 2016), or Ensembl (Zerbino et al. 2018). Visualization is the adapter layer between the realm of human cognitive capabilities and the realm of computational resources. Databases, data mining, and AI are all locked within silicon and lack the ability to progress without human understanding. Clear visualizations are needed to facilitate human reasoning about the complex interactions between structural variants in the data and research questions. This bridge brings together the best of both worlds. Insight without evidence is impotent. Data without reasoning is useless.

Big Data has been a major focus of 2010-2020, but significantly less has been said about scalable Big Data Visualization. Visualizations must continue to scale in order for new genomic data to be useful. The switch from a single genome, to a genome alignment graph is not simply a change in quantity, but a change in kind which requires a completely new toolset. Graph genomes allow researchers to access new types of features that would have been removed or invisible in previous sequencing techniques.

Pantograph's major contribution to the field is the ability to treat structural rearrangements as a single point feature that can be shared by many individuals in the same manner as a SNP can be shared and be assigned an allele frequency. Component Segmentation makes this ability far more powerful by identifying that many similar rearrangements in different individuals could originate from a common ancestor. Clustering by shared ancestry using haplotypes is also what enables the y-axis to scale to rendering possibly thousands of rows if each individual no longer needs their own visible row.

The second key lesson of Pantograph scalability is "acceptable losses" when zooming out to whole chromosomes. Binning based on Node sort order is not mathematically perfect, but it creates a comprehensible representation for large scale features at the cost of small scale features being lost in the averages. Unrolling discards some Links, but maintains the largest, most informative rearrangements. This is designed to mimic human vision of distant objects, which start small and blurry but become gradually more detailed as they move closer. There are two ways Pantograph implements this. First, through physical size controlled by Row Height and Column Width as well as browser zoom. Second, through semantic zooming implemented by Bin Size (5.3.7) and Unrolling (5.3.9) which actually reduces the level of detail presented to the user. A key aspect of the user interface is that it clearly labels what level of detail is visible, e.g. 10bp bin size, so users know features below that size will be indistinguishable.

The third lesson is the difficulty of picking a performant online rendering platform. For Pantograph's 2020 implementation, the team chose ReactJS with the intention of integrating with JBrowse2 in order to support the largest possible array of annotation formats (Cain and Buels 2020). Once we began collaborating with covid19.genenetwork.org and began loading pangenome with over 300 individuals, Pantograph became prohibitively slow. ReactJS, while faster than HTML, still required optimizations for the number of elements on screen, requiring significant extra development time. The contrast became most apparent at ISMB when Pantograph was presented alongside GenomeSpy, which had similar features for SNPs but was implemented in WebGL (Lavikka et al. 2020). GenomeSpy has no buffers or chunking, it simply loads the entire dataset as a single file over the Internet and renders all elements at once with smooth real-time animations. By comparison, Pantograph was much slower. The difference is due to taking advantage of the hardware acceleration of graphics cards which have benefitted from 40 years of R&D in the video game industry. A similar change in the Matrix rendering component of Pantograph could change its implemented rendering speed by 100-1,000x. GenomeSpy also uses the Vega visualization notation framework for streaming updates which may contribute to its responsiveness (Satyanarayan et al. 2016).

In general, genome graphs are not planar (Weisstein 2021); hence any 2D rendering will have overlapping edges whereas in 3D, edges will not overlap. Consider Point A inside a square perimeter of edges which needs to connect to Point B outside the perimeter. In 2D, there is no path where the edge does not cross the perimeter, whereas in 3D it can be connected to infinitely many points above or below the perimeter. Pantograph's design makes a conscious attempt to keep the 3D nature of a graph contained in rectangular grids on a 2D surface. This is strictly a usability consideration because working with a 3D object on a 2D monitor is so inefficient.

This will continue to be the case as long as the dominant mode of interaction is using monitors and mice, rather than augmented or virtual reality. Data visualization and user interface expert, Bret Victor has spent his career designing what an ideal interface might look like, called Seeing Spaces (Victor 2014). Ideally, we would construct a building out of our knowledge that one could walk through, explore, and learn. Memory experiments have found that storing facts in a mental memory palace is best for retaining and retrieving memories at later dates (Roediger 1980). This is likely because human brains are adapted to work in real 3D spaces with objects, containers, and journeys. The abstract land of computer interfaces is a poor match and foreign to our primate brains. For now, we continue to be constrained by the tools at hand.

5.5.3 Benefits of Graph Genomes as a Storage Medium

I predict that Graph Genomes will gradually become the data type underlying all genome data stores. The reason is the unavoidable reality that data can always be put through a lossy transformation for convenient consumption, but once data is discarded it can never be recovered. This reality has led to the archiving of short read data; a practice which requires enormous amounts of storage space and limits the applicability of sequencing technology as a field if it cannot scale (Pavlichin and Weissman 2018). The best solutions to this problem require using short read data as a temporary intermediary to an alignment with complete markup for all variants in the population such as (Kelleher et al. 2019) which is capable of storing all of humanity's SNPs in less than an extra gigabyte.

Graph Genomes have the capability to store any kind of genomic variation. When stored in RDF, the same database can contain any other annotation and metadata built directly into the pangenome. This degree of flexibility means that short reads can be stored as a series of small updates to a reference Graph Genome. Once all possible metadata is wrung from the short reads, they can be discarded to save space for the next batch of individuals to be sequenced. This trade-off becomes more practical as sequencing costs drop at a faster rate than compute and storage costs.

Any desired FASTA flat file can be served from a Graph Genome. One genome FASTA is a single set of paths from a pangenome. This paper has focused on variant graphs but Graph Genomes can also store an assembly graph of one individual. A genome assembly is a destructive flattening and linearization of an assembly graph. It is prudent to keep the original assembly graph so researchers can update it with new sequencing or use different assembly parameters to acquire a linear genome. Repeat annotations (which often account for more than 50% of the genome) are also driven by alignments which can be preserved as edges in the graph.

5.5.3.1 Science Enabled by Semantic Graph Genomes

Taken as a whole, we can envision a near future where Graph Genomes enable a much better-connected bioinformatics. Biologists will be free to sequence specimens with abandon without concerns for running out of storage space. All new information is stored as updates to the species pangenome and then raw data is discarded. Bioinformatics is currently the largest cost factor in sequencing and automation is the only cure. Human reasoning still stays in the loop using interactive visualizations and browsers such as Pantograph to highlight new features added by the latest rounds of sequencing.

Sequence classification can be separated using a decision tree composed of convolutional neural networks trained by user picked sequence examples (Frosst and Hinton 2017). In a more advanced future application, the topology of the decision tree could be learned through observing a user's actions in a pipeline environment like Galaxy (Giardine et al. 2005) using Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al. 2016). The filtering and action flowchart is then amplified by AI which mimics the researcher's actions across all similar regions of the pangenome. The AI then returns results of interest based on a scoring metric learned from watching researchers use Pantograph and tag areas of interest to them. Any new knowledge gained can be published in papers which reference the specific nodes in RDF for reproducibility. This type of collaborative human machine environment is an ongoing area of research in DARPA (Draper 2021).

More radically, micro-publications in the form of metadata added directly to the pangenome RDF can communicate new knowledge in a way that appears in genome browsers and is machine readable. These micro-publications can be federated through SPARQL and hosted on journal or university servers. Pantograph focuses on the scalability of a visualization technique for thousands of individual specimens, but there's a more fundamental scalability problem: the scalability of the human mind. Every day over 20,000 biology publications are published and after May 2020 there were an average of 282 COVID-19 publications every day (Sarkar et al. 2020; Pacchioni 2018). No human could possibly keep up with this rate of discovery and so we compensate through subspecialization and hopefully filtering by quality. However, research has shown that replicability (the most objective metric of quality) has no correlation with how often a paper is cited and published in high impact journals (Y. Yang, Youyou, and Uzzi 2020; Camerer et al. 2018). I propose machine readable micro-publications as a more realistic solution to the larger scalability challenge of comparative genomics as a whole (Clark et al. 2014; Raciti et al. 2018).

5.5.4 Cross Species Graph Genomes

This thesis started with the goal of comparing similar Ash tree species to understand how large-scale changes in genome content occur after a whole genome duplication. Pantograph's presence/absence matrix appears to be the ideal tool to visualize the repeatability of gene loss between 27 *Fraxinus* species (Chapter 4). So why is it not present in the Results section? Pantograph is primarily a browser for existing datasets. Graph construction is a separate and challenging topic which bears some discussion here. In order for the *Fraxinus* pangenome to be visualized in Pantograph, we must first construct a cross species Graph Genome. Constructing a *Fraxinus* genus pangenome graph could be another entire thesis topic, as explained below.

5.5.4.1 Graph Construction Challenges

All of the applications discussed so far have been restricted to within species pangenomes because additional technical challenges arise when comparing genomes across species. For Graph Genome approaches to become the default solution for comparative genomics, they must be applied to a diverse group of species, for example the 200 mammal alignments currently under development (Birren and Karlsson 2021). As of 2020, there was no turnkey solution for cross-species graph construction of gigabase-size genomes.

Most aligners still use a bottom up approach which matches small kmers, identifies syntenic kmers, and builds larger blocks from there. Examples of graph construction programs are vg, REVEAL graph, and seqwish (Novak, Garrison, and Paten 2017). A problem is introduced when SNPs are approximately as common as the base kmer size. The fundamental building blocks for alignments become disrupted and the graph dissolves into an array of variant nodes with few shared connections between species, even when overall sequence identity is high. Programs like BLAST solved this problem decades ago by allowing edits for short sequences, but those solutions aren't necessarily performant on gigabases of sequence. It is likely just a matter of getting the right team with time and funding to implement a complete solution to this engineering problem (Brown et al. 2020; Pritt et al. 2018).

5.5.4.2 Within vs. Across Species

When constructing a graph containing multiple species, do we include one reference individual from each species or a diverse panel of individuals? Likely both approaches will be used because of their advantages and disadvantages. Using only one reference per species is the simplest approach and reduces the chances of information overload by ensuring that every variant encountered is a difference between two species. However, this could be misleading.

There exist trans-species polymorphisms; which is variation present both within one species and also between species. For example, ABO blood types are preserved in primates (Ségurel et al. 2012). This can be created by incomplete lineage sorting and maintained by balancing selection. Trans-species polymorphisms will only be visible in a Graph Genome containing multiple individuals from each species. Within a few years, we should be able to detect and catalog large numbers of a new category of variants using Graph Genomes: trans-species structural variants.

Without Graph Genomes, it is possible to study trans-species structural variants. The base sequences a researcher would use to identify the trans-species polymorphisms are present in the per species genome assemblies. Existing genomes can be aligned to create LiftOvers which can be mined for these polymorphisms. Literature can be searched for already identified instances. However, the metadata tagging a trans-species polymorphisms in a machine readable format is not necessarily available. Currently examples of trans-species structural variants are few and far between for example, the FLCD variant is restricted to within *Drosophila* species (Yassin et al. 2016). Convergent fractionation reported in 4.3.3.1 could be considered a large scale trans-species structural variant, whereas non-convergent fractionation is not.

5.5.4.3 Steps for a *Fraxinus* Graph Genome

Building a Graph Genome of the 27 *Fraxinus* species discussed in Chapters 2 and 4 would certainly be thwarted by the technical limitations discussed above (5.5.4.1). Until tools such as seqwish, REVEAL graph, or vg have received further development this level of species diversity in a single graph is simply not practical. While all of *Fraxinus* is technically within the same genus, Section 3.3.6 established that *F. pennsylvanica* and *F. excelsior* are too diverged to use for scaffolding, which reflects on their number of relative rearrangements. *Fraxinus* has been diverging over at least 34 million years (3.3.3) and plants can tolerate a higher degree of genome change than mammals (Kejnovsky, Leitch, and Leitch 2009).

To be mathematically precise, excluding indels, the number of SNPs in aligned regions is approximately 1 in 10 between *F. pennsylvanica* and *F. excelsior*. Given an assembly size of 961 megabases (Table 3.2) a kmer of 15 nucleotides will only occur by chance once in the whole genome (4^{15}=1,073,741,824), making shared kmers between the two genomes evidence of homology. However, a kmer of 15 nucleotides is longer than the average distance between SNPs, meaning the majority of kmers will be disrupted before they can seed a node in the graph. A more robust graph construction method is required in these difficult cases.

5.5.4.4 Pantograph Cross Species Scalability

Finally, we should consider whether the Pantograph visualization design will scale across species and in applications with high divergence. It is difficult to be completely certain without a dataset to test against, but specific aspects can be tested. The matrix scales very well with a high number of SNPs still maintaining a rectangular coordinate frame. Rare variants can be hidden using x-scaling (5.3.10) without limit. An increasing number of rearrangements is more problematic if rearrangements are not shared between individuals as would be likely across species. In this case, Link Columns suffer from the Density Problem (5.1.2) while an increasing amount of vertical space is taken up by their ribbon stacking as we've seen in Sequence Tubemap (Supplemental 3).

Binning is disrupted by Cmponent boundaries, which can be prevented by a good sort. It is an open research question whether or not trans-species pangenomes have a well-ordered sort. The fact that chromosome painting is possible at all is positive evidence for a feasible global sort (Serov et al. 2005; Kemkemer et al. 2006). Unrolling would mainly be useful for simplifying the chromosome level views across species by removing information about small similarities that cannot be reconciled in the sort. These similarities would still be browsable locally when zoomed in. Other design issues will have to be addressed when test datasets become available. Current development is being carried out using synthetic genomes. Further design work is needed in order to hide irrelevant data to scale Pantograph to cross species comparative genomics.

5.6 Conclusions

Researchers will face significant scalability challenges in the years ahead in the transition from flat files of a reference individual to a network of annotated population variation in a species. Graph Genomes are the correct data structure to represent the full range of genetic variation possible. Previous tools for visualizing Graph Genomes cannot scale to thousands of individuals and megabases of sequence. Pantograph achieves this through precomputed analysis of shared variation to create syntenic blocks connected by key structural rearrangements. Presence/absence matrices are extended to SNPs and structural rearrangements and binning allows Pantograph to zoom out to chromosome scales.

Pantograph 1.0 is currently able to interactively browse annotated pangenomes of 200+ SARS-CoV-2 individuals and 24 aligned *A. thaliana* individuals. Future planned development will bring in more complete support for haplotypes, minor allele scaling, unrolling distributed repeats, and richer annotation support.

Graph Genome features identified in Pantograph are available through SPARQL queries for integration into any other program. Federated queries are machine readable and can be used to augment annotations as micro-publications. Pantograph is composed of a pipeline of software modules designed for reuse and future collaborations. It is our hope that tools such as Pantograph will enable faster research and discovery, as well as more open knowledge sharing s that resources invested in scientific research can have a multiplicative return to the public.

Software Availability

Pantograph is Open Source, under the Apache 2 license and available on GitHub in the Graph-Genome organization github.com/graph-genome/. You can learn more about the project at GraphGenome.org and browse a live demo of 169 SARS-CoV-2 individuals at graph-genome.github.io/Schematize/. Contributions and feedback are welcome on our GitHub page.

Grant Information

Pantograph design and development research were funded through the 1001 Genomes project under grant BBSRC BB/S004661/1.

JS was employed by Max Planck Institute during 2020 as part of the same 1001 Genomes project to continue work on Pantograph.

Computomics GmbH funded a second hackathon to implement the data pipeline in November 2019.

NBDC/DBCLS BioHackathon 2019 was supported by the Integrated Database Project (Ministry of Education, Culture, Sports, Science and Technology of Japan) and hosted by the National Bioscience Database Center (NBDC) and the Database Center for Life Science (DBCLS).

TY’s work was funded in part by JSPS KAKENHI (grant numbers 16H06279 and 19J21608). HM was supported by AMED under grant number JP19kk0205014.

The deNBI Cloud has provided VMs for developing this project and is currently providing a VM of 28 cores, 64GB RAM and 1TB storage for our COVID19 pangenome visualization.

Chapter 6
Conclusions

6.1 Accessible Visualizations

In this thesis, I have addressed core challenges in the scalability of visualization, exploration, and interaction with genome sequence datasets. A particular emphasis has been placed on making the visualizations as close to the raw sequence data as possible, rather than seeing only the annotations generated by programs. In comparative genomics, this allows researchers to see individual SNPs and indels even at the scale of megabases. Crucially, it also keeps human judgement in the feedback loop as we move to greater degrees of automation. FluentDNA and Pantograph both allow researchers to manually double check the outputs of their assembly as seen with the discovery of endophytes and mitochondria in 3.3.2 and major chromosomal rearrangements in 5.4.3.

FluentDNA allows users to visualize entire gigabase genomes on a single screen and explore them interactively while retrieving sequences. This level of accessibility has been used for museum displays and lab posters to enable interactions that simply would not have happened beforehand. This is not simple iterative improvement but a step change in what is possible in communicating about genomics. Similarly, Graph Genomes are not browsable by a human in their current state except as a way to gather statistics about population heterozygosity. Pantograph will enable scalable pangenomic studies of eukaryotic populations that would not happen otherwise.

6.2 Polyploidy and Fractionation

In the arena of paleopolyploidy, I was able to demonstrate through comparative genomics of 28 assemblies (including *Olea europaea*) that the same gene families were repeatedly lost in independent lines after a shared whole genome duplication. Gene families with greater than five losses were enriched 4.8x over null model expectations. Thus, the process of fractionation (the loss of redundant gene copies after a whole genome duplication) is a somewhat replicable evolutionary process. A similar study was recently published in cotton, based on five species genomes (Chen et al. 2020). The methodologies laid out in this thesis are scalable to large numbers of genomes and larger clades. As every plant clade is the descendant of at least one WGD, many more similar studies could be done. The method could also be extended to account for the compounding effects of multiple WGD, allowing a systematic analysis of taxa fractionation rates to be conducted.

Accounting for the total effects of duplication and fractionation of genes across the history of plants would be a major contribution to our understanding of plant evolution. WGD can have an exponential compounding effect on gene family size, affecting the genes that survive fractionation disproportionately to create areas of sequence space with abundant diversity, such as transcription factors and secreted plant volatiles (Dudareva et al. 2013; Mindrebo et al. 2016). This storehouse of genetic diversity plus plant's inherent physiological and genomic robustness combined with their ability to evolve by whole genome duplication and fractionation may give plants an evolutionary tool largely unavailable to other kingdoms of life. This may help answer fundamental questions, such as: if tree generation times are orders of magnitude slower than insect and fungal generational times how is it that trees have been able to keep up in the evolutionary arms race against pests and pathogens over millions of years? The study of fractionation in *Fraxinus* shows that evolution is a repeatable process which affects a subset of genes while leaving others untouched. The implications of this for pest and pathogen resistance are a promising area for future exploration.

