Title
Enhancer evolution in the Drosophila montium subgroup

Author
Bronski, Michael J

Publication Date
2018

Peer reviewed|Thesis/dissertation

Enhancer evolution in the *Drosophila montium* subgroup

By

Michael J. Bronski

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Eisen, Chair
Professor Iswar Hariharan
Professor Nipam Patel
Professor Doris Bachtrog

Fall 2018

Enhancer evolution in the *Drosophila montium* subgroup

Abstract

Enhancer evolution in the *Drosophila montium* subgroup

by

Michael J. Bronski

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Michael Eisen, Chair

Enhancers drive spatiotemporal patterns of gene expression, and play critical roles in development, disease, and evolution. Decades of research have yielded key insights, but many questions remain unanswered. A hallmark of enhancer evolution is functional conservation in the presence of extensive sequence divergence. However, identifying important mutational events between divergent sequences has been challenging. To overcome this challenge, I adopted a comparative genomic approach: sequence and assemble dozens of closely related species, and study enhancer evolution at the earliest stages of divergence. Such a data set provides an unprecedented opportunity to identify key changes and events (along with their context) before they are obscured by additional mutations. I started by sequencing and assembling 23 genomes from the *Drosophila montium* subgroup, a large group of closely related species. I also aligned each *montium* assembly to the extensively annotated *D. melanogaster* genome. The average scaffold NG50 is 76 kb, but varies widely (400 - 19 kb) depending on repeat content and heterozygosity levels. Despite large differences in contiguity, all *montium* assemblies contain high percentages of known genes and enhancers - demonstrating their suitably for this comparative genomic approach. To support my subsequent analyses, I also reconstructed the *montium* subgroup phylogeny using 20 Bicoid-dependent enhancers.

Next, I leveraged this new genomic resource to study enhancer evolution across 24 *montium* species and *D. melanogaster*. I started with the extensively characterized *eve* stripe 2 enhancer, and showed how patterns of (apparent) conservation and variation could be used to direct targeted mutagenesis experiments, and to inform models of enhancer grammar. To study binding site turnover on a large scale, I investigated hundreds of ChIP peaks for the transcription factors Bicoid, Krüppel, and Zelda. I treated groups of orthologous binding site scores as continuous traits, reconstructed

ancestral scores at each node of the species tree, and then calculated score changes along each branch of the tree. For all three factors, binding sites were more likely to be gained along branches of the tree that also lost a binding site. This was true for both conserved and non-conserved sites, and most differences were statistically significant. However, I observed similar patterns when I repeated the analyses using shuffled matrices, leaving me unable to conclude these were meaningful changes in transcription factor binding. Future analyses will focus on mitigating the effects of several confounding factors, including non-functional *montium* sequences, the forced gradualism of the Brownian motion model, and ancestral character estimation with a single species tree in the presence of widespread incomplete lineage sorting and / or introgression.

Finally, in collaboration with Carolyn Elya and Michael Eisen, I worked on assembling the genome of the *Drosophila*-manipulating fungus *Entomophthora muscae* 'Berkeley'. This is an excellent system with which to study the mechanistic basis of parasite-induced manipulations. Infected flies exhibit a suite of behavioral changes, including summit disease, proboscis extension / attachment, and raised / spread wings. Compared to most previously sequenced fungi, the genome is extremely large and repetitive. The total scaffold length is 1.24 Gb, but the haploid genome size might be around 650 Mb. Polyploidy appears to be common among related entomopathogenic fungi, so estimating the haploid genome size in the absence of additional experimental data is challenging. At least 85 % of the genome is repeats. In fact, the genome is so repeat-rich that aligning any pair of scaffolds produces characteristic X-alignments, where the forward strand of the first scaffold also aligns to the reverse complement of the second scaffold. The assembly appears to be missing many known fungal genes, but the significance of this is unclear. For genes that are present, the genome often appears to contain two distinct haplotypes. In many cases these haplotypes were assembled independently on different scaffolds, but many were also collapsed into single sequences. The alignment of PacBio long-reads to the assembly suggests that it contains numerous mis-assemblies. This was probably unavoidable given the genome's dense repeat structure. Future efforts will focus on improving the assembly. Going forward, the *E. muscae* 'Berkeley' genome will support our efforts to understand the molecular basis of fungal-induced behavioral manipulations in *D. melanogaster*.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Georgann Bronski: When I became seriously ill during graduate school, my mother - a registered nurse - sold her home in New York, and moved to Berkeley. If it wasn't for her love, expertise, and willingness to sacrifice, I would never have completed this thesis.

Michael Eisen: Thank you for the opportunity to work on such exciting projects, and for your extraordinary understanding and flexibility. During an extremely challenging time in my life, science became an oasis. I'll always be deeply grateful.

Rebecca Wilbanks, Concetta Spirio, and Mary Young: Thank you for all of your help, support, and sacrifice over the years. Without it, this endeavor would have ended long ago.

Iswar Hariharan, Nipam Patel, and Doris Bachtrog: Thank you for providing guidance and direction over the years.

The Department of Molecular and Cell Biology, especially Matt Welch: Thank you for being so helpful and accommodating over the years. It really made a difference.

Carolyn Elya, Ciera Martinez, Holli Weld, Kelly Schiabor, Alli Quan, Elizabeth Roeske, Peter Combs, Matt Davis, Jenna Haines, Ashley Albright, Xiao-Yong Li, Michael Stadler, Augusto Berrocal, Aaron Hardin, Steven Kuntz,, Matt Macmanes, Jackie Villalta, Mathilde Paris, Wendy Ingram, Tim Howes, Angus Chandler, and the rest of the Eisen Lab: Thank you for being amazing labmates. It was a privilege working with such talented people. I'll always remember the "Bronski Wars" video.

Friends and family, near and far, especially Mario Pietromonaco, Peter Bronski, Kelli Bronski, Marin Bronski, Charlotte Bronski, Timothy Bronski, Ron Star, Connie Star, Adam Star, and Cate Star: Thank you for all the visits, check-ins, and well wishes. They made all the difference.

Stefan Prost: Thank you for the helpful workshops and discussions on genome assembly.

Rick Harrison and Jeff Doyle: Thank you for the opportunity to work in your labs so many years ago, and for starting my training as a biologist.