In terms of scale, plants can produce fertile offspring by the millions and still grow to soak in sunlight, weighing in at metric tons as adults. The potential scale increase means that evolutionarily successful combinations can pull in a great deal of energy and produce many more offspring than the most successful mammal ever could. But evolution tends to work in small steps, which is again why WGD has the potential to be advantageous by creating larger selectable units of change which are nonetheless more likely to be clustered around functional sequence space (1.1.2). Duplicated genes are functional homeologs. Duplicated regulatory regions maintain their relative dosage since the entire regulatory network is copied together (1.4.3). Allopolyploid hybrids can show immediate gene expression novelty (1.2.2.3) or even express different parent genomes per tissue type (1.2.2.4). No other mutation mechanism generates this wide range of selectable features genome wide while not being immediately lethal to the organism.

An organism changing its gene copy number can be a type of adaptation via adjusting gene dosage. It also provides branches for future evolutionary search (1.1.1). Second, allopolyploids have massive epistatic innovations in their regulatory machinery, particularly transposon suppression networks, which can lead to immediate transgressive expression and long-term biased or unbiased fractionation (1.2.2.4, 1.2.2.6, 1.2.3). Or in lay terms, they can be brought together to create new and useful combinations, which we can thank for many of our crop breeds today (Stace 1987; Gornicki et al. 2014; Darrow 1955).

These adaptations may explain why plants have been able to outlast the constant bombardment of pests and microbes who would benefit from eating them while plants must simply endure (1.2; Wininger and Rank 2017). Humans rely on plants for every aspect of their lives and we need to better understand this robustness as we go about changing the planet so that we do not accidentally threaten their continued existence.

6.3 The Biological Reality of Phylogenetic Networks

Chapters 2 and 5 of this thesis are focused on how the bioinformatic tools researchers use influence the science and the fact that virtually every tool requires a strictly bifurcating species tree. It seems only appropriate then to note that in section 3.3.1.2 (Notable Species), we excluded a number of hybrids from the study since they occur frequently within the genus, yet in sections 3.3.3 and 4.2.2 the species tree assumes complete separation of species after speciation. We know this is not the case because of evidence for hybridization and incomplete lineage sorting. The deepest nodes of the *Fraxinus* species tree show much lower support for separating the relationships between clades based on concordance values (Kelly et al. 2019). Zohren (2016) demonstrated that *Fraxinu*s species often hybridize and facilitate introgression into distant populations, tying together previous ecological studies (Gerard, Fernandez-Manjarres, and Frascaria-Lacoste 2006; Thomasset et al. 2011). Except for *F. cuspidata*, the branch lengths between clades are short enough to allow incomplete lineage sorting.

It is my hope that tools like PhyloNetworks begin to gain popularity in bioinformatics communities as they better reflect the complex biological realities (Solís-Lemus, Bastide, and Ané 2017; Huson 1998). While not intended by Darwin, the neo-Darwinian synthesis posits the existence of a single globally optimal tree which explains the sequence alignment of all species (Theobald 2010). In practice, this is impossible as locally optimal sequence alignments won't agree with each other, and thus never agree with global optimums. There will always be boundary lines at multiple scales, such as haploblocks (5.3.8) and linkage disequilibrium where the set of "nearest neighbors" have to be reordered (Pook et al. 2019). This means the construction of graph genomes in chapter 5 from local alignments is non-trivial and fraught with the same technical challenges as phylogenetics. The difference is that graph topologies have many more degrees of freedom and thus are capable of expressing patterns of relatedness that species trees cannot.

6.4 The Future of Comparative Genomics

The future of genomic research will increasingly involve comparisons of multiple whole genomes. This is driven by the simple mathematical reality that there are hundreds of thousands of species to sequence and most species have millions of individuals with millions of informative variants to correlate to phenotypes. While studying a single individual reference genome and doing extensive lab experiments can slowly lead to some biological insights, the insights that can be gained from comparative genomics are exponential in nature. Much of biology is built on a simple principle: what is learned in one species can possibly be generalized to thousands of other species. We can set up our infrastructure to maximize these cross-species gains.

The very nature of the Tree of Life is ideal for automated discovery, deep learning, and knowledge networks. However, that scalability would require highly consistent data formats and more importantly, consistent methodologies. The Pantograph team made the decision to begin using RDF and semantic web technologies in order to help facilitate the transition to a more unified and standardized future in bioinformatics. Methodological consistency is still a key unresolved problem which has led to the ongoing replication crisis on multiple fronts (Ioannidis 2005; Engber 2017; Baker 2016). Enabling automated knowledge discovery does not replace the need for researchers but rather amplifies the return on investment in work time and education (Elhai 2011). Clear visualizations for browsers are still necessary to facilitate human reasoning about the complex interactions between structural variants and clinical questions.

Comparative genomics is essential in particular for two important applications: human health and crops. In human health, genetics is faced with the constraint that genetic experimentation is completely disallowed for ethical and practical reasons. However, we make regular use of transgenic mice, rats, bacteria, and human cell colonies. The out of bounds nature of human genetic engineering redirects the substantial needs and resource of human health into interspecies studies. Comparative genomics is the necessary bridge to translate insights gained in these experimental systems to a clinical application. The success or failure of early medical research is dependent on the ability to jump the gap from an experimental setup to human biology.

Analyzing existing variation in the human population serves as a stand-in for creating variation through experiments. Complete coverage of single gene knockout phenotypes have been obtained through forward genetics mutagenesis experiments in *Drosophila melanogaster* and *Arabidopsis thaliana* (Adams and Sekelsky 2002; Alonso and Ecker 2006). With the advancement of CRISPR-Cas9 technology any nucleotide position knockout can be made in an experimental organism. *However*, locating a living person who has a specific mutation requires thorough sequencing of millions of people (All of Us Research Program Investigation 2019). Even after sequencing the whole human population it will not be possible to find particular combinations of three or more SNPs due to the exponential drop-off in the probability of large combinations called the Waiting Time Problem (1000 Genomes Project Consortium et al. 2015; Chatterjee et al. 2014; Tuğrul et al. 2015). Epistatic interactions are observed by sets of variants interacting with each other, either in an individual or compared across species. If we cannot access organismal information about human epistatic interaction then the only remaining avenue is comparative genomics for similar variants in other species with the same pathways.

Interspecies knowledge is even more important in crops because much of the world's food production comes from a surprisingly small number of genera. Cauliflower, broccoli, swede, turnip, kohlrabi, cabbage, collard greens, kale, brussels sprouts, mustard, and rapeseed are all in the genus *Brassica*. This means that studies in one species are rapidly translated to others and have a multiplicative return on investment. Similarly, *Solanum* includes tomato, potato, and eggplant while *Triticeae* contains wheat, spelt, barley, and rye. Furthermore, the nature of molecular biology can lead to surprising generalizations. Research into plant polyploidy helped scientists to better understand genome duplication effects which led to insights in childhood cancers (Storchova and Pellman 2004).

As we move into a future of Big Data, AI, and increasing automation, tools like Pantograph will help to keep human insight central to the process of scientific discovery. This thesis has demonstrated that comparative genomics on the order of tens to hundreds of genomes is now practical. The presence or absence of genes across millions of years of fractionation can be analyzed and simulated to test the degree of repeatability in evolution. Graph Genomes are a likely future of genomics and will lead to a more traceable, machine readable form of genome science. This transition is requiring the development of a new set of tools. As a scientific community, we will also need to continue to address issues of reproducibility and accessibility; and RDF federation is a major tool along that road. The end prize is a new kind of science infrastructure with exponential benefits in every arena from human health, to crops, to basic science.

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.

Adams, Keith L., and Jonathan F. Wendel. 2005. "Polyploidy and Genome Evolution in Plants." *Current Opinion in Plant Biology* 8 (2): 135–41.

Adams, Melissa D., and Jeff J. Sekelsky. 2002. "From Sequence to Phenotype: Reverse Genetics in Drosophila Melanogaster." *Nature Reviews. Genetics* 3 (3): 189–98.

Aharoni, Amir, Leonid Gaidukov, Shai Yagur, Lilly Toker, Israel Silman, and Dan S. Tawfik. 2004. "Directed Evolution of Mammalian Paraoxonases PON1 and PON3 for Bacterial Expression and Catalytic Specialization." *Proceedings of the National Academy of Sciences of the United States of America* 101 (2): 482–87.

Ahmed A., Alejandro G., Daniel C., Dreycey A., Eric D., Fawaz D., Glenn H., Michael J., Sagayamary S., Zeng Q. n.d. *Structural Variation with Annotated Graph Genomes (SWAGG).* Github. Accessed October 20, 2020. https://github.com/collaborativebioinformatics/swagg.

All of Us Research Program Investigators, Joshua C. Denny, Joni L. Rutter, David B. Goldstein, Anthony Philippakis, Jordan W. Smoller, Gwynne Jenkins, and Eric Dishman. 2019. "The 'All of Us' Research Program." *The New England Journal of Medicine* 381 (7): 668–76.

Alonso, Jose M., and Joseph R. Ecker. 2006. "Moving Forward in Reverse: Genetic Technologies to Enable Genome-Wide Phenomic Screens in Arabidopsis." *Nature Reviews. Genetics* 7 (7): 524–36.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Arakawa, Kazuharu, Satoshi Tamaki, Nobuaki Kono, Nobuhiro Kido, Keita Ikegami, Ryu Ogawa, and Masaru Tomita. 2009. "Genome Projector: Zoomable Genome Map with Multiple Views." *BMC Bioinformatics* 10 (January): 31.

Arrigo, Nils, and Michael S. Barker. 2012. "Rarely Successful Polyploids and Their Legacy in Plant Genomes." *Current Opinion in Plant Biology* 15 (2): 140–46.

Asaf, Sajjad, Abdul Latif Khan, Muhammad Aaqil Khan, Ahmed Al-Harrasi, and In-Jung Lee. 2018. "Complete Genome Sequencing and Analysis of Endophytic Sphingomonas Sp. LK11 and Its Potential in Plant Growth." *3 Biotech* 8 (9): 389.

Bahlo, Melanie, Mark F. Bennett, Peter Degorski, Rick M. Tankard, Martin B. Delatycki, and Paul J. Lockhart. 2018. "Recent Advances in the Detection of Repeat Expansions with Short-Read next-Generation Sequencing." *F1000Research* 7 (June). https://doi.org/10.12688/f1000research.13980.1.

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." May 25, 2016. https://doi.org/10.1038/533452a.

Ballouz, Sara, Alexander Dobin, and Jesse A. Gillis. 2019. “Is It Time to Change the Reference Genome?” *Genome Biology* 20 (1): 159.

Barker, Michael S., Nolan C. Kane, Marta Matvienko, Alexander Kozik, Richard W. Michelmore, Steven J. Knapp, and Loren H. Rieseberg. 2008. “Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years.” *Molecular Biology and Evolution* 25 (11): 2445–55.

Barreda, Viviana D., Luis Palazzesi, Maria C. Tellería, Eduardo B. Olivero, J. Ian Raine, and Félix Forest. 2015. “Early Evolution of the Angiosperm Clade Asteraceae in the Cretaceous of Antarctica.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (35): 10989–94.

Barsky, Aaron, Tamara Munzner, Jennifer Gardy, and Robert Kincaid. 2008. “Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context.” IEEE Transactions on Visualization and Computer Graphics 14 (6): 1253–60.

Bataillon, T. 2000. “Estimation of Spontaneous Genome-Wide Mutation Rate Parameters: Whither Beneficial Mutations?” *Heredity* 84 (Pt 5) (May): 497–501.

Bekaert, Michaël, Patrick P. Edger, J. Chris Pires, and Gavin C. Conant. 2011. “Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints.” *The Plant Cell* 23 (5): 1719–28.

Belancio, Victoria P., Astrid M. Roy-Engel, and Prescott L. Deininger. 2010. “All Y’all Need to Know 'Bout Retroelements in Cancer.” *Seminars in Cancer Biology* 20 (4): 200–210.

Beyer, Wolfgang, Adam M. Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, and Daniel R. Zerbino. 2019. “Sequence Tube Maps: Making Graph Genomes Intuitive to Commuters.” *Bioinformatics* 35 (24): 5318–20.

Bially, T. 1969. “Space-Filling Curves: Their Generation and Their Application to Bandwidth Reduction.” *IEEE Transactions on Information Theory / Professional Technical Group on Information Theory* 15 (6): 658–64.

Bierkandt, Katrin, and Jens Bierkandt. 2009. “DNA Rainbow.” DNA Rainbow. 2009. https://www.dna-rainbow.org/.

Biggs, N. L., E. K. Lloyd, and R. J. Wilson. 1986. *Graph Theory 1736 - 1936*. Oxford.

Birchler, James A., Nicole C. Riddle, Donald L. Auger, and Reiner A. Veitia. 2005. “Dosage Balance in Gene Regulation: Biological Implications.” *Trends in Genetics: TIG* 21 (4): 219–26.

Birren, Bruce, and Elinor Karlsson. 2021. “The 200 Mammals Project: Sequencing Genomes by a Novel Cost-Effective Method, Yielding a High Resolution Annotation of the Human Genome.” 10087672. Broad Institute, Inc. https://grantome.com/grant/NIH/R01-HG008742-03S1.

Blanc, Guillaume, and Kenneth H. Wolfe. 2004. “Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes.” *The Plant Cell*. https://doi.org/10.1105/tpc.021345.

Böndel, Katharina B., Susanne A. Kraemer, Toby Samuels, Deirdre McClean, Josianne Lachapelle, Rob W. Ness, Nick Colegrave, and Peter D. Keightley. 2019. “Inferring the Distribution of Fitness Effects of Spontaneous Mutations in Chlamydomonas Reinhardtii.” *PLoS Biology* 17 (6): e3000192.

Bottani, Samuel, Nicolae Radu Zabet, Jonathan F. Wendel, and Reiner A. Veitia. 2018. “Gene Expression Dominance in Allopolyploids: Hypotheses and Models.” *Trends in Plant Science* 23 (5): 393–402.

Bowers, John E., Brad A. Chapman, Junkang Rong, and Andrew H. Paterson. 2003. “Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events.” *Nature* 422 (6930): 433–38.

British Association for the Advancement of Science. 1915. “Report of the British Association for the Advancement of Science” 84th Meeting (1914). https://www.biodiversitylibrary.org/item/95821.

Brocchieri, Luciano, and Samuel Karlin. 2005. “Protein Length in Eukaryotic and Prokaryotic Proteomes.” *Nucleic Acids Research* 33 (10): 3390–3400.

Brown, C. Titus, Dominik Moritz, Michael P. O’Brien, Felix Reidl, Taylor Reiter, and Blair D. Sullivan. 2020. “Exploring Neighborhoods in Large Metagenome Assembly Graphs Using Spacegraphcats Reveals Hidden Sequence Diversity.” *Genome Biology* 21 (1): 164.

Buchanan, K. L., and J. W. Murphy. 1998. “What Makes Cryptococcus Neoformans a Pathogen?” *Emerging Infectious Diseases* 4 (1): 71–83.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Buels, Robert, Eric Yao, Colin M. Diesh, Richard D. Hayes, Monica Munoz-Torres, Gregg Helt, David M. Goodstein, et al. 2016. “JBrowse: A Dynamic Web Platform for Genome Visualization and Analysis.” *Genome Biology* 17 (April): 66.

Buggs, Richard J. A., Srikar Chamala, Wei Wu, Jennifer A. Tate, Patrick S. Schnable, Douglas E. Soltis, Pamela S. Soltis, and W. Brad Barbazuk. 2012. “Rapid, Repeated, and Clustered Loss of Duplicate Genes in Allopolyploid Plant Populations of Independent Origin.” *Current Biology: CB* 22 (3): 248–52.

Buggs, Richard J. A., Linjing Zhang, Nicholas Miles, Jennifer A. Tate, Lu Gao, Wu Wei, Patrick S. Schnable, W. Brad Barbazuk, Pamela S. Soltis, and Douglas E. Soltis. 2011. “Transcriptomic Shock Generates Evolutionary Novelty in a Newly Formed, Natural Allopolyploid Plant.” *Current Biology: CB* 21 (7): 551–56.

Bungard, Dixie, Jacob S. Copple, Jing Yan, Jimmy J. Chhun, Vlad K. Kumirov, Scott G. Foy, Joanna Masel, Vicki H. Wysocki, and Matthew H. J. Cordes. 2017. “Foldability of a Natural *De Novo* Evolved Protein.” *Structure* 25 (11): 1687–96.e4.

"CAFE Tutorial: Computational Analysis of Gene Family Evolution." 2016. https://iu.app.box.com/v/cafetutorial-pdf.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–44.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael a. Calderwood, Muhammed a. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. "Proto-Genes and *de Novo* Gene Birth." *Nature* 487 (7407): 370–74.

Chang, Peter L., Brian P. Dilkes, Michelle McMahon, Luca Comai, and Sergey V. Nuzhdin. 2010. "Homoeolog-Specific Retention and Use in Allotetraploid Arabidopsis Suecica Depends on Parent of Origin and Network Partners." *Genome Biology* 11 (12): R125.

Chatterjee, Krishnendu, Andreas Pavlogiannis, Ben Adlam, and Martin A. Nowak. 2014. "The Time Scale of Evolutionary Innovation." *PLoS Computational Biology* 10 (9): e1003818.

Cheng, Feng, Chao Sun, Jian Wu, James Schnable, Margaret R. Woodhouse, Jianli Liang, Chengcheng Cai, Michael Freeling, and Xiaowu Wang. 2016. "Epigenetic Regulation of Subgenome Dominance Following Whole Genome Triplication in Brassica Rapa." *The New Phytologist* 211 (1): 288–99.

Cheng, Feng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. 2018. "Gene Retention, Fractionation and Subgenome Differences in Polyploid Plants." *Nature Plants* 4 (5): 258–68.

Chen, Hong, Lin Xu, and Zhenglong Gu. 2008. "Regulation Dynamics of WGD Genes during Yeast Metabolic Oscillation." *Molecular Biology and Evolution* 25 (12): 2513–16.

Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87.

Chin, Jason. 2019. "Constructing A Graph for Genome Comparison Swiftly - Jason Chin - Medium." Medium. November 4, 2019. https://medium.com/@infoecho/constructing-a-graph-for-genome-comparison-swiftly-d47dcd7eae5d.

Clark, James W., and Philip C. J. Donoghue. 2017. "Constraining the Timing of Whole Genome Duplication in Plant Evolutionary History." *Proceedings. Biological Sciences / The Royal Society* 284 (1858). https://doi.org/10.1098/rspb.2017.0912.

Clark, Tim, Paolo N. Ciccarese, and Carole A. Goble. 2014. "Micropublications: A Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications." *Journal of Biomedical Semantics* 5 (July): 28.

Coate, Jeremy E., Jessica A. Schlueter, Adam M. Whaley, and Jeff J. Doyle. 2011. “Comparative Evolution of Photosynthetic Genes in Response to Polyploid and Nonpolyploid Duplication.” *Plant Physiology* 155 (4): 2081–95.

Conti, Gregory, Sergey Bratus, Anna Shubina, Andrew Lichtenberg, Roy Ragsdale, Robert Perez-Alemany, Benjamin Sangster, and Matthew Supan. 2010. “A Visual Study of Primitive Binary Fragment Types.” *White Paper, Black Hat USA*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.4782&rep=rep1&type=pdf.

Cortesi, Aldo. 2011. “Visualizing Binaries with Space-Filling Curves.” Binvis.io. December 23, 2011. https://corte.si/posts/visualisation/binvis/index.html.

Cross, J. C. 2005. “How to Make a Placenta: Mechanisms of Trophoblast Cell Differentiation in Mice--a Review.” *Placenta* 26 Suppl A (April): S3–9.

Crusoe, Michael R., Hussien F. Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, et al. 2015. “The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis.” *F1000Research* 4 (September): 900.

Darlington, Cyril Dean, and Others. 1937. “Recent Advances in Cytology.” *Recent Advances in Cytology.*, no. 2nd Ed. https://www.cabdirect.org/cabdirect/abstract/19371601553.

De Bie, Tijl, Nello Cristianini, Jeffery P. Demuth, and Matthew W. Hahn. 2006. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” *Bioinformatics* 22 (10): 1269–71.

Delaneau, O., M. Zazhytska, C. Borel, G. Giannuzzi, G. Rey, C. Howald, S. Kumar, et al. 2019. “Chromatin Three-Dimensional Interactions Mediate Genetic Effects on Gene Expression.” *Science* 364 (6439): eaat8266.

Deschavanne, P. J., A. Giron, J. Vilain, G. Fagot, and B. Fertil. 1999. “Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences.” *Molecular Biology and Evolution* 16 (10): 1391–99.

De Smet, Riet, Keith L. Adams, Klaas Vandepoele, Marc C. E. Van Montagu, Steven Maere, and Yves Van de Peer. 2013. “Convergent Gene Loss Following Gene and Genome Duplications Creates Single-Copy Families in Flowering Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (8): 2898–2903.

D’Hont, Angélique, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, Olivier Garsmeur, Benjamin Noel, et al. 2012. “The Banana (Musa Acuminata) Genome and the Evolution of Monocotyledonous Plants.” *Nature* 488 (7410): 213–17.

Dodsworth, Steven, Mark W. Chase, and Andrew R. Leitch. 2016. “Is Post-Polyploidization Diploidization the Key to the Evolutionary Success of Angiosperms?” *Botanical Journal of the Linnean Society. Linnean Society of London* 180 (1): 1–5.

Douglas E. Soltis, María Claudia Segovia-Salcedo, Ingrid Jordon-Thaden, Lucas Majure, Nicolas M. Miles, Evgeny V. Mavrodiev, Wenbin Mei, María Beatriz Cortez, Pamela S.

Soltis, and Matthew A. Gitzendanner. 2014. “Are Polyploids Really Evolutionary Dead-Ends (again)? A Critical Reappraisal of Mayrose et Al. (2011).” *The New Phytologist* 202 (4): 1105–17.

Douglas, Gavin M., Gesseca Gos, Kim A. Steige, Adriana Salcedo, Karl Holm, Emily B. Josephs, Ramesh Arunkumar, et al. 2015. “Hybrid Origins and the Earliest Stages of Diploidization in the Highly Successful Recent Polyploid Capsella Bursa-Pastoris.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1412277112.

Doyle, Jeff J., and Jeremy E. Coate. 2019. “Polyploidy, the Nucleotype, and Novelty: The Impact of Genome Doubling on the Biology of the Cell.” *International Journal of Plant Sciences* 180 (1): 1–52.

Draper, Bruce. n.d. “Communicating with Computers.” DARPA. Accessed March 8, 2021. https://www.darpa.mil/program/communicating-with-computers.

Duarte, Jill M., Liying Cui, P. Kerr Wall, Qing Zhang, Xiaohong Zhang, Jim Leebens-Mack, Hong Ma, Naomi Altman, and Claude W. dePamphilis. 2006. “Expression Pattern Shifts Following Duplication Indicative of Subfunctionalization and Neofunctionalization in Regulatory Genes of Arabidopsis.” *Molecular Biology and Evolution* 23 (2): 469–78.

Ecco, Gabriela, Michael Imbeault, and Didier Trono. 2017. “KRAB Zinc Finger Proteins.” *Development* 144 (15): 2719–29.

Ecco, Gabriela, Marco Cassano, Annamaria Kauzlaric, Julien Duc, Andrea Coluccio, Sandra Offner, Michaël Imbeault, Helen M. Rowe, Priscilla Turelli, and Didier Trono. 2016. “Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues.” *Developmental Cell* 36 (6): 611–23.

Edger, Patrick P., Michael R. McKain, Kevin A. Bird, and Robert VanBuren. 2018. “Subgenome Assignment in Allopolyploids: Challenges and Future Directions.” *Current Opinion in Plant Biology* 42 (April): 76–80.

Edger, Patrick P., and J. Chris Pires. 2009. “Gene and Genome Duplications: The Impact of Dosage-Sensitivity on the Fate of Nuclear Genes.” *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 17 (5): 699–717.

Egea, N., and N. Lang-Unnasch. 1996. “Phylogeny of the Large Extrachromosomal DNA of Organisms in the Phylum Apicomplexa.” *The Journal of Eukaryotic Microbiology* 43 (2): 158.

Eizenga, Jordan, and Adam M. Novak. 2020. “Precision Pangenomics.” *Annual Reviews in Genomics and Human Genetics*.

Eizenga, Jordan M., Adam M. Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, Simon Heumos, Glenn Hickey, Vincenza Colonna, Benedict Paten, and Erik Garrison. 2020. “Efficient Dynamic Variation Graphs.” *Bioinformatics*, July. https://doi.org/10.1093/bioinformatics/btaa640.

Elhai, Jeff. 2011. “Humans, Computers, and the Route to Biological Insights: Regaining Our Capacity for Surprise.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (7): 867–78.

Ellison, Christopher E., and Doris Bachtrog. 2015. “Non-Allelic Gene Conversion Enables Rapid Evolutionary Change at Multiple Regulatory Sites Encoded by Transposable Elements.” *eLife* 4 (February). https://doi.org/10.7554/eLife.05899.

Emery, Marianne, M. Madeline S. Willis, Yue Hao, Kerrie Barry, Khouanchy Oakgrove, Yi Peng, Jeremy Schmutz, et al. 2018. “Preferential Retention of Genes from One Parental Genome after Polyploidy Illustrates the Nature and Scope of the Genomic Conflicts Induced by Hybridization.” *PLoS Genetics* 14 (3): e1007267.

Emms, David M., and Steven Kelly. 2015. “OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy.” *Genome Biology* 16 (August): 157.

Emms, D.M. and Kelly, S. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Emms, D., Seaman, J. 2020. “Permission to use a figure derivative.” November 2020. https://github.com/davidemms/OrthoFinder/issues/418

Engber, Daniel. 2017. “Daryl Bem Proved ESP Is Real. Which Means Science Is Broken.” Slate. June 7, 2017. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.

Flagel, Lex E., and Jonathan F. Wendel. 2009. “Gene Duplication and Evolutionary Novelty in Plants.” *The New Phytologist* 183 (3): 557–64.

Flagel, Lex, Joshua Udall, Dan Nettleton, and Jonathan Wendel. 2008. “Duplicate Gene Expression in Allopolyploid Gossypium Reveals Two Temporally Distinct Phases of Expression Evolution.” *BMC Biology* 6 (April): 16.

Flemming, A. J., Z. Z. Shen, A. Cunha, S. W. Emmons, and A. M. Leroi. 2000. “Somatic Polyploidization and Cellular Proliferation Drive Body Size Evolution in Nematodes.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (10): 5285–90.

Forney, G. David. 1973. “The Viterbi Algorithm.” *Proceedings of the IEEE* 61 (3): 268–78.

Freeling, Michael. 2009. “Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition.” *Annual Review of Plant Biology* 60: 433–53.

Freeling, Michael, and Brian C. Thomas. 2006. “Gene-Balanced Duplications, like Tetraploidy, Provide Predictable Drive to Increase Morphological Complexity.” *Genome Research* 16 (7): 805–14.

Frosst, Nicholas, and Geoffrey Hinton. 2017. “Distilling a Neural Network Into a Soft Decision Tree.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1711.09784.

Galbraith, David W., Kristi R. Harkins, and Steven Knapp. 1991. “Systemic Endopolyploidy in Arabidopsis Thaliana.” *Plant Physiology* 96: 985–89.

Gardner, M., D. Williamson, and R. Wilson. 1991. “A Circular DNA in Malaria Parasites Encodes an RNA Polymerase like that of Prokaryotes and Chloroplasts☆.” *Molecular and Biochemical Parasitology*. https://doi.org/10.1016/0166-6851(91)90227-w.

Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, et al. 2018. “Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference.” *Nature Biotechnology* 36 (9): 875–79.

Garsmeur, Olivier, James C. Schnable, Ana Almeida, Cyril Jourda, Angélique D’Hont, and Michael Freeling. 2014. “Two Evolutionarily Distinct Classes of Paleopolyploidy.” *Molecular Biology and Evolution* 31 (2): 448–54.

Gerard, Pierre R., Juan F. Fernandez-Manjarres, and Nathalie Frascaria-Lacoste. 2006. “Temporal Cline in a Hybrid Zone Population between Fraxinus Excelsior L. and Fraxinus Angustifolia Vahl.” *Molecular Ecology* 15 (12): 3655–67.

Giardine, Belinda, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. “Galaxy: A Platform for Interactive Large-Scale Genome Analysis.” *Genome Research* 15 (10): 1451–55.

Gómez, John, Leyla J. García, Gustavo A. Salazar, Jose Villaveces, Swanand Gore, Alexander García, Maria J. Martín, et al. 2013. “BioJS: An Open Source JavaScript Framework for Biological Data Visualization.” *Bioinformatics* 29 (8): 1103–4.

Gordon, Sean P., Bruno Contreras-Moreira, Daniel P. Woods, David L. Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, et al. 2017. “Extensive Gene Content Variation in the Brachypodium Distachyon Pan-Genome Correlates with Population Structure.” *Nature Communications* 8 (1): 2184.

Grob, Stefan, Marc W. Schmid, and Ueli Grossniklaus. 2014. “Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the Flamenco Locus of Drosophila.” *Molecular Cell* 55 (5): 678–93.

Gryder, B. E., M. E. Yohe, H. C. Chou, X. Zhang, and J. Marques. 2017. “PAX3–FOXO1 Establishes Myogenic Super Enhancers and Confers BET Bromodomain Vulnerability.” *Cancer Discovery*. http://cancerdiscovery.aacrjournals.org/content/7/8/884.abstract.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. “Cooperative Inverse Reinforcement Learning.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1606.03137.

Halo, Boshra Ahmed, Abdul Latif Khan, Muhammad Waqas, Ahmed Al-Harrasi, Javid Hussain, Liaqat Ali, Muhammad Adnan, and In-Jung Lee. 2015. “Endophytic Bacteria (Sphingomonas Sp. LK11) and Gibberellin Can Improve Solanum Lycopersicum Growth and Oxidative Stress under Salinity.” *Journal of Plant Interactions* 10 (1): 117–25.

Han, Mira V., Gregg W. C. Thomas, Jose Lugo-Martinez, and Matthew W. Hahn. 2013. “Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3.” *Molecular Biology and Evolution* 30 (8): 1987–97.

Harris, Robert S. 2007. “Improved Pairwise Alignmnet of Genomic DNA.” Edited by Webb Colby Miller. PhD, Pennsylvania State. https://etda.libraries.psu.edu/catalog/7971.

Haug-Baltzell, Asher, Sean A. Stephens, Sean Davey, Carlos E. Scheidegger, and Eric Lyons. 2017. “SynMap2 and SynMap3D: Web-Based Whole-Genome Synteny Browsers.” Bioinformatics 33 (14): 2197–98.

Haverkort, Herman, and Freek van Walderveen. 2010. “Locality and Bounding-Box Quality of Two-Dimensional Space-Filling Curves.” *Computational Geometry: Theory and Applications* 43 (2): 131–47.

Hawkins, Jennifer S., Stephen R. Proulx, Ryan A. Rapp, and Jonathan F. Wendel. 2009. “Rapid DNA Loss as a Counterbalance to Genome Expansion through Retrotransposon Proliferation in Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (42): 17811–16.

Hegarty, Matthew J., Richard J. Abbott, and Simon J. Hiscock. 2012. “Allopolyploid Speciation in Action: The Origins and Evolution of Senecio Cambrensis.” In *Polyploidy and Genome Evolution*, edited by Pamela S. Soltis and Douglas E. Soltis, 245–70. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hegarty, Matthew J., Gary L. Barker, Ian D. Wilson, Richard J. Abbott, Keith J. Edwards, and Simon J. Hiscock. 2006. “Transcriptome Shock after Interspecific Hybridization in Senecio Is Ameliorated by Genome Duplication.” *Current Biology: CB* 16 (16): 1652–59.

Hellsten, Uffe, Kevin M. Wright, Jerry Jenkins, Shengqiang Shu, Yaowu Yuan, Susan R. Wessler, Jeremy Schmutz, John H. Willis, and Daniel S. Rokhsar. 2013. “Fine-Scale Variation in Meiotic Recombination in Mimulus Inferred from Population Shotgun Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (48): 19478–82.

Hennig, André, Jörg Bernhardt, and Kay Nieselt. 2015. “Pan-Tetris: An Interactive Visualisation for Pan-Genomes.” *BMC Bioinformatics* 16 Suppl 11 (August): S3.

Henry, Nathalie, Jean-Daniel Fekete, and Michael J. McGuffin. 2007. “NodeTrix: A Hybrid Visualization of Social Networks.” *IEEE Transactions on Visualization and Computer Graphics* 13 (6): 1302–9.

Hickey, Glenn, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. 2013. “HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments.” *Bioinformatics* 29 (10): 1341–42.

Hohmann, Nora, Eva M. Wolf, Martin A. Lysak, and Marcus A. Koch. 2015. “A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History.” *The Plant Cell* 27 (10): 2770–84.

Hossain, Mahmud Shahriar, Christopher Andrews, Naren Ramakrishnan, and Chris North. 2011. “Helping Intelligence Analysts Make Connections.” In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. aaai.org. https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewPaper/3937.

Hovav, Ran, Adi Faigenboim-Doron, Noa Kadmon, Guanjing Hu, Xia Zhang, Joseph P. Gallagher, and Jonathan F. Wendel. 2015. “A Transcriptome Profile for Developing Seed of Polyploid Cotton.” *The Plant Genome* 8. https://doi.org/10.3835/plantgenome2014.08.0041.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. “Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists.” *Nucleic Acids Research* 37 (1): 1–13.

Huntley, Rachael P., Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O’Donovan. 2015. “The GOA Database: Gene Ontology Annotation Updates for 2015.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1113.

Huson, D. H. 1998. “SplitsTree: Analyzing and Visualizing Evolutionary Data.” *Bioinformatics* 14 (1): 68–73.

Imbeault, Michaël, Pierre-Yves Helleboid, and Didier Trono. 2017. “KRAB Zinc-Finger Proteins Contribute to the Evolution of Gene Regulatory Networks.” *Nature*, March. https://doi.org/10.1038/nature21683.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124.

Ioannidis, John P. A., Marcus R. Munafò, Paolo Fusar-Poli, Brian A. Nosek, and Sean P. David. 2014. “Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence, and Prevention.” *Trends in Cognitive Sciences* 18 (5): 235–41.

Jaillon, Olivier, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, et al. 2007. “The Grapevine Genome Sequence Suggests Ancestral Hexaploidization in Major Angiosperm Phyla.” *Nature* 449 (7161): 463–67.

Jakubowska, Joanna, Ela Hunt, Matthew Chalmers, Martin McBride, and Anna F. Dominiczak. 2007. “VisGenome: Visualization of Single and Comparative Genome Representations.” *Bioinformatics* 23 (19): 2641–42.

Jeff J. Doyle, Lex E. Flagel, Andrew H. Paterson, Ryan A. Rapp, Douglas E. Soltis, Pamela S. Soltis, and Jonathan F. Wendel. 2008. “Evolutionary Genetics of Genome Merger and Doubling in Plants.” *Annual Review of Genetics*. ftp://ftp.ufv.br/DBG/Filogenia_molecular/usuarios/karla/Lyderson/2010/artigos/Hibridos/Doyle2008.pdf.

Jiao, Yuannian, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, et al. 2011. “Ancestral Polyploidy in Seed Plants and Angiosperms.” *Nature* 473 (7345): 97–100.

Joseph, Jijoy, and Roschen Sasikumar. 2006. “Chaos Game Representation for Comparison of Whole Genomes.” *BMC Bioinformatics* 7 (May): 243.

Joshi, N. A., J. N. Fass, and Others. 2011. “Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33)[Software].”

Julca, Irene, Marina Marcet-Houben, Pablo Vargas, and Toni Gabaldon. 2017. “Phylogenomics of the Olive Tree (Olea Europaea) Disentangles Ancient Allo- and Autopolyploidizations in Lamiales.” *bioRxiv*. https://doi.org/10.1101/163063.

Julca, Irene, Marina Marcet-Houben, Pablo Vargas, and Toni Gabaldón. 2018. “Phylogenomics of the Olive Tree (Olea Europaea) Reveals the Relative Contribution of Ancient Allo- and Autopolyploidization Events.” *BMC Biology* 16 (1): 15.

Kashkush, Khalil, Moshe Feldman, and Avraham A. Levy. 2002. “Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid.” *Genetics* 160 (4): 1651–59.

Katoh, Kazutaka, John Rozewicki, and Kazunori D. Yamada. 2017. “MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization.” *Briefings in Bioinformatics*, September. https://doi.org/10.1093/bib/bbx108.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Kaya, Naz, and Helen Epps. n.d. “Color-Emotion Associations: Past Experience and Personal Preference.” In , 31–34. Jose Luis Caivano.