Chapter 1: Introduction

Enhancers

Enhancers - cis regulatory modules (CRMs) that drive spatiotemporal patterns of gene expression - play critical roles in development [1], evolution [2], and disease [3]. Since their discovery in the SV40 genome nearly 40 years ago [4, 5], they have been the subject of intensive research. While much has been learned, key questions remain unanswered.

Enhancers are relatively small non-coding sequences, ranging in size from several hundred to one thousand base pairs. They interact with promoters and the basal transcriptional machinery (e.g., RNA polymerase II) to activate gene expression through a mechanism that likely involves looping [6]. Enhancers can be located far from the promoters / genes they regulate (up to 1 Mb in mammals [7]), and can even ignore adjacent genes as they interact with distant promoters [8]. The relative orientation of the enhancer sequence is unimportant for its function [4]. Enhancers also contain numerous transcription factor binding sites for integrating regulatory input from multiple factors, each of which can act as an activator or repressor of transcription [e.g., 9].

Regulation of the *eve* stripe 2 enhancer

The pair-rule gene *even-skipped* (*eve*) is expressed in seven transverse stripes in the early *Drosophila* embryo. The entire pattern is controlled by the combined action of five enhancers, each of which drives expression in one or two stripes [10]. The stripe 2 enhancer is arguably the most extensively studied regulatory sequence in all of biology. A detailed summary of its structure, function, and evolution will be illustrative.

Early experiments provided key insights into the structure and function of the *eve* stripe 2 enhancer [9, 11-14]. Briefly, the enhancer is activated by the maternal morphogen Bicoid and the gap gene *Hunchback*, and repressed by the gap genes *Giant* and *Krüppel*. Bicoid and Hunchback form gradients that are concentrated in the anterior half of the embryo. *Giant* is expressed in two bands located in the anterior and posterior of the embryo, while Krüppel protein is localized to a single band close to the midline. Robust *eve* stripe 2 expression occurs within a narrow stripe where the activators Bicoid and Hunchback are present, but the repressors Giant and Krüppel are absent. Giant and Krüppel expression define the anterior and posterior boundaries of the stripe, respectively. Binding sites for these factors are found in clusters of overlapping Bicoid-Krüppel and Hunchback-Giant sites. Subsequent experiments showed that repression of *eve* stripe 2 in the anterior tip of the embryo is mediated by redundant mechanisms, including repression by Sloppy-paired 1, and the downregulation of Bicoid activity by Torso [15]. The pioneer factor Zelda is also necessary for normal activity [16].

Comparative analysis of the *eve* stripe 2 enhancer

Early comparative studies of orthologous *eve* stripe 2 enhancers across closely and distantly related species provided key insights into enhancer evolution [17-19]. Despite the fact that the ablation of individual transcription factor binding sites in *D. melanogaster* causes aberrant expression [13, 14], detailed investigation of orthologous enhancers in *D. yakuba*, *D. erecta*, and *D. pseudoobscura* showed that binding sites were frequently gained and lost on evolutionary timescales, a process known as binding site turnover [18]. Furthermore, the size of the enhancer varied between species, as well as the distances between key binding sites. Remarkably, when *eve* stripe 2 enhancers from these species were transfected into transgenic *D. melanogaster*, they drove the endogenous stripe 2 pattern [18]. An impressive result given that *D. melanogaster* and *D. pseudoobscura* diverged approximately 60 million years ago [20]. Despite this result, differences between species do matter. Chimeric sequences constructed from the 5' and 3' halves of stripe 2 enhancers from *D. melanogaster* and *D. pseudoobscura* drove aberrant expression patterns. This indicated that divergent compensatory mutations had evolved in each species [19].

Hare et al. [21] extended these findings to sepsid flies, which diverged from *Drosophila* more than 100 millions years ago. Despite the deep divergence time, early embryonic patterning appears to be conserved. Expression patterns for the key development regulators *Giant*, *Hunchback*, and *Krüppel* are nearly identical in *D. melanogaster* and *Themira minor* (a Sepsid species). Amazingly, when sepsid *eve* stripe 2, stripe 3+7, stripe 4+6, and muscle-heart enhancers were tested in transgenic *D. melanogaster* embryos, they drove the endogenous expression patterns. This is an astounding result given that these species diverged at least 100 million years ago. To better understand how these enhancers changed over time, Hare et al. compared computationally-predicted binding sites across *Drosophila* and sepsid species. Relative to *D. melanogaster*, binding sites in sepsids were nearly completely rearranged, dramatically highlighting functional conservation in the absence of sequence conservation. They also showed that overlapping binding sites tended to be more highly conserved than sites that were only close or isolated.

Simulating binding site turnover in the *eve* stripe 2 enhancer

The overlap of binding sites for activators and repressors in the *eve* stripe 2 enhancer suggests a simple mechanistic explanation for how transcription factors can create precise expression patterns [9, 12-14]. Furthermore, the fact that overlapping binding sites appear to be conserved across distantly related species suggests that such arrangements are functionally important [21]. However, other explanations are possible. Lusk and Eisen [22] simulated the evolution of the *eve* stripe 2 enhancer using simple rules. Mutations were only accepted if they preserved the overall number of sites for a given factor. (For example, a mutation that destroyed a Bicoid site was only accepted if the enhancer had already acquired a new Bicoid site.) Interestingly, they found a marked enrichment for overlapping Bicoid-Krüppel sites, even though selection acted

only to preserve the total number of sites, not their specific spatial arrangement. (Once formed, overlapping sites tended to have long half-lives.) Given the deletion bias present in *Drosophila*, they also showed that binding sites tended to cluster together over time. When comparing binding sites across species, this process could lead to the misleading conclusion that clustered sites were conserved. While these simulations do not preclude the functional importance of overlapping or clustered binding sites, they highlight important caveats. Such features can also arise as byproducts of neutral processes, or selection acting on the overall number of sites, not their local arrangement.