Keilwagen, Jens, Frank Hartung, and Jan Grau. 2019. “GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-Seq Data.” *Methods in Molecular Biology* 1962: 161–77.

Keilwagen, Jens, Michael Wenk, Jessica L. Erickson, Martin H. Schattat, Jan Grau, and Frank Hartung. 2016. “Using Intron Position Conservation for Homology-Based Gene Prediction.” *Nucleic Acids Research* 44 (9): e89.

Kejnovsky, Eduard, Ilia J. Leitch, and Andrew R. Leitch. 2009. “Contrasting Evolutionary Dynamics between Angiosperm and Mammalian Genomes.” *Trends in Ecology & Evolution* 24 (10): 572–82.

Kelleher, Jerome, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. 2019. “Inferring Whole-Genome Histories in Large Population Datasets.” *Nature Genetics* 51 (9): 1330–38.

Kelly, Laura J., William J. Plumb, David W. Carey, Mary E. Mason, Endymion D. Cooper, William Crowther, Alan T. Whittemore, Stephen J. Rossiter, Jennifer L. Koch, and Richard J. A. Buggs. 2020. “Convergent Molecular Evolution among Ash Species Resistant to the Emerald Ash Borer.” *Nature Ecology & Evolution* 4 (8): 1116–28.

Kemkemer, Claus, Matthias Kohn, Hildegard Kehrer-Sawatzki, Peter Minich, Josef Högel, Lutz Froenicke, and Horst Hameister. 2006. “Reconstruction of the Ancestral Ferungulate Karyotype by Electronic Chromosome Painting (E-Painting).” *Chromosome*

Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology 14 (8): 899–907.

Khouri-Saba, Paul, Antoine Vandecreme, Mary Brady, Kiran Bhadriraju, and Peter Bajcsy. 2013. “Deep Zoom Tool for Advanced Interactivity with High-Resolution Images.” Spienewsroom. https://isg.nist.gov/deepzoomweb/resources/nist/paper/deepZoom_published004866_10.pdf.

Khurana, Himanshu, Jim Basney, Mehedi Bakht, Mike Freemon, Von Welch, and Randy Butler. 2009. “Palantir: A Framework for Collaborative Incident Response and Investigation.” In *Proceedings of the 8th Symposium on Identity and Trust on the Internet*, 38–51. IDtrust ’09. New York, NY, USA: ACM.

Kidd, Jeffrey M., Tina Graves, Tera L. Newman, Robert Fulton, Hillary S. Hayden, Maika Malig, Joelle Kallicki, Rajinder Kaul, Richard K. Wilson, and Evan E. Eichler. 2010. “A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms.” *Cell* 143 (5): 837–47.

Kimura, M. 1979. “The Neutral Theory of Molecular Evolution.” *Scientific American* 241 (5): 98–100, 102, 108 passim.

Kistler, Amy L., Dale R. Webster, Silvi Rouskin, Vince Magrini, Joel J. Credle, David P. Schnurr, Homer A. Boushey, Elaine R. Mardis, Hao Li, and Joseph L. DeRisi. 2007. “Genome-Wide Diversity and Selective Pressure in the Human Rhinovirus.” *Virology Journal* 4 (1): 40.

Knuth, Donald E. 1970. “The Analysis of Algorithms.” In *Actes Du Congres International Des Mathématiciens*. Vol. 3. history.cs.ncl.ac.uk. http://history.cs.ncl.ac.uk/seminars/139.pdf.

Konrad, Anke, Ashley I. Teufel, Johan a. Grahnen, and David a. Liberles. 2011. “Toward a General Model for the Evolutionary Dynamics of Gene Duplicates.” *Genome Biology and Evolution* 3 (January): 1197–1209.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research* 19 (9): 1639–45.

Kuhn, Robert M., David Haussler, and W. James Kent. 2013. “The UCSC Genome Browser and Associated Tools.” *Briefings in Bioinformatics* 14 (2): 144–61.

Kuraku, Shigehiro. 2011. “Hox Gene Clusters of Early Vertebrates: Do They Serve as Reliable Markers for Genome Evolution?” *Genomics, Proteomics & Bioinformatics* 9 (3): 97–103.

Laetsch, Dominik R., and Mark L. Blaxter. 2017. “BlobTools: Interrogation of Genome Assemblies.” *F1000Research* 6 (July). https://doi.org/10.12688/f1000research.12232.1.

Lamesch, Philippe, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. “The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools.” Nucleic Acids Research 40 (Database issue): D1202–10.

Landis, Jacob B., Douglas E. Soltis, Zheng Li, Hannah E. Marx, Michael S. Barker, David C. Tank, and Pamela S. Soltis. 2018. “Impact of Whole-Genome Duplication Events on Diversification Rates in Angiosperms.” *American Journal of Botany* 105 (3): 348–63.

Langham, Richard J., Justine Walsh, Molly Dunn, Cynthia Ko, Stephen A. Goff, and Michael Freeling. 2004. “Genomic Duplication, Fractionation and the Origin of Regulatory Novelty.” *Genetics* 166 (2): 935–45.

Leader, David P. 2004. “BugView: A Browser for Comparing Genomes.” *Bioinformatics* 20 (1): 129–30.

Lechat, Pierre, Erika Souche, and Ivan Moszer. 2013. “SynTView - an Interactive Multi-View Genome Browser for next-Generation Comparative Microorganism Genomics.” *BMC Bioinformatics* 14 (September): 277.

Lee, Hyun O., Jean M. Davidson, and Robert J. Duronio. 2009. “Endoreplication: Polyploidy with Purpose.” *Genes & Development* 23 (21): 2461–77.

Leitch, A. R., and I. J. Leitch. 2012. “Ecological and Genetic Factors Linked to Contrasting Genome Dynamics in Seed Plants.” *The New Phytologist* 194 (3): 629–46.

Leitch, Ilia J., and Steven Dodsworth. 2001. “Endopolyploidy in Plants.” In *eLS*. John Wiley & Sons, Ltd.

Leitch, Ilia J., and Andrew R. Leitch. 2013. “Genome Size Diversity and Evolution in Land Plants.” In *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*, edited by Johann Greilhuber, Jaroslav Dolezel, and Jonathan F. Wendel, 307–22. Vienna: Springer Vienna.

Levin, Donald A. 1975. “Minority Cytotype Exclusion In Local Plant Populations.” *Taxon* 24 (1): 35–43.

Levin, Donald A. 1983. “Polyploidy and Novelty in Flowering Plants.” *The American Naturalist* 122 (1): 1–25.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.

Liverpool, Layal. 2019. “Genetic Studies Have Missed Important Gene Variants in African People.” *New Scientist*, October 31, 2019. https://www.newscientist.com/article/2221957-genetic-studies-have-missed-important-gene-variants-in-african-people/.

Li, Yong-Hua, Wei Zhang, and Yong Li. 2015. “Transcriptomic Analysis of Flower Blooming in Jasminum Sambac through *De Novo* RNA Sequencing.” *Molecules* 20 (6): 10734–47.

Li, Zhen, Jonas Defoort, Setareh Tasdighian, Steven Maere, Yves Van de Peer, and Riet De Smet. 2016. “Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms.” *The Plant Cell* 28 (2): 326–44.

Lynch, Michael, and John S. Conery. 2003. “The Evolutionary Demography of Duplicate Genes.” *Journal of Structural and Functional Genomics* 3 (1-4): 35–44.

Lyons, Eric Harold. 2008. *CoGe, a New Kind of Comparative Genomics Platform: Insights into the Evolution of Plant Genomes*. University of California, Berkeley.

Lyons, Eric, Brent Pedersen, Josh Kane, and Michael Freeling. 2008. “The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy That Predates the Rosids.” *Tropical Plant Biology* 1 (3-4): 181–90.

Lyons, Eric, and Michael Freeling. 2008. “How to Usefully Compare Homologous Plant Genes and Chromosomes as DNA Sequences.” The Plant Journal: For Cell and Molecular Biology 53 (4): 661–73.

Maaten, and Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9: 2579–2605.

Maere, Steven, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. 2005. “Modeling Gene and Genome Duplications in Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (15): 5454–59.

Makino, Takashi, and Aoife McLysaght. 2010. “Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9270–74.

Mahendran, Aravindh and Vedaldi, Andrea. 2016. “Visualizing Deep Convolutional Neural Networks Using Natural Pre-images.” *International Journal of Computer Vision* 120, 233–255

Mandáková, Terezie, Zheng Li, Michael S. Barker, and Martin A. Lysak. 2017. “Diverse Genome Organization Following 13 Independent Mesopolyploid Events in Brassicaceae Contrasts with Convergent Patterns of Gene Retention.” *The Plant Journal: For Cell and Molecular Biology* 91 (1): 3–21.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

Mayrose, Itay, Shing H. Zhan, Carl J. Rothfels, Karen Magnuson-Ford, Michael S. Barker, Loren H. Rieseberg, and Sarah P. Otto. 2011. “Recently Formed Polyploid Plants Diversify at Lower Rates.” *Science* 333 (6047): 1257.

Mehta, Ravi, and Rui Juliet Zhu. 2009. “Blue or Red? Exploring the Effect of Color on Cognitive Task Performances.” *Science* 323 (5918): 1226–29.

Metheringham, Carey. 2018. “Within and Between Species Methods to Identify Loci of Resistance to Ash Dieback.”

Miller, Judith, Stephen Engelberg, and William J. Broad. 2001. “U.S. Germ Warfare Research Pushes Treaty Limits.” *The New York Times*, September 4, 2001.

Mindrebo, Jeffrey T., Charisse M. Nartey, Yoshiya Seto, Michael D. Burkart, and Joseph P. Noel. 2016. “Unveiling the Functional Diversity of the Alpha/beta Hydrolase

Superfamily in the Plant Kingdom." *Current Opinion in Structural Biology* 41 (December): 233–46.

Moran, Reid. 1996. "Flora of Guadalupe Island, Mexico." http://agris.fao.org/agris-search/search.do?recordID=US201300054803.

Moustafa, Ahmed M., Vlad Korolev, Surya Saha, Robert Davey, Jerven Bolleman, Robert A. Edwards, Alex Gener, et al. In Review. "Eight Community Recommendations for Genome and Knowledge Graph Infrastructure Elements." Accessed November 24, 2020.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability and Its Applications* 9 (1): 141–42.

Nartey, C. M., J. Aljadeff, I. Fernandez, and C. Laurendon. 2017. "Quantifying the Thermostability Landscape Separating Plant Sesquiterpene Synthase Orthologs Using Maximum Entropy." *bioRxiv*. https://www.biorxiv.org/content/10.1101/172437v1.abstract.

Naumann, Julia, Karsten Salomo, Joshua P. Der, Eric K. Wafula, Jay F. Bolin, Erika Maass, Lena Frenzke, et al. 2013. "Single-Copy Nuclear Genes Place Haustorial Hydnoraceae within Piperales and Reveal a Cretaceous Origin of Multiple Parasitic Angiosperm Lineages." *PloS One* 8 (11): e79204.

Neiman, Maurine, Margaret J. Beaton, Dag O. Hessen, Punidan D. Jeyasingh, and Lawrence J. Weider. 2017. "Endopolyploidy as a Potential Driver of Animal Ecology and Evolution." *Biological Reviews of the Cambridge Philosophical Society* 92 (1): 234–47.

Neugebauer, Tomasz, Eric Bordeleau, Vincent Burrus, and Ryszard Brzezinski. 2015. "DNA Data Visualization (DDV): Software for Generating Web-Based Interfaces Supporting Navigation and Analysis of DNA Sequence Data of Entire Genomes." *PloS One* 10 (12): e0143615.

"NIH Clinical Trial of Investigational Vaccine for COVID-19 Begins." 2020. National Institutes of Health (NIH). March 16, 2020. https://www.nih.gov/news-events/news-releases/nih-clinical-trial-investigational-vaccine-covid-19-begins.

Novikova, Polina Yu, Takashi Tsuchimatsu, Samson Simon, Viktoria Nizhynska, Viktor Voronin, Robin Burns, Olga M. Fedorenko, et al. 2017. "Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis Suecica." *Molecular Biology and Evolution* 34 (4): 957–68.

Oberdorf, Richard, and Tanja Kortemme. 2010. "Complex Topology Rather Than Complex Membership Is a Determinant of Protein Dosage Sensitivity." *Biophysical Journal*. https://doi.org/10.1016/j.bpj.2009.12.085.

O'Donoghue, Seán I., Benedetta Frida Baldi, Susan J. Clark, Aaron E. Darling, James M. Hogan, Sandeep Kaur, Lena Maier-Hein, et al. 2018. "Visualization of Biomedical Data." *Annual Review of Biomedical Data Science* 1 (1): 275–304.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Olm, Matthew R., Cristina N. Butterfield, Alex Copeland, T. Christian Boles, Brian C. Thomas, and Jillian F. Banfield. 2017. "The Source and Evolutionary History of a

Microbial Contaminant Identified Through Soil Metagenomic Analysis." *mBio* 8 (1). https://doi.org/10.1128/mBio.01969-16.

Olofsson, Jill K., Isabel Cantera, Céline Van de Paer, Cynthia Hong-Wa, Loubab Zedane, Luke T. Dunning, Adriana Alberti, Pascal-Antoine Christin, and Guillaume Besnard. 2019. "Phylogenomics Using Low-Depth Whole Genome Sequencing: A Case Study with the Olive Tribe." *Molecular Ecology Resources* 19 (4): 877–92.

"OpenSeadragon." n.d. Accessed March 5, 2018. https://openseadragon.github.io/.

Ottesen, Andrea R., Antonio González Peña, James R. White, James B. Pettengill, Cong Li, Sarah Allard, Steven Rideout, et al. 2013. "Baseline Survey of the Anatomical Microbial Ecology of an Important Food Plant: Solanum Lycopersicum (tomato)." *BMC Microbiology* 13 (May): 114.

Pacchioni, Gianfranco. 2018. *The Overproduction of Truth: Passion, Competition, and Integrity in Modern Science*. Oxford University Press.

Paladin, Lisanna, Mathieu Schaeffer, Pascale Gaudet, Monique Zahn-Zabal, Pierre-André Michel, Damiano Piovesan, Silvio C. E. Tosatto, and Amos Bairoch. 2020. "The Feature-Viewer: A Visualization Tool for Positional Annotations on a Sequence." *Bioinformatics* 36 (10): 3244–45.

Pannell, John R., Darren J. Obbard, and Richard J. A. Buggs. 2004. "Polyploidy and the Sexual System: What Can We Learn from Mercurialis Annua?" *Biological Journal of the Linnean Society. Linnean Society of London* 82 (4): 547–60.

Parker, Joe, Georgia Tsagkogeorga, James A. Cotton, Yuan Liu, Paolo Provero, Elia Stupka, and Stephen J. Rossiter. 2013. "Genome-Wide Signatures of Convergent Evolution in Echolocating Mammals." *Nature* 502 (7470): 228–31.

Paten, Benedict, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. 2017. "Genome Graphs and the Evolution of Genome Inference." *Genome Research* 27 (5): 665–76.

Paterson, Andrew H., Brad A. Chapman, Jessica C. Kissinger, John E. Bowers, Frank A. Feltus, and James C. Estill. 2006. "Many Gene and Domain Families Have Convergent Fates Following Independent Whole-Genome Duplication Events in Arabidopsis, Oryza, Saccharomyces and Tetraodon." *Trends in Genetics: TIG* 22 (11): 597–602.

Pavlichin, Dmitri, and Tsachy Weissman. 2018. "The Desperate Quest for Genomic Compression Algorithms." *IEEE Spectrum*, September 2018. https://spectrum.ieee.org/computing/software/the-desperate-quest-for-genomic-compression-algorithms.

Poland, Therese M., and Deborah G. McCullough. 2006. "Emerald Ash Borer: Invasion of the Urban Forest and the Threat to North America's Ash Resource." *Journal of Forestry* 104 (3): 118–24.

Pook, Torsten, Martin Schlather, Gustavo de Los Campos, Manfred Mayer, Chris Carolin Schoen, and Henner Simianer. 2019. "HaploBlocker: Creation of Subgroup Specific Haplotype Blocks and Libraries." *Genetics*, May. https://doi.org/10.1534/genetics.119.302283.

Pritt, Jacob, Nae-Chyun Chen, and Ben Langmead. 2018. “FORGe: Prioritizing Variants for Graph Genomes.” *Genome Biology* 19 (1): 220.

Procter, James B., G. Mungo Carstairs, Ben Soares, Kira Mourão, T. Charles Ofoegbu, Daniel Barton, Lauren Lui, et al. 2021. “Alignment of Biological Sequences with Jalview.” *Methods in Molecular Biology* 2231: 203–24.

Putnam, Nicholas H., Brendan L. O’Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, et al. 2016. “Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage.” *Genome Research* 26 (3): 342–50.

Rabier, Charles-Elie, Tram Ta, and Cécile Ané. 2014. “Detecting and Locating Whole Genome Duplications on a Phylogeny: A Probabilistic Approach.” *Molecular Biology and Evolution* 31 (3): 750–62.

Raciti, Daniela, Karen Yook, Todd W. Harris, Tim Schedl, and Paul W. Sternberg. 2018. “Micropublication: Incentivizing Community Curation and Placing Unpublished Data into the Public Domain.” *Database: The Journal of Biological Databases and Curation* 2018 (January). https://doi.org/10.1093/database/bay013.

Raina, S. N., A. Parida, K. K. Koul, S. S. Salimath, M. S. Bisht, V. Raja, and T. N. Khoshoo. 1994. “Associated Chromosomal DNA Changes in Polyploids.” *Genome.* 37 (4): 560–64.

Ramsey, Justin, and Douglas W. Schemske. 1998. “Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants.” *Annual Review of Ecology and Systematics* 29 (1): 467–501.

Rasmussen, Matthew D., Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. 2014. “Genome-Wide Inference of Ancestral Recombination Graphs.” *PLoS Genetics* 10 (5): e1004342.

Ravid, Katya, Jun Lu, Jeffrey M. Zimmet, and Matthew R. Jones. 2002. “Roads to Polyploidy: The Megakaryocyte Example.” *Journal of Cellular Physiology* 190 (1): 7–20.

Recht, Benjamin, Christopher Re, Stephen Wright, and Feng Niu. 2011. “Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent.” Advances in Neural Information Processing Systems 24: 693–701.

Renny-Byfield, Simon, Lei Gong, Joseph P. Gallagher, and Jonathan F. Wendel. 2015. “Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution.” *Molecular Biology and Evolution* 32 (4): 1063–71.

Ren, Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, and Ji Qi. 2018. “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms.” *Molecular Plant* 11 (3): 414–28.

Rensing, Stefan a. 2014. “Gene Duplication as a Driver of Plant Morphogenetic Evolution.” *Current Opinion in Plant Biology* 17 (February): 43–48.

Resig, John, and Others. 2006. “JQuery.” Online. http://ajaxexperience.techtarget.com/images/Presentations/Resig_John_jQueryAdvanced.pdf.

Rice, John A. 2006. *Mathematical Statistics and Data Analysis*. Cengage Learning.

Roalson, Eric H., and Wade R. Roberts. 2016. "Distinct Processes Drive Diversification in Different Clades of Gesneriaceae." *Systematic Biology* 65 (4): 662–84.

Robertson, Fiona M., Manu Kumar Gundappa, Fabian Grammes, Torgeir R. Hvidsten, Anthony K. Redmond, Sigbjørn Lien, Samuel A. M. Martin, Peter W. H. Holland, Simen R. Sandve, and Daniel J. Macqueen. 2017. "Lineage-Specific Rediploidization Is a Mechanism to Explain Time-Lags between Genome Duplication and Evolutionary Diversification." *Genome Biology* 18 (1): 111.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26.

Rodgers-Melnick, Eli, Shrinivasrao P. Mane, Palitha Dharmawardhana, Gancho T. Slavov, Oswald R. Crasta, Steven H. Strauss, Amy M. Brunner, and Stephen P. Difazio. 2012. "Contrasting Patterns of Evolution Following Whole Genome versus Tandem Duplication Events in Populus." *Genome Research* 22 (1): 95–105.

Roediger, Henry L. 1980. "The Effectiveness of Four Mnemonics in Ordering Recall." *Journal of Experimental Psychology. Human Learning and Memory* 6 (5): 558–67.

Rombouts, Wannes. 2014. "Veles - Binary Analysis Tool." CodiSec. August 25, 2014. https://codisec.com/veles/.

Rosenbloom, Kate R., Cricket A. Sloan, Venkat S. Malladi, Timothy R. Dreszer, Katrina Learned, Vanessa M. Kirkup, Matthew C. Wong, et al. 2013. "ENCODE Data in the UCSC Genome Browser: Year 5 Update." *Nucleic Acids Research* 41 (Database issue): D56–63.

Rudolph, Jan Daniel, Marjo de Graauw, Bob van de Water, Tamar Geiger, and Roded Sharan. 2016. "Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks." *Cell Systems* 3 (6): 585–93.e3.

Ruprecht, Colin, Rolf Lohaus, Kevin Vanneste, Marek Mutwil, Zoran Nikoloski, Yves Van de Peer, and Staffan Persson. 2017. "Revisiting Ancestral Polyploidy in Plants." *Science Advances* 3 (7): e1603195.

Sagan, Hans. 1994. "Continuous Images of a Line Segment." *Universitext*. https://doi.org/10.1007/978-1-4612-0871-6_6.

Salmon, Armel, Malika L. Ainouche, and Jonathan F. Wendel. 2005. "Genetic and Epigenetic Consequences of Recent Hybridization and Polyploidy in Spartina (Poaceae)." *Molecular Ecology* 14 (4): 1163–75.

Sánchez, J., and I. Lopez-Villasenor. 2006. "A Simple Model to Explain Three-base Periodicity in Coding DNA." *FEBS Letters*. https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2006.10.056.

Sanderson, Michael J. 2003. "r8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock." *Bioinformatics* 19 (2): 301–2.

Sarkar, Biplab, R. P. Dip, Anusheel Munshi, Bhaswar Ghosh, Tharmarnadar Ganesh, Dabr Arjunan Manikandan, R. P. Dip, et al. 2020. “Dynamics of the COVID -19 Related Publications.” Apollo Gleneagles Hospitals, Kolkata, India . https://doi.org/10.1101/2020.08.05.237313.

Satyanarayan, Arvind, Ryan Russell, Jane Hoffswell, and Jeffrey Heer. 2016. “Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization.” *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 659–68.

Schatz, Michael C., Adam M. Phillippy, Daniel D. Sommer, Arthur L. Delcher, Daniela Puiu, Giuseppe Narzisi, Steven L. Salzberg, and Mihai Pop. 2013. “Hawkeye and AMOS: Visualizing and Assessing the Quality of Genome Assemblies.” *Briefings in Bioinformatics* 14 (2): 213–24.

Schnable, James C., Nathan M. Springer, and Michael Freeling. 2011. “Differentiation of the Maize Subgenomes by Genome Dominance and Both Ancient and Ongoing Gene Loss.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (10): 4069–74.

Schneider, Harald, Eric Schuettpelz, Kathleen M. Pryer, Raymond Cranfill, Susana Magallón, and Richard Lupia. 2004. “Ferns Diversified in the Shadow of Angiosperms.” *Nature* 428 (6982): 553–57.

Schranz, M. Eric, Setareh Mohammadin, and Patrick P. Edger. 2012. “Ancient Whole Genome Duplications, Novelty and Diversification: The WGD Radiation Lag-Time Model.” Current Opinion in Plant Biology 15 (2): 147–53.

Schulte, Eric, Zachary P. Fry, Ethan Fast, Westley Weimer, and Stephanie Forrest. 2014. “Software Mutational Robustness.” *Genetic Programming and Evolvable Machines* 15 (3): 281–312.

Seaman, Josiah D., and John C. Sanford. 2009. “Skittle: A 2-Dimensional Genome Visualization Tool.” *BMC Bioinformatics* 10 (December): 452.

Seaman, Josiah, and Richard J. A. Buggs. 2020. “FluentDNA: Nucleotide Visualization of Whole Genomes, Annotations, and Alignments.” *Frontiers in Genetics* 11 (April): 292.

Ségurel, Laure, Emma E. Thompson, Timothée Flutre, Jessica Lovstad, Aarti Venkat, Susan W. Margulis, Jill Moyse, et al. 2012. “The ABO Blood Group Is a Trans-Species Polymorphism in Primates.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (45): 18493–98.

Selmecki, Anna M., Yosef E. Maruvka, Phillip A. Richmond, Marie Guillet, Noam Shoresh, Amber L. Sorenson, Subhajyoti De, et al. 2015. “Polyploidy Can Drive Rapid Adaptation in Yeast.” *Nature* 519 (7543): 349–52.

Serov, O. L., B. Chowdhary, J. E. Womack, and J. A. M. Graves. 2005. “Comparative Gene Mapping, Chromosome Painting and the Reconstruction of the Ancestral Mammalian Karyotype.” In *Mammalian Genomics*, edited by A. Ruvinsky and J. A. Marshall Graves, 349–92. Wallingford: CABI.

Shaked, Hezi, Khalil Kashkush, Hakan Ozkan, Moshe Feldman, and Avraham A. Levy. 2001. “Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible

Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat.” *The Plant Cell*. https://doi.org/10.1105/tpc.010083.

Sicard, Anne, Yannis Michalakis, Serafín Gutiérrez, and Stéphane Blanc. 2016. “The Strange Lifestyle of Multipartite Viruses.” *PLoS Pathogens* 12 (11): e1005819.

Smit, A. F. A., R. Hubley, and P. Green. 2015. “RepeatMasker Open-4.0. 2013--2015.”

Smith, Stephen, Joseph W. Brown, and Joseph F. Walker. 2017. “So Many Genes, so Little Time: Comments on Divergence-Time Estimation in the Genomic Era.” *bioRxiv*. https://doi.org/10.1101/114975.

Solís-Lemus, Claudia, Paul Bastide, and Cécile Ané. 2017. “PhyloNetworks: A Package for Phylogenetic Networks.” *Molecular Biology and Evolution* 34 (12): 3292–98.

Sollars, Elizabeth. 2017. “The Genome and Epigenome of the European Ash Tree (Fraxinus Excelsior).” Queen Mary University of London. https://qmro.qmul.ac.uk/xmlui/handle/123456789/25977.

Sollars, Elizabeth S. A., Andrea L. Harper, Laura J. Kelly, Christine M. Sambles, Ricardo H. Ramirez-Gonzalez, David Swarbreck, Gemy Kaithakottil, et al. 2017. “Genome Sequence and Genetic Diversity of European Ash Trees.” *Nature* 541 (7636): 212–16.

Spoelhof, Jonathan P., Michael Chester, Roseana Rodriguez, Blake Geraci, Kweon Heo, Evgeny Mavrodiev, Pamela S. Soltis, and Douglas E. Soltis. 2017. “Karyotypic Variation and Pollen Stainability in Resynthesized Allopolyploids Tragopogon Miscellus and T. Mirus.” *American Journal of Botany* 104 (10): 1484–92.

Stahl, Bjorn. n.d. *Senseye - Dynamic Visual Debugging / Reverse Engineering Toolsuite*. Github. Accessed June 7, 2019. https://github.com/letoram/senseye/tree/1bab4b9c60ad43302e460a24de14a0ac136bea7f.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Stanke, Mario, and Burkhard Morgenstern. 2005. “AUGUSTUS: A Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints.” *Nucleic Acids Research* 33 (Web Server issue): W465–67.

Storchova, Zuzana, and David Pellman. 2004. “From Polyploidy to Aneuploidy, Genome Instability and Cancer.” *Nature Reviews. Molecular Cell Biology* 5 (1): 45–54.

Sun, Honghe, Shan Wu, Guoyu Zhang, Chen Jiao, Shaogui Guo, Yi Ren, Jie Zhang, et al. 2017. “Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes.” *Molecular Plant*, September. https://doi.org/10.1016/j.molp.2017.09.003.

Sussillo, David, Anshul Kundaje, and Dimitris Anastassiou. 2004. “Spectrogram Analysis of Genomes.” *EURASIP Journal on Advances in Signal Processing* 2004 (1): 790248.

Swarbreck, David, Christopher Wilks, Philippe Lamesch, Tanya Z. Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, et al. 2008. “The Arabidopsis

Information Resource (TAIR): Gene Structure and Function Annotation." *Nucleic Acids Research* 36 (Database issue): D1009–14.

Swigoňová, Z., J. Lai, J. Ma, and W. Ramakrishna. 2004. "Close Split of Sorghum and Maize Genome Progenitors." *Genome*. https://genome.cshlp.org/content/14/10a/1916.short.

Tamura, Koichiro, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. 2011. "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." *Molecular Biology and Evolution* 28 (10): 2731–39.

Tank, David C., Jonathan M. Eastman, Matthew W. Pennell, Pamela S. Soltis, Douglas E. Soltis, Cody E. Hinchliff, Joseph W. Brown, Emily B. Sessa, and Luke J. Harmon. 2015. "Nested Radiations and the Pulse of Angiosperm Diversification: Increased Diversification Rates Often Follow Whole Genome Duplications." The New Phytologist 207 (2): 454–67.

Ta, Tram, Charles-Elie Rabier, Cecile Ane, and Maintainer Tram Ta. 2013. "Package 'WGDgc.'" http://pages.stat.wisc.edu/~ane/wgd/WGDgc-manual_1.2.pdf.

Teichmann, Sarah Amalia, and Reiner Albert Veitia. 2004. "Genes Encoding Subunits of Stable Complexes Are Clustered on the Yeast Chromosomes: An Interpretation from a Dosage Balance Perspective." *Genetics* 167 (4): 2121–25.

Tekaia, Fredj. 2016. "Inferring Orthologs: Open Questions and Perspectives." *Genomics Insights* 9 (February): 17–28.

Telentia, Amalio, Levi C. T. Piercea, William H. Biggsa, Julia di Iulioa, Emily H. M. Wonga, Martin M. Fabania, Ewen F. Kirknessa, et al. 2016. "Deep Sequencing of 10,000 Human Genomes." *Proceedings of the National Academy of Sciences of the United States of America*, August. https://www.pnas.org/content/pnas/early/2016/10/03/1613365113.full.pdf.

Teufel, Ashley I., Mackenzie M. Johnson, Jon M. Laurent, Aashiq H. Kachroo, Edward M. Marcotte, and Claus O. Wilke. 2018. "The Many Nuanced Evolutionary Consequences of Duplicated Genes." *Molecular Biology and Evolution*, November. https://doi.org/10.1093/molbev/msy216.

The Computational Pan-Genomics Consortium. 2018. "Computational Pan-Genomics: Status, Promises and Challenges." *Briefings in Bioinformatics* 19 (1): 118–35.

Theobald, Douglas L. 2010. "A Formal Test of the Theory of Universal Common Ancestry." *Nature* 465 (7295): 219–22.

Thomas, Brian C., Brent Pedersen, and Michael Freeling. 2006. "Following Tetraploidy in an Arabidopsis Ancestor, Genes Were Removed Preferentially from One Homeolog Leaving Clusters Enriched in Dose-Sensitive Genes." *Genome Research* 16 (7): 934–46.

Thomasset, M., J. F. Fernández-Manjarrés, G. C. Douglas, N. Frascaria-Lacoste, and T. R. Hodkinson. 2011. "Hybridisation, Introgression and Climate Change: A Case Study of the Tree Genus Fraxinus (Oleaceae).” *Climate Change, Ecology and Systematics*, 320–42.

Thornton, J. W., and R. DeSalle. 2000. “Gene Family Evolution and Homology: Genomics Meets Phylogenetics.” Annual Review of Genomics and Human Genetics 1: 41–73.

Tiley, George P., Cécile Ané, and J. Gordon Burleigh. 2016. “Evaluating and Characterizing Ancient Whole-Genome Duplications in Plants with Gene Count Data.” *Genome Biology and Evolution* 8 (4): 1023–37.

“TimeTree :: The Timescale of Life.” n.d. Accessed April 23, 2019. http://timetree.org.

Tomato, The, and Genome Consortium. 2012. “The Tomato Genome Sequence Provides Insights into Fleshy Fruit Evolution.” *Nature* 485 (7400): 635–41.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols* 7 (3): 562–78.

Tuğrul, Murat, Tiago Paixão, Nicholas H. Barton, and Gašper Tkačik. 2015. “Dynamics of Transcription Factor Binding Site Evolution.” *PLoS Genetics* 11 (11): e1005639.

Unver, Turgay, Zhangyan Wu, Lieven Sterck, Mine Turktas, Rolf Lohaus, Zhen Li, Ming Yang, et al. 2017. “Genome of Wild Olive and the Evolution of Oil Biosynthesis.” *Proceedings of the National Academy of Sciences*, October. https://doi.org/10.1073/pnas.1708621114.

Van Dongen, Stijn Marinus. 2000. “Graph Clustering by Flow Simulation.” https://dspace.library.uu.nl/handle/1874/848.

Vanneste, Kevin, Yves Van de Peer, and Steven Maere. 2013. “Inference of Genome Duplications from Age Distributions Revisited.” *Molecular Biology and Evolution* 30 (1): 177–90.

Vicient, Carlos M., and Josep M. Casacuberta. 2017. “Impact of Transposable Elements on Polyploid Plant Genomes.” *Annals of Botany* 120 (2): 195–207.

Vitte, Clémentine, Olivier Panaud, and Hadi Quesneville. 2007. “LTR Retrotransposons in Rice (Oryza Sativa, L.): Recent Burst Amplifications Followed by Rapid DNA Loss.” *BMC Genomics* 8 (July): 218.

Wallander, Eva. 2012. “Systematics and Floral Evolution in Fraxinus (Oleaceae).” *Belgische Dendrologie Belge* 2012: 39–58.

Wang, Haifeng, Chunce Guo, Hong Ma, and Ji Qi. 2019. “Reply to Zwaenepoel et Al.: Meeting the Challenges of Detecting Polyploidy Events from Transcriptomic Data.” *Molecular Plant*.

Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. “Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench.” *Bioinformatics* 25 (9): 1189–91.

Watson, Geoffrey S. 1964. “Smooth Regression Analysis.” *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26 (4): 359–72.

Weisstein, Eric W. 2021. “Planar Graph.” MathWorld. Accessed June 10, 2021. https://mathworld.wolfram.com/PlanarGraph.html.

Wendel, Jonathan F., Damon Lisch, Guanjing Hu, and Annaliese S. Mason. 2018. “The Long and Short of Doubling down: Polyploidy, Epigenetics, and the Temporal Dynamics of Genome Fractionation.” *Current Opinion in Genetics & Development* 49 (April): 1–7.

Werth, Charles R., and Michael D. Windham. 1991. “A Model for Divergent, Allopatric Speciation of Polyploid Pteridophytes Resulting from Silencing of Duplicate-Gene Expression.” *The American Naturalist* 137 (4): 515–26.

Wick, Ryan R., Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. 2015. “Bandage: Interactive Visualization of *de Novo* Genome Assemblies.” *Bioinformatics* 31 (20): 3350–52.

Wilcoxon, Frank, S. K. Katti, and Roberta A. Wilcox. 1970. “Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test.” *Selected Tables in Mathematical Statistics* 1: 171–259.

Wild, Chris, and George Seber. 1999. “Wilcoxon Rank Sum.” In *CHANCE ENCOUNTERS: A First Course in Data Analysis and Inference*. John Wiley & Sons, New York.

Wininger, Kerry, and Nathan Rank. 2017. “Evolutionary Dynamics of Interactions between Plants and Their Enemies: Comparison of Herbivorous Insects and Pathogens.” Annals of the New York Academy of Sciences 1408 (1): 46–60.

Wolf, Yuri I., and Eugene V. Koonin. 2013. “Genome Reduction as the Dominant Mode of Evolution.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35 (9): 829–37.

Woodhouse, Margaret R., Feng Cheng, J. Chris Pires, Damon Lisch, Michael Freeling, and Xiaowu Wang. 2014. “Origin, Inheritance, and Gene Regulatory Consequences of Genome Dominance in Polyploids.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (14): 5283–88.

Woodhouse, Margaret R., James C. Schnable, Brent S. Pedersen, Eric Lyons, Damon Lisch, Shabarinath Subramaniam, and Michael Freeling. 2010. “Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs.” *PLoS Biology* 8 (6): e1000409.

“World Checklist of Selected Plant Families: Royal Botanic Gardens, Kew.” n.d. Accessed June 24, 2020. https://wcsp.science.kew.org/namedetail.do?name_id=370564.

Wright, B., J. Payne, M. Steckman, and S. Stevson. 2009. “Palantir: A Visualization Platform for Real-World Analysis.” In *2009 IEEE Symposium on Visual Analytics Science and Technology*, 249–50. ieeexplore.ieee.org.