The *sparkling* enhancer

Along with the *eve* stripe 2 enhancer, the *sparkling* (*spa*) enhancer of *dPax2* is one of the most extensively studied enhancers [23, 24]. In *Drosophila*, this 362 bp sequence drives expression in cone cells in the developing eye. It is regulated by the known factors Lz, PntP2/Yan, and Su(H), but also contains many novel regulatory motifs [23]. Systematic mutagenesis of the enhancer [23], combined with comparative analyses across *Drosophila* species (including chimeric *D. melanogaster* - *D. pseudoobscura* constructs) [24], have yielded key insights. For example, *spa* contains a “remote control” element (RCE) that is necessary for long-distance enhancer-promoter interactions [23]. Comparative analyses also showed that *spa* is rapidly evolving, and has undergone significant reorganization [24]. However, despite extensive sequence divergence, the function of the enhancer has been conserved: *D. melanogaster* and *D. pseudoobscura* orthologs drive identical expression patterns in transgenic *D. melanogaster*. Grammar elements (local arrangements of binding sites) were also identified, and these elements could be flexibly rearranged within *spa*. Additionally, the loss of a binding site for one factor could be compensated for by the acquisition of a site for a different factor. Weak sites are also important for *spa* function [24].

The interplay between binding affinity and syntax

Beyond the *eve* stripe 2 and *sparkling* enhancers, detailed investigations of enhancers in other model systems have led to importance observations. For example, Farley et al. [25] investigated notochord enhancers in *Ciona* embryos, and identified a relationship between binding affinity and syntax. Here, syntax refers to the order, orientation, and spacing of binding sites. Notochord enhancers are activated by the transcription factors ETS and ZicL. They found that optimal spacing (11 bp) and orientation (oppositional) of linked ETS - ZicL sites, so-called syntax, could compensate for weak binding sites. Conversely, high-affinity binding sites were able to adopt more flexible spatial arrangements. This study highlighted the importance of weak binding sites. It also showed how focusing on high-affinity binding sites could obscure underlying syntax constraints.

Genome-wide enhancer screens

Historically, testing sequences for enhancer function was time-consuming and laborious. Recently, new assays like STARR-seq (self-transcribing active regulatory region sequencing) [26] have made it possible to simultaneously test millions of candidate enhancer sequences in a single experiment. Briefly, candidate sequences are placed downstream of minimal promoters, where they can potentially drive their own expression. The activity of an enhancer is therefore proportional to the level of its own RNA. After libraries are transfected into cell lines, polyadenylated RNA is extracted, reverse-transcribed, PCR-amplified, and deep-sequenced. Finally, reads are mapped back to the genome to quantify enrichment.

The application of STARR-seq to five *Drosophila* species (*melanogaster*, *yakuba*, *ananassae*, *pseudoobscura*, and *willistoni*) [27] showed that many *D. melanogaster* enhancers are functionally conserved, even in distantly related species. Despite this functional conservation, compensatory binding site turnover between species is common. Hybrid enhancers created by fusing halves from different species yielded sequences with different activities. Entirely new enhancers also arise frequently, even between closely related species. In the approximately 11 million years since the divergence of *D. melanogaster* and *D. yakuba*, hundreds of sequences acquired enhancer function.

Enhancer models

Three enhancer models have been proposed. They differ on whether transcription factor binding is cooperative, and the importance of syntax or grammar. The first enhancer model was based on the virus-inducible enhancer (really the promoter) for the human interferon-β (IFNβ) gene [28, 29]. This 55 bp sequence is bound by eight factors forming a highly ordered and continuous complex known as an enhanceosome, after which the model is named. These factors bind cooperatively, and the precise spacing and orientation of sites is critical. In fact, this sequence is almost perfectly conserved across 100 million years of mammalian evolution . Unsurprisingly, enhanceosomes are intolerant of binding site turnover . In contrast to the enhanceosome model, the information display or “billboard” model does not depend on large-scale cooperative binding, or the precise arrangement of sites [14, 30, 31]. Instead, subelements within the enhancer bind transcription factors (activators and/or repressors) independently. These regions display their information to the basal transcriptional machinery - either iteratively or simultaneously - where it is interpreted. The information presented by different regions within the same enhancer can be contrasting. This model was informed by experiments on simple genetic switch elements in *Drosophila*. It is a highly flexible model that allows for extensive binding site turnover. Finally, the “TF collective” model is based on heart enhancers in *Drosophila*, which are regulated by at least five factors [32]. According to this model, transcription factors bind cooperatively, but unlike the enhanceosome, the arrangement of binding sites is unimportant. Motifs for some factors

can even be missing. Enhancers regulated in this way would be highly tolerant of binding site turnover.

Unanswered questions

Orthologous enhancers frequently exhibit functional conservation in the absence of underlying sequence conservation. The rules that allow enhancer sequences to diverge, while simultaneously maintaining the same function, are currently unknown. The relative importance of syntax or grammar in enhancer evolution is also unclear.

Chapter 2: Building the *Drosophila montium* subgroup into a genomic resource

Abstract

Enhancers can exhibit remarkable functional conservation despite extensive sequence divergence. The rules that allow enhancers to diverge, while simultaneously maintaining function, are currently unknown. Attempts to understand this process using divergent, yet functionally conserved, sequences are stymied by the large number of mutational events that have occurred. To overcome this challenge, I propose a comparative genomic approach: sequence and assemble dozens of closely related species, and study enhancer evolution at the earliest stages of divergence. The *Drosophila montium* subgroup contains roughly 100 species, and is well-suited to this approach. In this chapter, I describe the sequencing and assembly of 23 *montium* genomes. To make this endeavor financially feasible, I sequenced a single, small-insert library for each species. All genomes were initially assembled using MaSuRCA, before going through an extensive post-assembly pipeline. The average estimated genome size is 195 Mb, and most assemblies (total scaffold lengths) reach 85 - 95 % of the estimated genome size. The scaffold NG50s vary widely, depending in large part on repeat content and heterozygosity levels. The average scaffold NG50 is 76 kb, and individual assemblies range from 400 kb to 19 kb. Despite large differences in contiguity, all genomes contain at least 96 % of known single-copy Dipteran genes (BUSCOs). To facilitate the identification of enhancer sequences within *montium* species, and to further develop the *montium* clade as a genomic resource, I aligned each *montium* assembly to the extensively annotated *D. melanogaster* genome. I then used the whole-genome alignments to remap coordinates for thousands of *D. melanogaster* enhancers onto each *montium* assembly. Pairwise BLAST alignments between *D. melanogaster* and *montium* sequences showed that the vast majority of these sequences appear to be orthologous. Finally, I reconstructed the *montium* subgroup phylogeny using 20 Bicoid-dependent enhancers. Going forward, this new genomic resource will support my efforts to study enhancer evolution.

Introduction

A recurring theme in enhancer evolution is the conservation of function despite extensive sequence divergence. This has been demonstrated in detailed analyses of individual enhancers, such as *eve* stripe 2 [17-19, 21] and *sparkling* [24], along with massively parallel screens [27]. How these processes occur is still poorly understood.