Yachdav, Guy, Tatyana Goldberg, Sebastian Wilzbach, David Dao, Iris Shih, Saket Choudhary, Steve Crouch, et al. 2015. "Anatomy of BioJS, an Open Source Community for the Life Sciences." *eLife* 4 (July). https://doi.org/10.7554/eLife.07009.

Yang, Xiaofei, Wan-Ping Lee, Kai Ye, and Charles Lee. 2019. "One Reference Genome Is Not Enough." *Genome Biology* 20 (1): 104.

Yang, Yang, Wu Youyou, and Brian Uzzi. 2020. "Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence." *Proceedings of the National Academy of Sciences of the United States of America* 117 (20): 10762–68.

Yassin, Amir, Emily K. Delaney, Adam J. Reddiex, Thaddeus D. Seher, Héloïse Bastide, Nicholas C. Appleton, Justin B. Lack, et al. 2016. "The pdm3 Locus Is a Hotspot for Recurrent Evolution of Female-Limited Color Dimorphism in Drosophila." *Current Biology: CB* 26 (18): 2412–22.

Yokoyama, Toshiyuki T., Simon Heumos, Josiah Seaman, Dmytro Trybushnyi, Torsten Pook, Andrea Guarracino, Erik Garrison, and Jerven T. Bolleman. 2020. "Semantic Variation Graphs: Ontologies for Pangenome Graphs." In . ISCB.

Yokoyama, Toshiyuki T., and Masahiro Kasahara. 2020. "Visualization Tools for Human Structural Variations Identified by Whole-Genome Sequencing." *Journal of Human Genetics* 65 (1): 49–60.

Yokoyama, Toshiyuki T., Yoshitaka Sakamoto, Masahide Seki, Yutaka Suzuki, and Masahiro Kasahara. 2019. "MoMI-G: Modular Multi-Scale Integrated Genome Graph Browser." *bioRxiv*. https://doi.org/10.1101/540120.

Yoo, M-J, E. Szadkowski, and J. F. Wendel. 2013. "Homoeolog Expression Bias and Expression Level Dominance in Allopolyploid Cotton." *Heredity* 110 (2): 171–80.

Yu, Jingyin, Sadia Tehrim, Linhai Wang, Komivi Dossa, Xiurong Zhang, Tao Ke, and Boshou Liao. 2017. "Evolutionary History and Functional Divergence of the Cytochrome P450 Gene Superfamily between Arabidopsis Thaliana and Brassica Species Uncover Effects of Whole Genome and Tandem Duplications." *BMC Genomics* 18 (1): 733.

Zedane, L., C. Hong-Wa, J. Murienne, C. Jeziorski, B. G. Baldwin, and G. Besnard. 2016. "Museomics Illuminate the History of an Extinct, Paleoendemic Plant Lineage (Hesperelaea, Oleaceae) Known from an 1875 Collection from Guadalupe Island, Mexico." *Biological Journal of the Linnean Society. Linnean Society of London* 117 (1): 44–57.

Zedane, Louba. 2016. "Phylogenetic Hypothesis of the Oleeae Tribe (Oleaceae) Diversification and Molecular Evolution Patterns in Plastid and Nuclear Ribosomal DNA."

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. "Ensembl 2018." *Nucleic Acids Research* 46 (D1): D754–61.

Zohren, Jasmin. 2016. "Introgression in Betula Species of Different Ploidy Levels and the Analysis of the Betula Nana Genome." Edited by Nichols Buggs. Queen Mary University.

https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/24564/Zohren_J_PhD_final_080317.pdf?sequence=1.

Zwaenepoel, Arthur, Zhen Li, Rolf Lohaus, and Yves Van de Peer. 2019. “Finding Evidence for Whole Genome Duplications: A Reappraisal.” *Molecular Plant*.

Supplemental

S1: FluentDNA Museum Display

Supplement to Chapter 3

These instructions step through everything required to setup a FluentDNA museum display. The original image file generated by FluentDNA was edited in Photoshop to add the interpretation text and legend and then reprocessed by DeepZoom to ensure the monitor and poster match. Finally, since gene labels are not rendered on the poster, the color palette was modified so that genes were saturated colors and intergenic regions are lightened to draw attention to the genes.

The *Arabidopsis thaliana* genome required special processing for this exhibit. Chromosomes were broken into a series of genes with the FASTA entry name carrying the functional annotation for that gene. This was used as a specialized annotation retrieval since FluentDNA shows the name of each FASTA entry under the mouse. GO Slim was selected to minimize the length of technical jargon in the annotation. Centromeres, largely lacking genes, are easily visible in the final display.

The physical setup was a 2-meter-tall display case with a 1.5-meter poster behind plexiglass. Behind the poster was a 72cm x 124cm touch sensitive film manufactured by Displax. Since touch sensors use changes in electrostatics, they can sense touch through several layers of material. The Displax touch sensor sends touch events to a computer running FluentDNA with a copy of the exact same poster. The monitor positioned inside the poster shows a magnified version of the poster (Figure Supplemental-1.1).

First, sequence and functional annotation were downloaded from TAIR10 (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGO_and_PO_Annotations%2FGene_Ontology_Annotations) (accessed March 2019). GO Slim was selected to minimize the length of technical jargon in the annotation. The annotation was processed through Functional_Annotation_Arabidopsis.ipynb to remove all genes that did not have a known function. The remaining gene annotations had their numeric names replaced with their GO Slim function and placed in a new GFF file.

The function **write_functional_gene_contigs_from_chromosome()** reads nucleotide sequence and gene annotation together and outputs consecutive contigs with the fasta header as the function followed by the complete sequence inside the gene annotation. Regions with no annotation are all labeled “between genes - not yet understood”. Each FASTA file then contains one chromosome with a FASTA entry for each gene and intergenic region. There are some obvious downsides to this approach, primarily that overlapping genes compete for label space. However, this was a quick and effective way to get annotation mouseover in FluentDNA and more than accurate enough for a museum display.

Using the modified FASTA file, FluentDNA can be used to render a poster of one chromosome. In order to layout multiple chromosomes using the Ideogram layout, a modified version of the software must be used. The “wakehurst” branch on GitHub contains some special exceptions that must be modified for each project. The origin coordinates of each of the chromosomes are calculated and entered by hand to whatever poster layout the developer prefers. While it appears the poster uses HighlightedAnnotations.py it actually uses the FASTA headers themselves. Whenever a contig called "between genes - not yet understood" is rendered, the color palette is switched to a lighter shade which mimics highlighting. This allows attention to be drawn to the genes rendered in saturated dark colors while the intergenic regions are whitened out. Gene labels are not rendered since the monitor already serves that function.

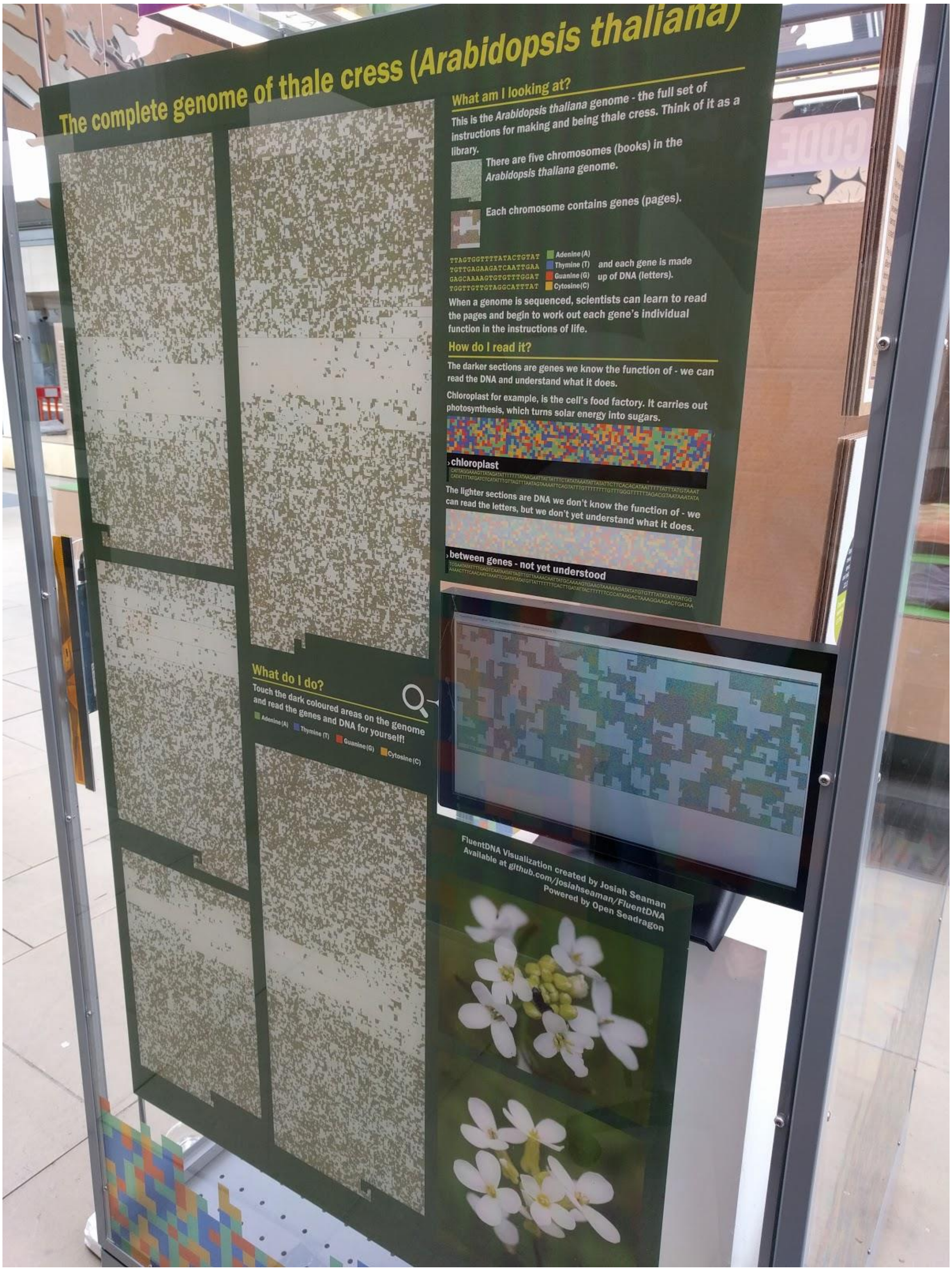
After the initial render, the now 150MB PNG file was brought into Photoshop where the rest of the poster design was carried out. Poster layout design was done by Samantha Seaman. Final color choice and fonts were done by Adomas Mockus at Rockbrook Engineering. The final poster design was re-rendered back into FluentDNA DeepZoom stack so that the monitor and poster matched perfectly.

When a user touches a point on the poster, Displax Connect interprets that into a x,y coordinate on the touch surface. This is sent through Windows HID service to a screen. The display screen is purely informative and doesn't match the User Interface represented by a poster. Therefore, a second logical screen was required to act as a digital counterpart to the poster. On HDMI and newer displays Windows logical screens always map to a physical device. However, Windows contains legacy support for VGA monitors that would not report their resolution. Declaring a monitor as a VGA monitor in Windows allows one to add another monitor in display settings and manually set the resolution without Windows requiring feedback from a physical monitor.

The small desktop used for the display did not have an onboard VGA port, so it was necessary to plug in a USB to VGA adapter. Bridging VGA pins 1 (Red) and 6 (Red Ground) with an insulated cable caused the computer to recognize it as a connected monitor of unknown resolution. The resolution was then set to exactly match the resolution of the Displax touch surface and place in Portrait orientation.

The faked logical monitor is then used to translate Displax touch coordinates into HTML coordinates for a portion of the webpage that generates JavaScript commands but is never seen by the user. The OpenSeadragon Navigator element is placed on this second screen since it never changes position and allows the user to navigate globally. In each display, the positioning of the navigator on the second screen will require manual adjustment to match the poster position and scale. This process was helped greatly by using TeamViewer on a laptop to see the contents of the virtual screen and Chrome's ability to live edit JavaScript source files.

Finally, in order to get the sequence to display, a mouseover event must be simulated in the middle of the screen. A quarter-second loop which creates a fake mouse hover event and asserts the zoom level was enough to trigger FluentDNA mouseover functionality. The whole setup is shut down at night and boots up every morning using a BIOS rule that checks for power. The program LaunchLater (https://jeffcox111.github.io/LaunchLater/) is used to launch the FluentDNA server, then a browser 5 minutes later. The Chrome browser is run in Kiosk mode which removes all the usual browser decorations and pointed directly to the locally served webpage. Chrome windows will preserve their size after reboots and can span multiple screens as long as they are not Maximized. Using this setup one browser window handles both the display and hidden UI elements.



The complete genome of thale cress (*Arabidopsis thaliana*)

What am I looking at?

This is the *Arabidopsis thaliana* genome - the full set of instructions for making and being thale cress. Think of it as a library.

There are five chromosomes (books) in the *Arabidopsis thaliana* genome.

Each chromosome contains genes (pages).

```
TTAGTGGTTTTATACTGTAT
TGTTGAGAAGATCAATTGAA
GAGCAAAAGTGTGTTTGGAT
TGGTTGTTGTAGGCATTTAT
```

🟩 Adenine (A)
🟦 Thymine (T)
🟥 Guanine (G)
🟨 Cytosine (C)

and each gene is made up of DNA (letters).

When a genome is sequenced, scientists can learn to read the pages and begin to work out each gene's individual function in the instructions of life.

How do I read it?

The darker sections are genes we know the function of - we can read the DNA and understand what it does.

Chloroplast for example, is the cell's food factory. It carries out photosynthesis, which turns solar energy into sugars.

›chloroplast

The lighter sections are DNA we don't know the function of - we can read the letters, but we don't yet understand what it does.

›between genes - not yet understood

What do I do?

Touch the dark coloured areas on the genome and read the genes and DNA for yourself!

🟩 Adenine (A) 🟦 Thymine (T) 🟥 Guanine (G) 🟨 Cytosine (C)

FluentDNA Visualization created by Josiah Seaman
Available at *github.com/josiahseaman/FluentDNA*
Powered by Open Seadragon

Figure Supplemental-1.1: Final poster displayed in the Millennium Seed Bank. The left two thirds of the poster are backed by Displax touch sensitive film (visible as an orange strip on the far left) for user input. On the right, an embedded computer monitor running FluentDNA with a digital copy of the poster. Touching the poster causes the monitor to zoom in on the corresponding place in the genome and display the sequence and function of the gene that was touched.

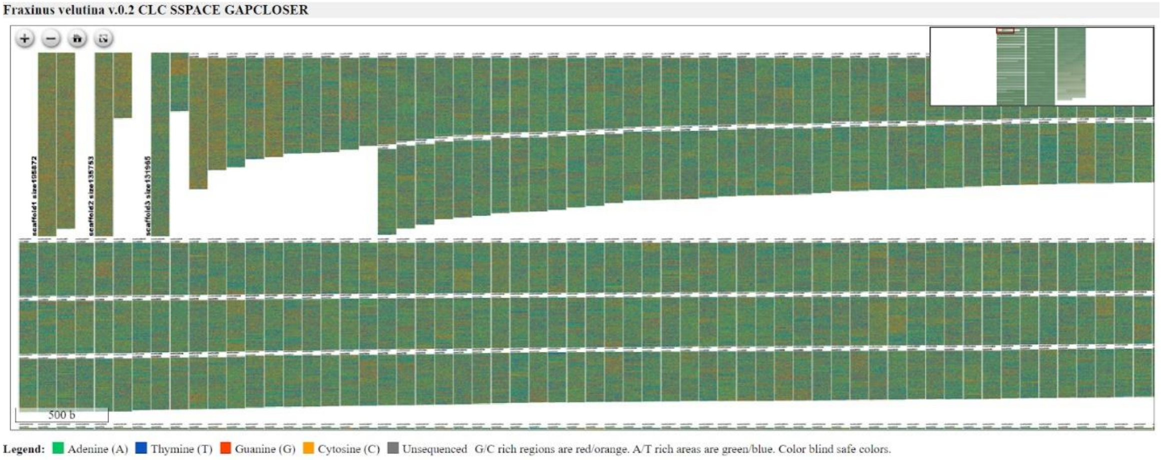
S2: *Fraxinus* Assembly Inspection Notes. Supplement to Chapter 2

This section includes the original research notebook from the inspection of *Fraxinus* assemblies using FluentDNA. This can be used as an example of how FluentDNA can be used in a practical way in a lab setting. Entries include screenshots with visible GC patterns. Clipped sequences were used as BLAST queries to check for matches in other species which would often be matches to mitochondria, plastids, or various contaminant sources. Specific scaffold names are mentioned taken from both the visual rendered names as well as the mouse over information.

F. mandshurica

Scaffold3&4 are the *Fraxinus* mitochondrion. There is a sizable collection of scaffolds <20 Kbp with a high GC content. These most likely represent another organism. At 5 Kbp sizes, it approaches ⅓ of all scaffolds.

F. velutina



Fraxinus velutina v.0.2 CLC SSPACE GAPCLOSER

Legend: 🟩 Adenine (A) 🟦 Thymine (T) 🟥 Guanine (G) 🟧 Cytosine (C) ⬛ Unsequenced G/C rich regions are red/orange. A/T rich areas are green/blue. Color blind safe colors.