What are the rules that allow the number, strength, and arrangement of transcription factor binding sites to change over time, while simultaneously preserving enhancer function? It's difficult to answer this question starting with distantly related species. So

many mutations have accumulated that the order of key events is unclear, and it's difficult to identify the changes that actually matter.

To answer this fundamental question, I propose a comparative genomic approach: sequence and assemble dozens of closely related species, compare hundreds / thousands of enhancers, and study enhancer evolution at the earliest stages of divergence. This approach has a number of advantages. At these distances, the most closely related species will differ at only one or two binding sites. So when an important event does occur - like the loss of a conserved site - candidate compensatory changes can be identified before they are obscured by additional mutations. Furthermore, by exploiting naturally occurring variation, I can gain valuable insight into changes that are allowed to occur, as well as those that are forbidden. These data can also be used to inform targeted mutagenesis experiments, yielding additional insights. Eventually, these data can be used to build models of enhancer evolution that can be tested using synthetic constructs.

The *Drosophila montium* species subgroup contains 98 species [33], and is well-suited to this approach. Species from the *montium* subgroup have been used to study a variety of evolutionary, ecological, and behavioral questions, including the genetic basis of female-limited color polymorphism [34], cold and desiccation resistance [35], adaptation to drought stress [36], and courtship behavior [37]. More than 30 species are available in culture, with more on the way. The *montium* subgroup diverged from *D. melanogaster* approximately 40 million years ago [20]. It's relative proximity to an extensively studied and annotated reference genome means that I can leverage existing tools to identify and study *montium* enhancers.

Two *montium* genomes have already been assembled. The *D. kikkawai* genome was sequenced to a depth of 182x coverage using a combination of 454 and Illumina technology. This produced a 164 Mb assembly with a scaffold N50 of 904 kb [38]. The *D. serrata* genome was sequenced to a depth of 63x coverage using PacBio long-reads. It yielded a 198 Mb assembly with a contig N50 of 943 kb [39]. While these approaches generated high-quality draft assemblies, the associated costs preclude sequencing dozens of *montium* species this way.

Therefore, I needed to assemble dozens of *montium* genomes in a cost-effective way, while also producing assemblies of sufficient quality and completeness to study enhancers genome-wide. In this chapter, I describe the sequencing and assembly of 23 *montium* genomes. To make this endeavor financially feasible, I sequenced a single, small-insert library for each species. This generally led to fragmented assemblies. However, despite significant variability in contiguity, I show that all assemblies contain high percentages of known genes. I also aligned each *montium* assembly to the *D. melanogaster* genome, and show that these alignments can be used to identify thousands of putatively orthologous *montium* enhancers. Finally, I reconstruct the *montium* subgroup phylogeny using 20 Bicoid-dependent enhancers.

Results

Genome sizes and assembly statistics

To assemble dozens of genomes in a cost-effective way, I sequenced a single, small-insert (350 bp), PCR-free, library to roughly 30x coverage for each species. The genomes were assembled using the Maryland Super Read Cabog Assembler (MaSuRCA), which combines de Bruijn graph and overlap-layout-consensus (OLC) approaches into a novel algorithm based on “super-reads” [40]. The genomes then went through an extensive post-assembly pipeline to greatly improve upon the primary assemblies. (See the Materials and Methods for an in-depth description of the entire pipeline.)

Table 2.1 reports genome size estimates and assembly statistics for 23 *montium* species. The scaffold / contig NG50 is analogous to the well-known N50, but substitutes the estimated genome size for the total assembly length [41, 42]. For example, a scaffold NG50 of 100,000 bp means that 50 % of the estimated genome size is present in scaffolds that are at least 100,000 bp. When this calculation is repeated for all integers from 1 to 100, the result is an “NG graph” [42]. Figure 2.1 is an NG graph showing the distribution of scaffold lengths for 23 *montium* assemblies. (While NG graphs have typically been used to compare different assemblies of the same species [42], I use one here to highlight a limited number of features across multiple species.)

The 23 *montium* genomes range in size from 153 Mb to 225 Mb (mean = 194.5 Mb; median = 198.5 Mb). These sizes are consistent with the previously assembled *D. kikkawai* [38] and *D. serrata* [39] genomes, with total sequence lengths of 164 Mb and 198 Mb, respectively. Genome size variation within the *montium* subgroup is related to repeat content (Figure 2.S1).

The scaffold NG50s vary widely, from the remarkably contiguous *D. kanapiae* assembly (402 kb), to the highly fragmented *D. triauraria* assembly (19 kb). The contiguity of the *D. kanapiae* assembly is somewhat surprising (given the use of a small-insert library), but is related to genome and sample characteristics described below. The average scaffold NG50 across all *montium* species is 76,154 bp (median = 58,240 bp).

Many factors influence the contiguity of an assembly, including repeat content, heterozygosity, and sequencing depth. Large, repeat-rich genomes are typically difficult to assemble, as are highly heterozygous samples. Given that the *montium* genomes were assembled using small-insert libraries, they are especially sensitive to repeat-content and heterozygosity. For the *montium* genomes, the most contiguous assemblies tend to come from relatively small, repeat-poor genomes, and samples with little heterozygosity. (This accurately describes the *D. kanapiae* genome / sample). On the other hand, repeat-rich genomes and/or highly heterozygous samples, proved the most difficult to assemble (as was the case with *D. triauraria*).

The total scaffold length for most *montium* assemblies reaches 85 % - 95 % of the estimated genome size (In Figure 2.1, this is where the lines intersect the x-axis). This was expected given that each genome was assembled using a small-insert library. In the absence of long-distance information, in the form of mate-pair libraries or long-reads, large repeats form unresolvable structures in the de Bruijn graph. This results in fragmented assemblies that are missing many repeat copies. Accordingly, the total scaffold length should be significantly shorter than the estimated genome size, with the magnitude of the difference proportional to the size / number of repeats in the genome. For example, *D. mayri* (225 Mb) and *D. pectinifera* (220 Mb) have the largest estimated genomes, but yielded the shortest assemblies relative to their genome sizes (74.5 % and 67.8 %, respectively). (The *D. pectinifera* sample was also heavily contaminated with bacteria. While the bacterial reads were filtered prior to assembly, there initial presence lowered the sequencing coverage of the fly genome.) In contrast, the relatively small and repeat-poor *D. kanapiae* genome (153 Mb) yielded an assembly that reaches 99.7 % of its estimated genome size.

The total scaffold lengths for *D. leontia*, *D. punjabiensis*, and *D. watanabe* actually exceed their estimated genome sizes. (In Figure 2.1, these lines never intersect the x-axis.) Two of these differences are large: 10.6 Mb for *D. punjabiensis*, and 16.9 Mb for *D. watanabe*. These differences may be related to heterozygosity. Given modest levels of heterozygosity, most assemblers collapse allelic variation into a single consensus sequence. As the divergence increases though, allelic variation can be assembled independently and placed on two different scaffolds (one usually much shorter than the other). This artificially inflates the total scaffold length. Consistent with this effect, all three of the above samples were highly heterozygous. However, several other samples were also highly heterozygous, but yielded assemblies that were significantly shorter than their estimated genome sizes.

Overall, the *montium* assemblies are fragmented, as evidenced by their modest scaffold NG50s. However, taken in isolation, the NG50s say little about the quality of the assemblies. Any single metric (especially the NG50) can be a poor predictor of the quality / utility of an assembly. It is best to evaluate assemblies using a variety of methods, with an eye towards the downstream application [42]. For example, it is often advantageous to sacrifice contiguity for accuracy, and many questions can be answered without knowing the detailed repeat structure of the genome. I turn now to evaluating the *montium* assemblies in ways that will tell me if they are of sufficient contiguity and quality to study genes and enhancers.

The vast majority of *montium* scaffolds are at least gene-sized

To study genes, a genome assembly should be present in at least gene-sized fragments [42, 43]. (By extension, such an assembly would also be useful for studying enhancers, since they are significantly smaller than most genes.) Based on existing annotations of the PacBio *D. serrata* genome, the average gene length is up to 6.3 kb [39, 44]. Figure 2.2 shows the relationship between the scaffold NG50 and the percentage of the total scaffold length present in scaffolds that are at least 6.3 kb. (Here I use the total scaffold

length instead of the estimated genome size. If the estimated genome size was used, the percentages would obviously decrease. However, most *montium* assemblies are significantly shorter their estimated genome sizes because they are missing repeats (see above). Therefore, I think it's reasonable to ask the related question: What percentage of the non-repetitive genome is present in at least gene-sized scaffolds?) Despite large differences in contiguity, all assemblies are present predominantly as scaffolds that are at least gene-sized. While there is a small downward trend with decreasing NG50 ($r = 0.58$, $p < 0.004$), this effect is modest. Even for the most fragmented assemblies, roughly 80 % of the assembly is present in at least gene-sized fragments.

All *montium* assemblies contain high percentages of known genes

The vast majority of scaffolds in each *montium* assembly are large enough to contain genes. Do the assemblies actually contain known genes though? One way to assess the quality of an assembly is by annotation: a good assembly should contain a high percentage of known genes. Benchmarking Universal Single-Copy Orthologs (BUSCOs) are single-copy genes present in more than 90 % of surveyed species [45, 46]. The Dipteran BUSCO set contains 2,799 genes, and is based on a survey of 25 species. Figure 2.3 shows the BUSCO assessment results for eight *montium* assemblies. These species were chosen from every major subclade within the *montium* subgroup, and represent a wide range of genome sizes and contiguities. They range from the small and highly contiguous *D. kanapiae* (153 Mb, NG50 402 kb), to the large and fragmented *D. triauraria* (210 Mb, NG50 19 kb). Strikingly, despite the wide range of contiguities, there is little variation in gene content: at least 96 % of BUSCOs are present and complete across all species. The *D. kanapiae* assembly exceeds 98 %. Ten BUSCOs are missing across all eight species, and likely represent lineage-specific loss events within Diptera. For comparison, the previously assembled *D. kikkawai* and *D. serrata* genomes, which approach scaffold / contig N50s of 1 Mb, reach 98 % and 96 %, respectively [39]. Once again, despite their relatively modest scaffold NG50s, my assemblies have performed well in metrics that matter for downstream analyses.

Whole-genome alignments of *montium* species to *D. melanogaster*

To facilitate the identification of enhancer sequences within *montium* species, and to further develop the *montium* clade as a genomic resource, I aligned each *montium* genome to *D. melanogaster* using a previously described whole-genome alignment pipeline. (See the Materials and Methods for the complete pipeline.) Briefly, repeats were soft-masked in the target and query genomes using RepeatMasker [47] and Tandem Repeat Finder (TRF) [48]. Each *montium* genome was then individually aligned to *D. melanogaster* using LASTZ [49]. The LASTZ alignments were processed into structures called "chains" and "nets" using a series of programs, described in detail by Kent et al. [50]. Gapless alignments ("blocks") were linked together into maximally scoring chained alignments, or chains. The order of blocks within chains must be the same in both target and query genomes. Blocks within chains can be separated by

insertions / deletions, inversions, duplications, or translocations. Gaps in high-scoring chains were filled in with lower scoring chains, creating hierarchies (parent-child relationships) known as nets. This pipeline ultimately produced a liftOver chain file. Given a set of coordinates for an annotated feature in the *D. melanogaster* genome, the program liftOver [51] returns coordinates for the (putatively) orthologous sequence in an aligned *montium* genome.

montium assemblies contain thousands of orthologous enhancer sequences

With the genomes aligned, I turned to looking for known enhancer sequences in the *montium* assemblies. I used a previously described set of 3,500 experimentally verified enhancers that drive expression in the *D. melanogaster* embryo [52]. Using liftOver [51], I remapped the *D. melanogaster* coordinates onto each *montium* assembly. Nearly all of the enhancer sequences were "lifted" successfully (Table 2.S1). However, do the corresponding *montium* coordinates contain orthologous sequence? The original DNA fragments (tiles) from *D. melanogaster* are roughly 2 kb in length. Non-coding sequence tends to be elongated in *montium* species, so orthologous sequence should generally be at least 2 kb. (This might not always be the case though. The draft genomes are fragmented, so a single enhancer could be divided across two scaffolds. The size of the sequence may have also changed significantly since the lineages diverged, especially if the sequence is non-functional in a *montium* species. Enhancers themselves are known to turnover, even between closely related species [51].) Accordingly, I focused on a subset of remapped enhancer sequences that were between 1.5 kb and 3.5 kb. For each *montium* assembly, roughly 87 % of remapped sequences fall within this range. I then aligned each *D. melanogaster* sequence to its putative *montium* ortholog using BLASTn. Figure 2.4 shows illustrative results for *D. lacteicornis*. Based on the query coverages, percent identities, and Expect values (E), it is extremely likely these sequences are orthologous. On average, 67 % of the *D. melanogaster* sequence aligns to sequence from *D. lacteicornis* (query coverage). The average percent identity is 75 %, and the E value for the vast majority of alignments is essentially zero. Based on these results, it is clear that I can remap coordinates for thousands of *D. melanogaster* enhancers onto any *montium* assembly, and with a high level of confidence extract putatively orthologous sequences.