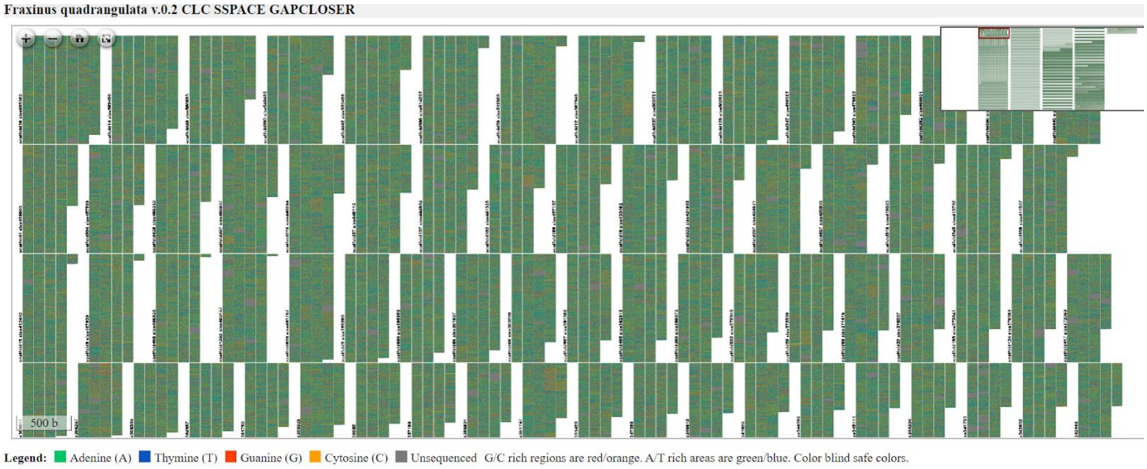
The two largest scaffolds are more GC rich than the rest of the genome and likely represent organellar genomes. Scaffold248 is an extreme outlier in GC, possibly representing contamination. The assembly lacks N's or large scaffolds because no long mate pair libraries were available.

scaffold1|size195872: (101 - 400) (BLAST: Mitochondrion)
CGTAATAGTGAACTCTTTCACAAGAGAGATAGAAGAAAGAAGATCTGTCTCGATTCTAT
CTCTATCTTTCGACATCTCATCTCTATAATACGATAATAGTGGGTGGAGAATCCTTGTGT
ATTCTACTGGTATAGAATCTTTGTTTATTACTAGGAAGGCGGGCTACTTCCGTCTAGCG
GTCATGGGAAAGCCAAAACTTATATATAATAAGTCAATACTGGGTCGGTCGAGACTCTT
TCTTAGTGAAGTGGGAAGACAGCACCGAATCAGACGGGCACAGAAGAAGAAGTGGTTT
CATCC

scaffold2|size135753: (301 - 600) (BLAST: Mitochondrion)
TTAGGATCGTAATAGCTCATACGAAAAATCTGCATAGATAGCGTTGGCGCTGTCAAGCA
ATCCTCGATATCCACTTTGTACTGAGTCCAAAAAATCTCTTGAGCCTTCTGGAAAACACC
GCCTAACAGACGAATATCCTGTTTCAGATAATCCAACAATTCTTCCCTATTATTCAGAAG
ATTCGACACTCCAACTTCGTCATGTTGGATAGAGCCTTTAACACCTAATTGAGGACATA
AAGCCTTAGCCAATGTTGCCAGACTACTACTGAGTAGAGTGTATGAATCTCTTATACGG
AAG

scaffold248|size21911: (1,201 - 1,500) (Delftia acidovorans)
GGGCAGGCTCGCTGCCATGCTGCAGCACCTGCAGGCACTGCAGCATGCCGGGCCCGCC
GTGGGTCTCGTACTTGTTGGCGCTGGCAACGCCGCGCACCTGGCAGAAGTCTTCCTGG
GGCAGACGCGCAATCCAGCCCTGGGCCGTCCAGGCGCGGTCAAAGCGCTCCACGCACA
GCACCTGCTGCCCGCCAAAGGTCTCGATGCTGCTGTCGGCCACGGCAAAGCCCAGCTCG
GCCAGCAGCCGGGCGCACAGCCACTCATTGGCCACTGACTCGCTCAGGTCGTAGTGCC
AGTGCGGCA

F. quadrangulata


Fraxinus quadrangulata v.0.2 CLC SSPACE GAPCLOSER
500 b
Legend: 🟩 Adenine (A) 🟦 Thymine (T) 🟥 Guanine (G) 🟧 Cytosine (C) ⬛ Unsequenced G/C rich regions are red/orange. A/T rich areas are green/blue. Color blind safe colors.

Larger scaffolds, containing small N gaps from long mate pair libraries (LMP) correctly used. Half the genome is covered by scaffolds that are larger than the 3rd largest scaffold in *F. velutina*. Scaffold7 is a GC outlier, possibly organellar:

scaffold7|size243779: (801 - 1,100) (BLAST: Mitochondrion)
CACTTATAGTCCAGTGGCAAGATCAATGTCTTTTATTGGCTTCTCTGTCCCTGGAACAAT
ATATCCTCTTTTATTTCTTTGTTTTTTTCAAATAAAGACCTCTTTCTGTCTTCTCGGGCAG
TGGATTTGCCAACTCTATTGCTCTTATATCGGTCTGAATTCCTCTTCATCTTGCTTTTCG
CTTTTGGTTTGTGGATTCTATTTCCCCCTGTGCTTCCACTTCCCCTGCCTCGAGATCGTG
CCTTTTTTTTACTCGACTCGCTCCCTGCTCAAAAAAGTCGATCTTATTCCGCAACTCGG

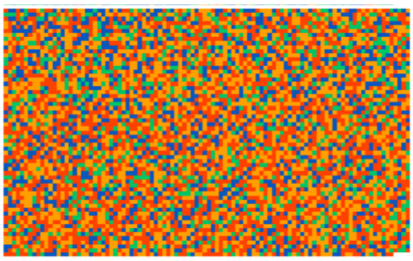
Scaffolds are large enough that multiple genomic elements with different kmer compositions are visible. For example, a low complexity GC rich region:

scaffold3395|size178564: (57,501 - 57,800) (BLAST: No matches, 14% match to Olea europaea var. sylvestris pentatricopeptide repeat-containing protein At4g19440, chloroplastic-like)
CGGAATCGTCTGCATAGTGAGCTGGTCAGGCGCGTCTGCCGGAAAAGTGGTCGGATCG
TGGTGAATCTGGGTTTGGAACCCAGATTGCTCAACTGCTTTCCGCCAACTTCCATCCTT
CCTGCCGTCGCCTGAAGTCCCCACCTCCACCCAGAGGTTCTCACCTGGGATTTCTGCTC
CGGTCTGAGCGGCGGTGCCAGGGTTTCCACCTAAGGCCGCCGCTTGTGAGTGTGGGTG
TTGGCCGGAATGGCCATAGGGGTGTGGCGTGAGGCCGGTAGAGGTGGGTTCGCCGGT
CGATACCTC

Scaffold1602 appears to be a minor outlier. Mouseover not available.

A large number of 10 - 20 Kbp scaffolds with a large N gap indicate LMP that did not chain scaffold to larger regions.

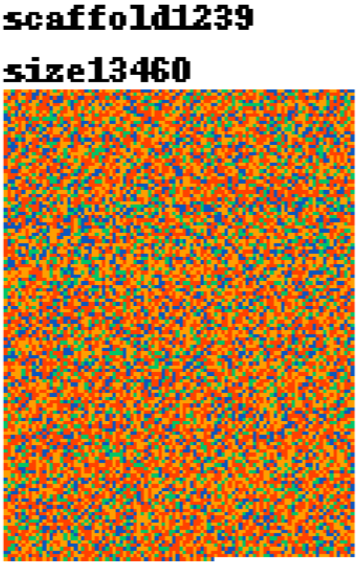
There are 6 scaffolds <10 Kbp that show extreme GC indicative of another organism. It could be an unremoved phiX Illumina control sequence. Mouseover unavailable.



Sequence from Image: GTTGTAGGCAGGTCTGAGCCTTATTCTCTGGACAC matches *Delftia acidovorans* soil bacteria (Olm et al. 2017).

F. paxiana

GT rich region on scaffold1147 matches with low significance to Carp genome.



CGTCAGCGCGGCGCTGCAGCTCGGCGCAGCAGATC matches *Delftia acidovorans* soil bacteria (Olm et al. 2017).

Processed around the same time. Were these collected and extracted in diverse geographic locations? Western arboretum. BlobTools.

Run phiX.

F. sieboldiana

“sieboldiana 10969.fa” doesn’t match anything in BLAST nr database.
CCTGCCTCGCCAAGTAAGTATGAGGGGCAACAATTAAGCCAATTCTAGATGCTTAAATT
TACTCCAATTGGCAGTCACCCTGCTGGCCATGGGATACATAGACCCTTTCACCGACCCC
CCGCACGGGCTACCGTGCCCCATGCAAAAGCCCACCCAATCCTTCGTATCCGTACCCTT
GGATACCCTGGTCACAGACCCTGAAAACGCCCGCTTACACCCAGATGCTAACTTGGATG
CCATCACCGGCAGCCTAGAGCTATTTGGCCAGGTTGAGCCCTTGGTGGTGCAGCAAGG
CACAC

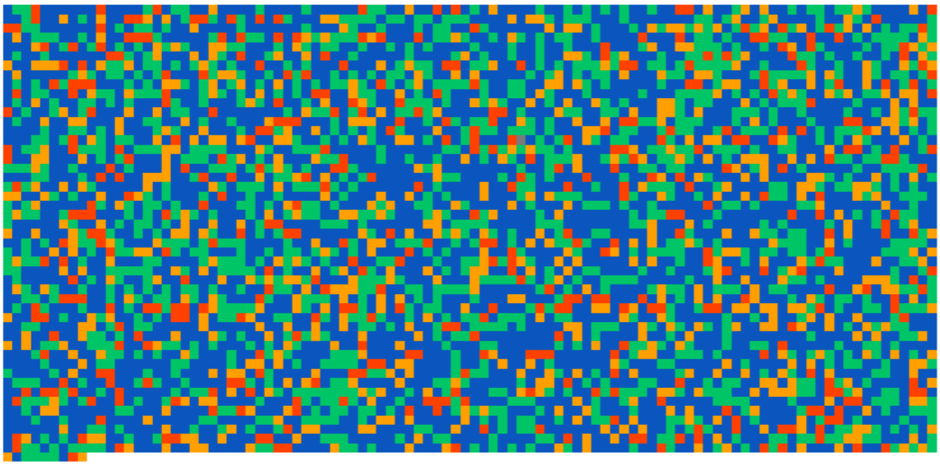
F. sieboldiana scaffold14 belongs to a relative of *Cryptococcus neoformans*, a fungal aerobe that lives in plants and animals (Buchanan and Murphy 1998).

Sacffold8 appears to be a sequence from *Aspergillus*, a genus of airborne molds.

F. angustifolia

Scaffold1 is the chloroplast. Scaffold58 and many fragments belong to the same *Delftia acidovorans* bacteria. The question is, was it lab contamination, or was this a prominent soil bacteria when they were all gathered from the same experimental forest?

Scaffold2-5 are mitochondria.



T-rich scaffolds

F. apertisquamifera

Scaffold 125 is mitochondria.

F. caroliniana

There’s a medium GC-rich element that appears as Scaffold 362 and also many small contigs. It doesn’t appear to match well to any sequence. Out of 300bp queries, 60-70 bp segments map to diverse mitochondria at ~80% identity. Diverse species include

Hannaella oryzae (fungi), *Ustilago triodiae* (fungi teliospore), *Vulgatibacter incomptus* (proteobacteria), *Rhodotorula glutinis* (pink yeast), and *Oerskovia* (gram-positive bacteria).

F. dipetala

scaffold54|size37094: (17,801 - 18,100)
CGTAACTAGATAGACTATGGGTAGTTCAGATGCGCTTAAAGTCTTATATAACGCATGCC
GATGTTGCCACCGACAAACGGTTGACTTTGACTGTGAAAAGCTCCAAATAGGCACACGG
TAGTTTAGAGTAGAAGGAGCTCGAAGTTGGAATTCAATAGTCTACAAGCCCGCCCCTGA
GAGCGAAGTTGGAAGAATCTCTGTTGCCGTATCCGATGATGCTAGCGCAGTCAGAAGA
ATGGGTCAAGTGGGAATTGATTCGTTTAAGTTTCCGAAGTAAGTTCGCCTAGTTTTAGG
TTTTAG

Scaffold54 is a mitochondrial sequence, which is not surprising, but it matches most closely to *Phoenix dactylifera* (date palm) and *Nepenthes ventricosa* (pitcher plant) mitochondria at 100% coverage at 98% identity.

Scaffold5 mitochondrial sequence most closely matches *Hesperelaea palmeri*, a now-extinct species from Tribe Oleeae found in Guadalupe Island ("World Checklist of Selected Plant Families: Royal Botanic Gardens, Kew" n.d.; L. Zedane et al. 2016).

Scaffold134 is also mitochondria matching to Olea europaea as expected.

F. latifolia

Mitochondria: scaffolds 1,2,4,5.

F. ornus

Ornus scaffold 1&2 mitochondrial genome may have sequences attached from a nuclear integration. The sequence shows a sharp change in GC content to nuclear background levels and that looks similar to the *F. pennsylvanica* integration. The sequence in question matches to *O. europeaea*; unfortunately, it is an unplaced scaffold. The same scaffold also matches to *Solanum lycopersicum cultivar* I-3 chromosome 10.

S3: Terms and Files Referenced in Chapter 2

DDVUtils.py - copy, fetch, filter, sort long lists of contigs

Fluentdna.py - server architecture, rendering modes, table of user input options

Layouts.py - modular layout design enables mouseover sequence for any layout style, user defined layouts

TileLayout.py - Standard rectangular layout visualization. Default is Powers of 10 Layout (Figs. 2, 4 & 5).

Index.html - what is packaged inside a FluentDNA visualization. Contig Spacing JSON, layout JSON.

nucleotideNumber.js - Inverse algorithm for each layout, chromosome sequence streaming

Annotations.py - Read Annotation GFF2 and GFF3. Filtering annotation types. Generate pseudosequence from a GFF.

HighlightedAnnotation.py - Uses shading in alpha channel to highlight regions covered by an annotation. Labels lay over top the sequence and scale by the area available. Larger annotations get bigger labels (Figs. 4 & 5).

AnnotatedTrackLayout.py - More traditional rectangles with gene name labels to run alongside sequence visualization (Fig. 3 & 8).

MultipleAlignmentLayout.py - Proteome MSA gallery. Each Fasta file gets one MSA and one block of visualization. MSA blocks are arranged in rows and laid out in 2D according to size (Fig. 6).

ParallelGenomeLayout.py - Handles the interlacing of multiple files for Annotation Tracks or Whole Genome Alignments. Renders boxes for columns of aligned genome pairs. Whole Genome Alignments are rendered from pseudosequences produced by ChainParser.py (Figs. 3, 7 & 8).

ChainFiles.py - parser for UCSC Chained LiftOver files. A Chain contiguous alignment represented by a series of Chain entries which each have a size, gap_query, and gap_ref.

ChainParser.py - Handles the main logic of parsing whole genome alignments into a visualization. It generates two gapped sequences from reference and query genomes, then computes the differences between the two genomes (Figs. 3, 7 & 8).

AnnotatedAlignment.py - Apply gaps to GFF pseudosequence as a LiftOver visualization (Fig. 7).

Span.py - Utility class for intersecting ranges of coordinates and handling gaps inside them.

UniqueOnlyChainParser.py - See Unique sequence content by subtracting one genome from another.

Ideogram.py - Peano curve layout designed to look like a packed chromosome under the microscope. This layout preserves locality, causing gene regions to appear as bumpy regions like counties on a map (Figs. 2B & 4B).

Processing Scripts

Image_resize_script.py - Set the level of magnification for any image. Useful for generating figures upscaled larger than screen display size. Downsample Genome posters for printing.

Stats_Aggregator.ipynb - Collect stats on a whole genome alignment across many chromosomes.

RepeatAnnotations.py - Fetch all sequences from RepeatMasker output - Show Repeat Diversity within Human

AnnotationAlignment.py - Use chain file to perform RepeatAnnotations fetch on a query genome

TransposonLayout.py - Layout for RepeatAnnotations

S4: Sequence Tubemap Scalability Limits Supplement to Chapter 5

The first set of examples are a synthetic toy dataset with four nodes used to demonstrate the lack of reuse of Edges. While 4 nodes only have 16 edges between them, Sequence Tubemap creates a new entity for each Path which traverses the Edge. This leads to more of the screen being taken up by Path stacking as the number of individuals increases.

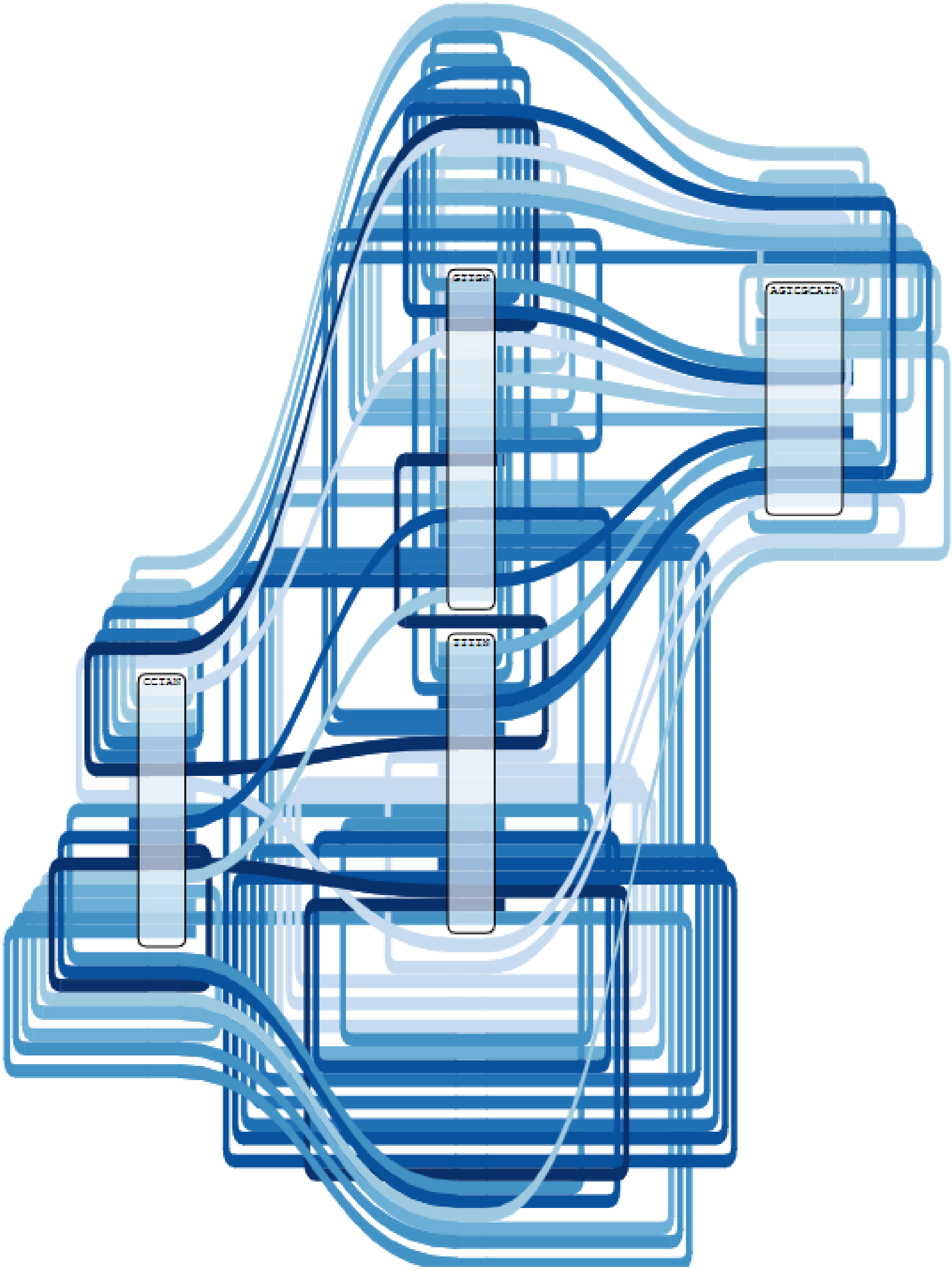


Figure: 4 Nodes, 16 Edges, 20 Paths traversing the Nodes in a random ordering.