Reconstructing the *montium* subgroup phylogeny

Finally, to support my study of enhancer evolution within the *Drosophila montium* species subgroup, I need an accurate phylogeny with non-coding branch lengths. Previous phylogenetic reconstructions of the *montium* subgroup - typically based on a small number of genes - have produced incongruent trees [34, 37, 53-56]. For this analysis, I added the previously sequenced *D. kikkawai* genome [38], bringing the total number of *montium* genomes to 24. Chen et al. [57] previously described 66 Bicoid-dependent enhancers in *D. melanogaster*. I identified a subset of 20 enhancers that are present in all 24 *montium* assemblies, and are also spread out across the genome.

Figure 2.5 is a maximum likelihood tree constructed using RAxML [58] for 20 Bicoid-dependent enhancers. Most branches are highly supported, but the support values are relatively low for three branches. (Also see Figure 2.S4 for a matrix of distances in the *montium* clade.)

Table 2.1. Genome Size Estimates and Assembly Statistics for 23 *montium* species.

Species	Est. Genome Size (Mb)	Total Scaffold Length (Mb)	Scaffold NG50 (bp)	Longest Scaffold (bp)	Contig NG50 (bp)	Longest Contig (bp)
D. kanapiae	153.2	152.8	402,179	2,274,126	306,590	2,274,126
D. birchii	168.8	156.7	212,219	1,501,252	165,727	1,183,551
D. truncata	182.5	168.8	105,823	830,117	82,968	827,712
D. punjabiensis	194.1	204.7	89,220	1,226,934	68,747	1,083,757
D. bunnanda	174.8	151.8	84,738	1,142,480	71,624	1,127,760
D. bocki	159.7	151.2	74,334	785,450	62,781	785,450
D. vulcana	206.1	188.9	67,245	530,507	54,192	472,464
D. asahinai	209.9	189.5	64,407	1,052,132	56,009	904,342
D. mayri	225.2	167.8	59,941	2,219,437	41,680	1,355,909
D. serrata	177.9	159.7	59,530	1,091,401	47,476	718,797
D. jambulina	192.1	164.0	58,874	873,064	50,457	756,687
D. lacteicornis	198.5	182.8	58,240	1,044,495	46,923	766,914

Species	Est. Genome Size (Mb)	Total Scaffold Length (Mb)	Scaffold NG50 (bp)	Longest Scaffold (bp)	Contig NG50 (bp)	Longest Contig (bp)
D. seguyi	213.7	179.7	57,313	891,413	48,820	891,413
D. pectinifera	220.2	149.3	52,632	528,734	41,478	467,725
D. tani	191.4	182.4	51,594	921,527	46,920	780,901
D. watanabe	180.1	197.0	42,067	1,045,963	37,763	656,929
D. auraria	216.9	200.7	41,506	765,970	38,774	533,726
D. rufa	215.4	186.2	40,495	498,065	36,275	498,065
D. bakoue	215.6	192.6	32,495	1,045,797	31,083	598,596
D. leontia	162.3	164.9	26,644	331,217	24,018	301,031
D. nikananu	205.4	192.4	26,619	626,542	24,470	574,375
D. burlai	199.2	175.7	24,026	628,960	23,163	628,960
D. triauraria	209.8	199.0	19,402	590,840	18,200	576,941

Genome size estimates were calculated by SGA Preqc [66]. They are based on the *k*-mer frequency spectrum of the unassembled reads, using a method that corrects for error *k*-mers. To calculate the scaffold NG50 [41, 42], scaffold lengths were ordered from longest to shortest, and then summed. The NG50 was the scaffold length that brought the sum above 50 % of the estimated genome size. Contig lengths were estimated conservatively by splitting scaffolds on every N (including single Ns).

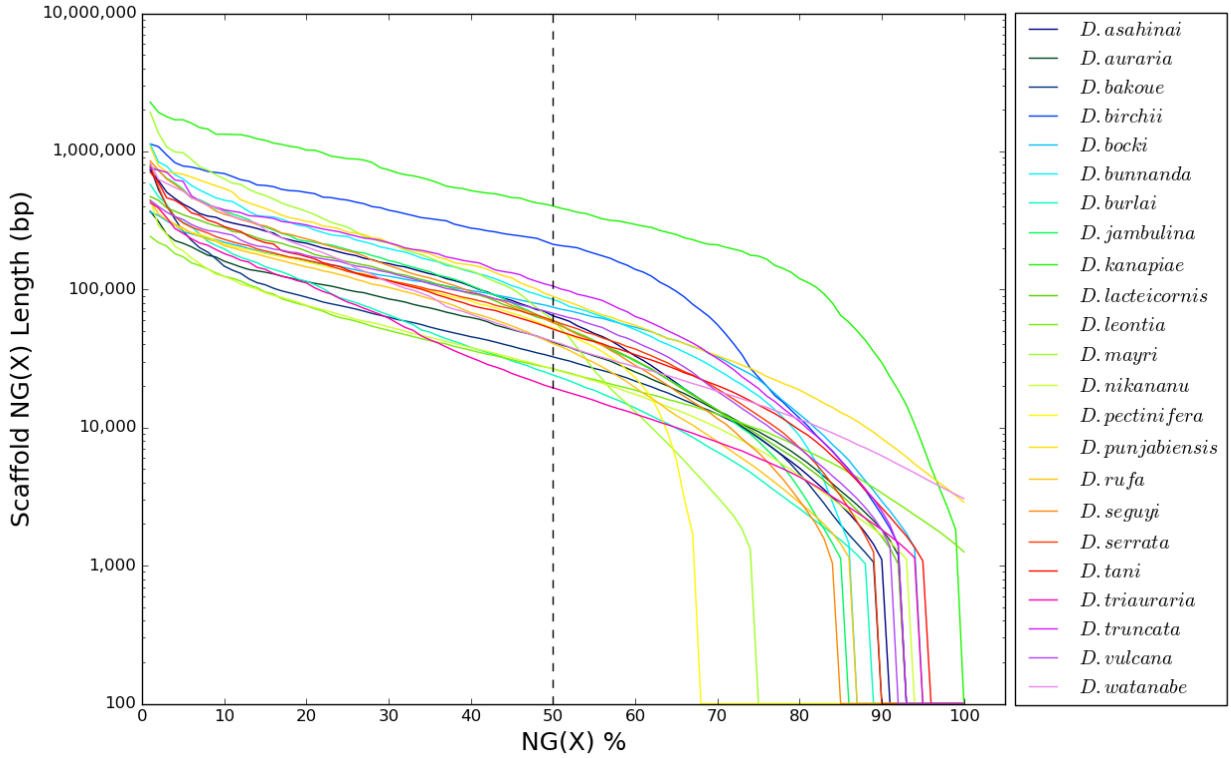


Figure 2.1. NG graph showing the distribution of scaffold lengths for 23 *montium* assemblies.
To calculate the scaffold NG50, scaffold lengths are ordered from longest to shortest, and then summed. The NG50 is the scaffold length that brings the sum above 50 % of the estimated genome size [41, 42]. When this calculation is repeated for all integers from 1 to 100, the result is an NG graph [42]. NG graphs were constructed for each *montium* species using the corresponding genome size estimates. When a series intersects the x-axis, it means the total scaffold length is shorter than the estimated genome size. Similarly, if the series never touches the x-axis, then the assembly is longer than the estimated genome size. Due to filtering, the shortest scaffold present in any assembly is 1 kb.

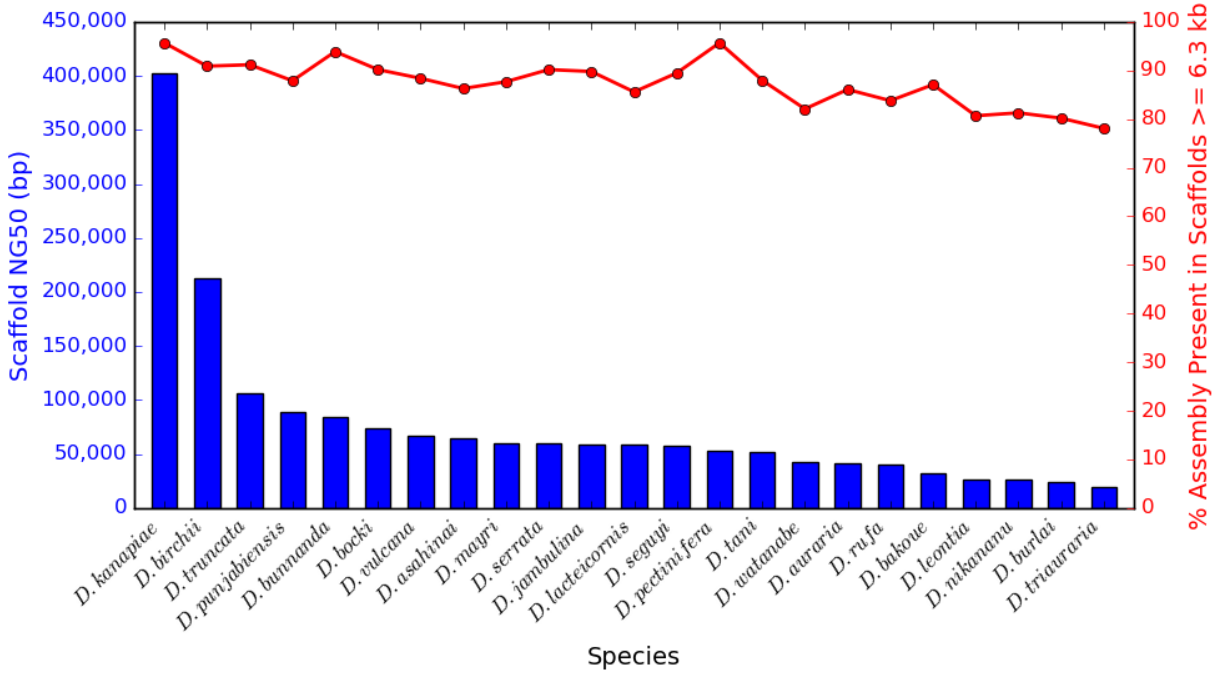


Figure 2.2. For all *montium* species, the vast majority of the assembly is present in at least gene-sized scaffolds, despite large differences in contiguity.
Based on annotations of the previously assembled *D. serrata* genome, the average gene length is up to 6.3 kb [39, 44]. For each *montium* species, the blue bar graph shows the scaffold NG50, and the red line graph shows the percentage of the total scaffold length (assembly) present in scaffolds that are at least 6.3 kb in length. Species are listed in order of scaffold NG50, starting with the longest.