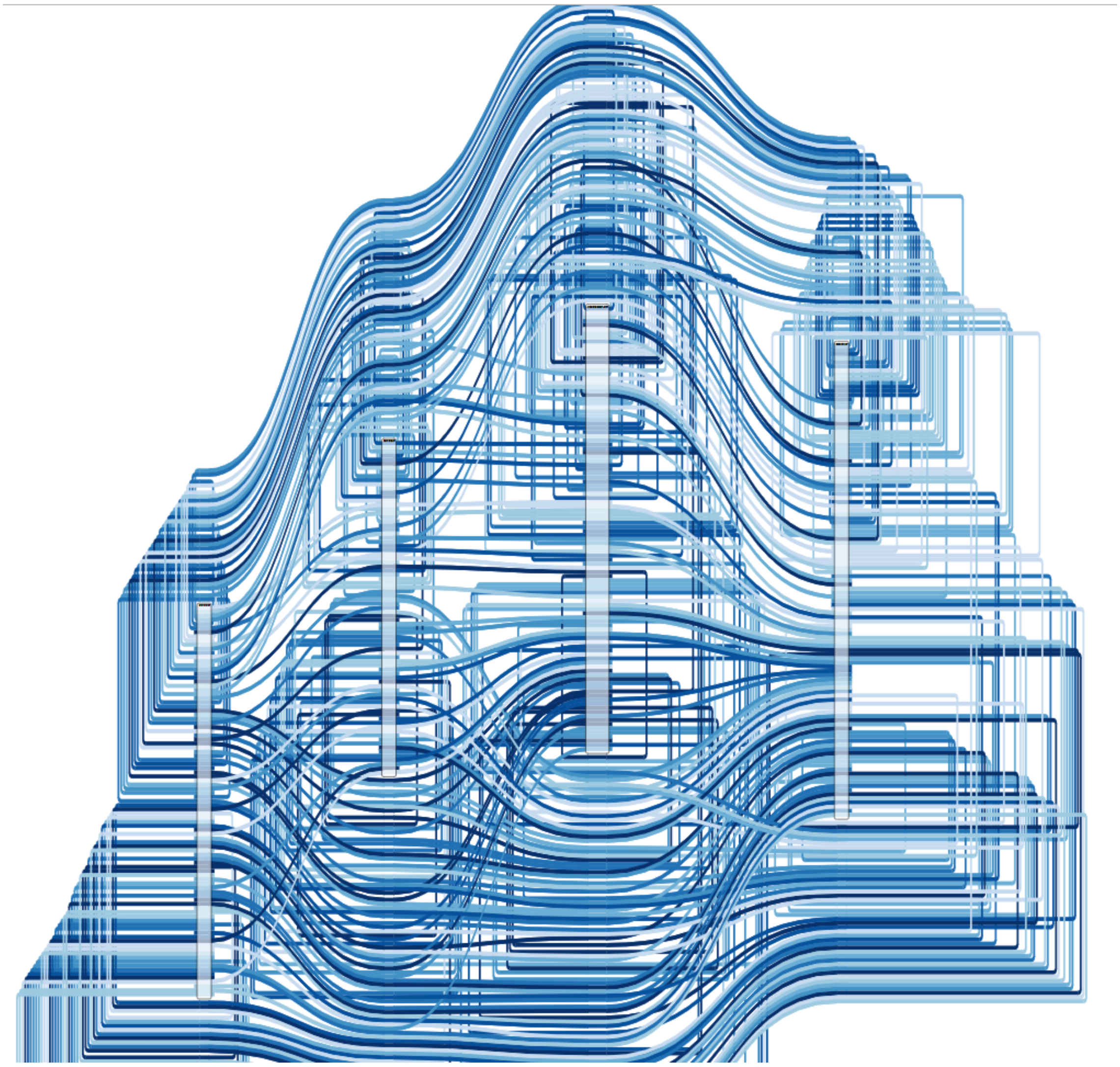


Figure: 4 Nodes, 16 Edges, 100 Paths traversing the Nodes in a random ordering.

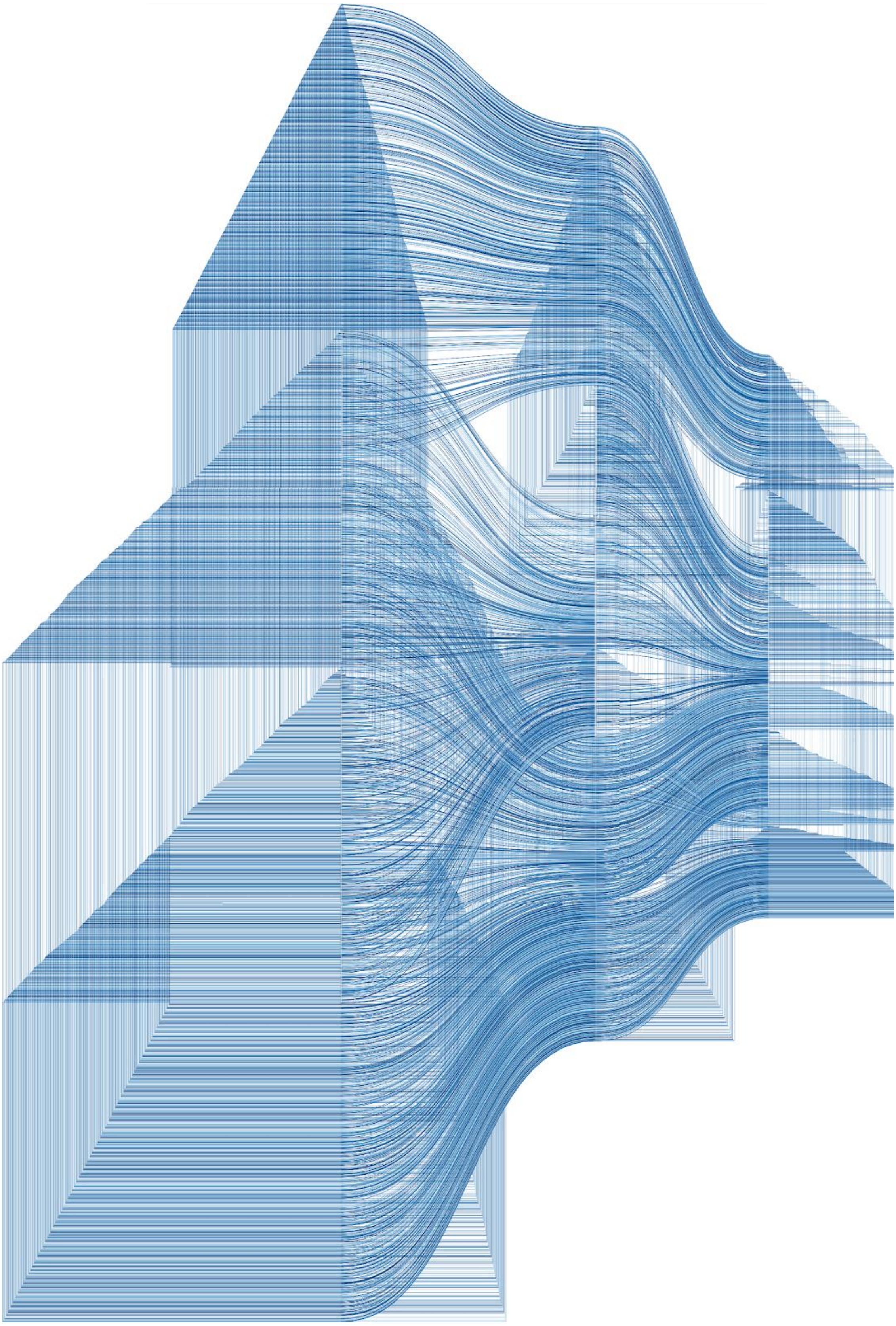
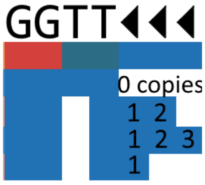


Figure: 4 Nodes, 16 Edges, 1000 Paths traversing the Nodes in a random ordering. This image was composited because it overloaded the browser's ability to zoom out. Sequence Tubemap has undeniable artistic appeal, even in failure conditions.

S5. Link Column Ordering

This section contains design details for developers for edge cases with Link Columns in Components with multiple traversals. It delves into technical design decisions that require a deep understanding of the subject matter. First, when interpreting a pantograph schematic, each Link is only followed once, otherwise anything with a repeat would be an infinite loop. Each component has sequence content in the middle, arrivals (upstream Links) on the left and departures (downstream Links) on the right.

Inversion Split Criteria



Inversions are read in the same direction, left to right but their sequence is interpreted as the reverse complement of the whole component sequence. In order for this to work, inversion must be a criteria for splitting of a Component. Nested inversions are possible through the use of Links. A whole Component row must be either inverted or not inverted. In reality, using binning we get a decimal percentage of inversion per bin e.g. 0.817 inversion. This is due to a necessary lack of resolution over an entire bin which will contain multiple nodes which are either inverted or not. For our criteria, an entire component row must be either above 0.5 inversion (colored inverted) or below 0.5 for the entire length.

Link Ordering stacks from Inside Out

When constructing a Matrix component, the sequence content starts in the middle flanked by arrivals on the left and departures on the right. Each time a path traverses the component, a pair of Link Columns is added to the outermost (far left and far right) of the component. Traversals read from the inside out. Following these rules, a single path that traverses the same Link multiple times would create multiple Link Columns with redundant Links of different colors. It is likely simpler to 1) reuse the same color and Link in multiple Link Columns or 2) collapse all traversals of an Link into one Link Column and assign it a copy number in the “coverage” metadata. Both of these compressions are lossy for duplications but not rearrangements. Which is the best implementation will depend on how repetitive the input data is.

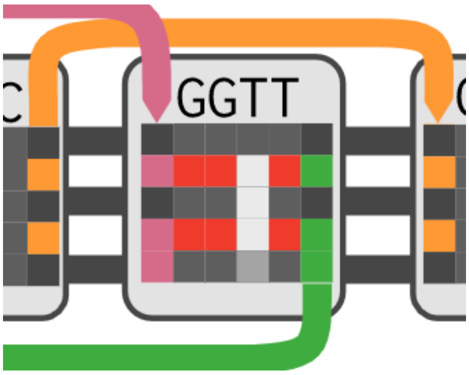
In our implementation, we use (1) because counting the number of active cells in a particular row is their copy number. Visually, this means every Link Column is also a histogram of copy number. Traversals aren’t ambiguous given a column number when following a Link (hyperlinked in the browser). Vertically, this has the same scalability as option (2) for the same reasons that Link visualization is better than Sequence Tubemap’s vertical Path stacking.

Incompatible Link Column Orderings

The Link Column ordering for a sorted pangenome schematic containing one path is guaranteed to correct and unambiguous. However, because Link Columns are created one path at a time and reused by successive paths, there are rare edge cases where the schematic implies the wrong ordering of Link Column traversals. A Link Column is created when the first path traverses that Link. Subsequent Paths that traverse the same Link will reuse the same column whose column order was determined by the first Path. It is conceivable that this will create cases where the interpretation will put the traversal of paths in the wrong order. This is a consequence of the lossy compression in favor of scalability. The intention is that the majority of cases are lossless, but there will always be some ambiguous cases when reusing the same visual space.

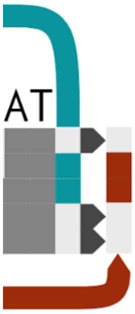
Ambiguous Variation in Copies

The most prominent case of lossy compression in the visualization is sequence variation within a repeat. Consider the last row in the figure on the right. The last individual has two traversals of the GGTT Component. We can see that in one case the sequence reads GGT and the other reads GGTT, but we will not be able to tell which traversal has the deletion from the visualization. There was no apparent solution to this ambiguity that does not sacrifice scalability. Mouseover will need to cover these shortcomings by listing the real sequences for each Component traversal in order.



The Implied Adjacent Column

The Adjacent Connector is a Link Column that is given special treatment because it is the most common and also the most unremarkable. All Components are sorted to maximize adjacency to their neighboring Component. Ideally, the most common Link is the charcoal colored Adjacent Connector between Components. In the typical case, there is a single Adjacent Link Column that is shared between the two adjacent Components. When one path has more than one traversal of the same two Components, additional Adjacent Columns have to be added to reflect the increased copy number. Additional Adjacent connectors are not yet implemented in the Pantograph 1.0 release.



S6. Example SPARQL Query from Semantic Variation Graph

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX uniprotkb: <http://purl.uniprot.org/uniprot/>
PREFIX uberon: <http://purl.obolibrary.org/obo/uo#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX sp: <http://spinrdf.org/sp#>
PREFIX SLM: <https://swisslipids.org/rdf/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX sh: <http://www.w3.org/ns/shacl#>
PREFIX schema: <http://schema.org/>
PREFIX rh: <http://rdf.rhea-db.org/>
PREFIX pubmed: <http://rdf.ncbi.nlm.nih.gov/pubmed/>
PREFIX patent: <http://data.epo.org/linked-data/def/patent/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX orthodb: <http://purl.orthodb.org/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX np: <http://nextprot.org/rdf#>
PREFIX nextprot: <http://nextprot.org/rdf/entry/>
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX mnet: <https://rdf.metanetx.org/mnet/>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>
PREFIX lscr: <http://purl.org/lscr#>
PREFIX keywords: <http://purl.uniprot.org/keywords/>
PREFIX identifiers: <http://identifiers.org/>
PREFIX glyconnect: <https://purl.org/glyconnect/>
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX genex: <http://purl.org/genex#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX eunisSpecies: <http://eunis.eea.europa.eu/rdf/species-
schema.rdf#>
PREFIX ensembltranscript:
<http://rdf.ebi.ac.uk/resource/ensembl.transcript/>
PREFIX ensemblterms: <http://rdf.ebi.ac.uk/terms/ensembl/>
PREFIX ensemblprotein:
<http://rdf.ebi.ac.uk/resource/ensembl.protein/>
PREFIX ensemblexon: <http://rdf.ebi.ac.uk/resource/ensembl.exon/>
PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/>
PREFIX ec: <http://purl.uniprot.org/enzyme/>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX cco: <http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX chebihash: <http://purl.obolibrary.org/obo/chebi#>
PREFIX CHEBI: <http://purl.obolibrary.org/obo/CHEBI_>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX allie: <http://allie.dbcls.jp/>
PREFIX GO: <http://purl.obolibrary.org/obo/GO_>
PREFIX orthodbGroup: <http://purl.orthodb.org/odbgroup/>
PREFIX vg: <http://biohackathon.org/resource/vg#>
PREFIX insdc: <http://ddbj.nig.ac.jp/ontologies/nucleotide/>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX up: <http://purl.uniprot.org/core/>
SELECT
    DISTINCT
       ?insdCDS #?insdCDSBegin ?insdCDSEnd ?step
       ?uniprot ?stepBeginInProteinSpace ?stepEndInProteinSpace
?annotationText
WHERE
{
#  Find CDS annotated by INDSC that do not match a UniProt protein.
  ?insdCDS insdc:translation ?sequence ;
            a insdc:Coding_Sequence ;
            faldo:location ?insdCDSLocation .
  MINUS {
      ?uniprotSequence rdf:value ?sequence .
  }

#   Get the range of this CDS and make sure the coordinates are on
the
#  path we need later
  ?insdCDSLocation faldo:begin [ faldo:reference ?path ;
                              faldo:position ?insdCDSBegin] ;
                  faldo:end [ faldo:reference ?path ;
```

```
                              faldo:position ?insdCDSEnd] .

  ?step a vg:Step ;
      vg:path/skos:closeMatch ?path ;
      vg:node ?node ;
      faldo:begin [ faldo:reference/skos:closeMatch ?path ;
                               faldo:position ?insdcStepBegin ] ;
      faldo:end [ faldo:reference/skos:closeMatch ?path ;
                         faldo:position ?insdcStepEnd ] .
## I always forget how to interval ranges :(
  FILTER ( (?insdcStepBegin >= ?insdCDSBegin && ?insdcStepBegin <=
?insdCDSEnd) ||
            (?insdCDSBegin >= ?insdcStepBegin && ?insdCDSBegin <=
?insdcStepEnd) ||
                  (?insdcStepEnd >= ?insdCDSEnd && ?insdcStepEnd <=
?insdCDSBegin) ||
            (?insdCDSEnd >= ?insdcStepEnd && ?insdCDSEnd <=
?insdcStepBegin) )
##  Then we look for a node close to the ones in the CDS in genome
graph space (one step)
  ?node vg:linksForwardToForward ?nextNode .
  ?step2 a vg:Step ;
      vg:path/skos:closeMatch ?nextPath ;
      vg:node ?nextNode .
##  Where that node is on a uniprot matching sequence
  ?nextinsdCDS insdc:translation ?nextSequence ;
            a insdc:Coding_Sequence ;
            faldo:location/faldo:begin/faldo:reference ?nextPath .
  ?uniprot up:sequence/rdf:value ?nextSequence .
  BIND(IF(?insdCDSBegin > ?insdcStepBegin, ?insdCDSBegin,
?insdcStepBegin - ?insdCDSBegin)/3 AS ?stepBeginInProteinSpace)
  BIND(IF(?insdCDSEnd > ?insdcStepEnd, ?insdcStepEnd, ?insdCDSBegin -
?insdcStepEnd)/3 AS ?stepEndInProteinSpace)
  ?uniprot up:annotation ?annotation .
  ?annotation a up:Active_Site_Annotation .
  ?annotation up:range ?annotationRegion .
  ?annotation rdfs:comment ?annotationText .
  ?annotationRegion faldo:begin/faldo:position ?annotationBegin .
  ?annotationRegion faldo:end/faldo:position ?annotationEnd .
  FILTER (?annotationBegin >= ?stepBeginInProteinSpace &&
?annotationEnd < ?stepEndInProteinSpace )
}
```