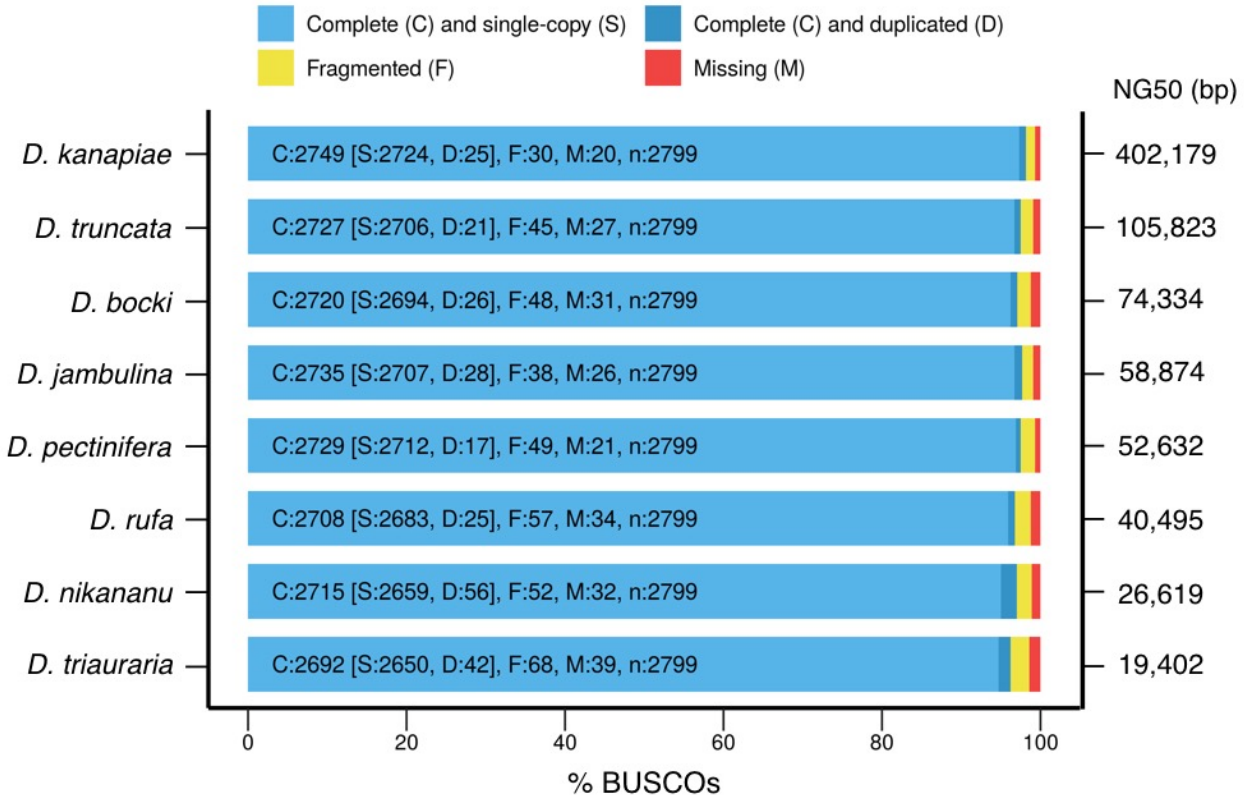


Figure 2.3. All *montium* assemblies contain a high percentage of known genes despite large differences in contiguity.
BUSCO [45, 46] assessment results for eight *montium* genomes representing a diversity of genomes / assemblies. The Dipteran BUSCO set contains 2,799 genes. For each assembly, the bar graph reports the number of BUSCOs that are complete and single-copy, complete and duplicated, fragmented, and missing. The scaffold NG50 for each assembly is shown on the right.

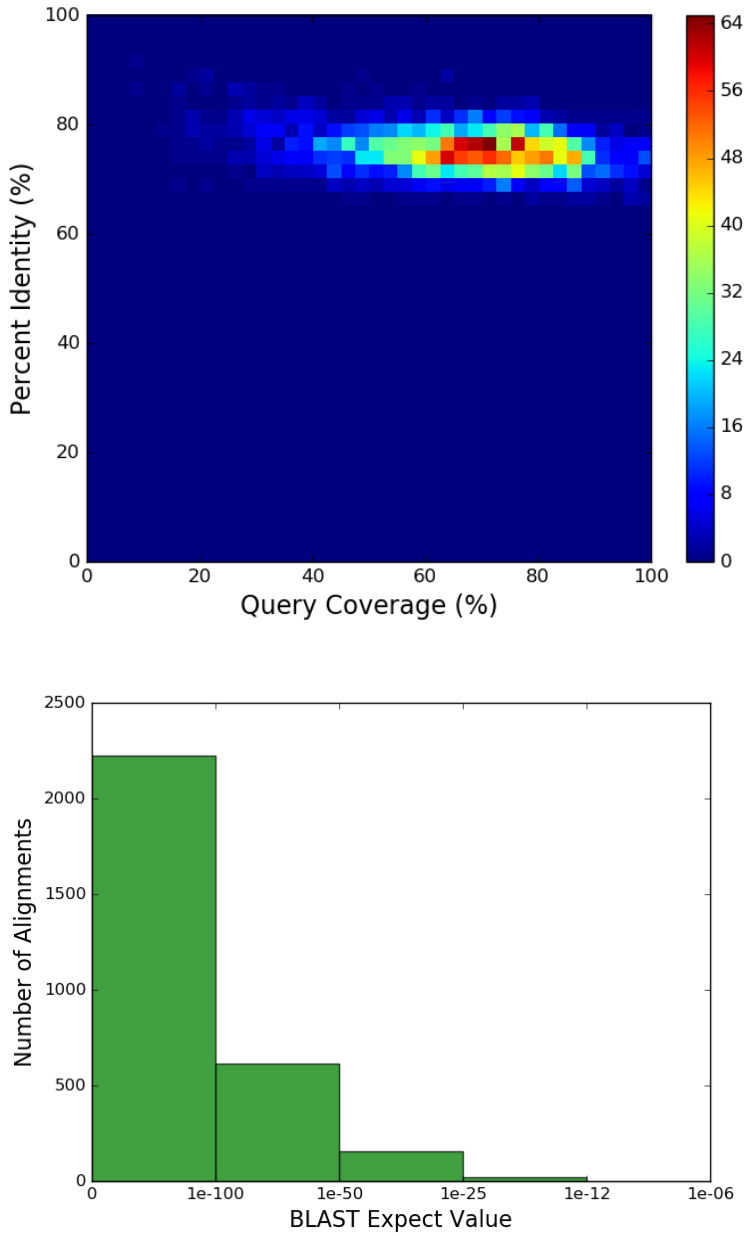


Figure 2.4. Thousands of orthologous *montium* enhancers can be identified by remapping *D. melanogaster* enhancer coordinates onto *montium* assemblies. Approximately 3,500 experimentally verified *D. melanogaster* enhancers from [52] were remapped onto the *D. lacteicornis* assembly using liftOver [51]. The *D. melanogaster* sequences were approximately 2 kb, and the putative *D. lacteicornis* orthologs were filtered to contain sequences between 1.5 kb and 3.5 kb. In total, 3,010 pairs of *D. melanogaster* and *D. lacteicornis* sequences were aligned using BLASTn [59]. *D. lacteicornis* was chosen for illustrative purposes because it is the assembly with the median scaffold NG50. A) 2D histogram showing query coverage and percent identity for 3,010 pairwise BLASTn [59] alignments. Query coverage is the percentage of *D. melanogaster* sequence that is aligned to *D. lacteicornis* sequence. B) Distribution of Expect values (E) for alignments in Part A.

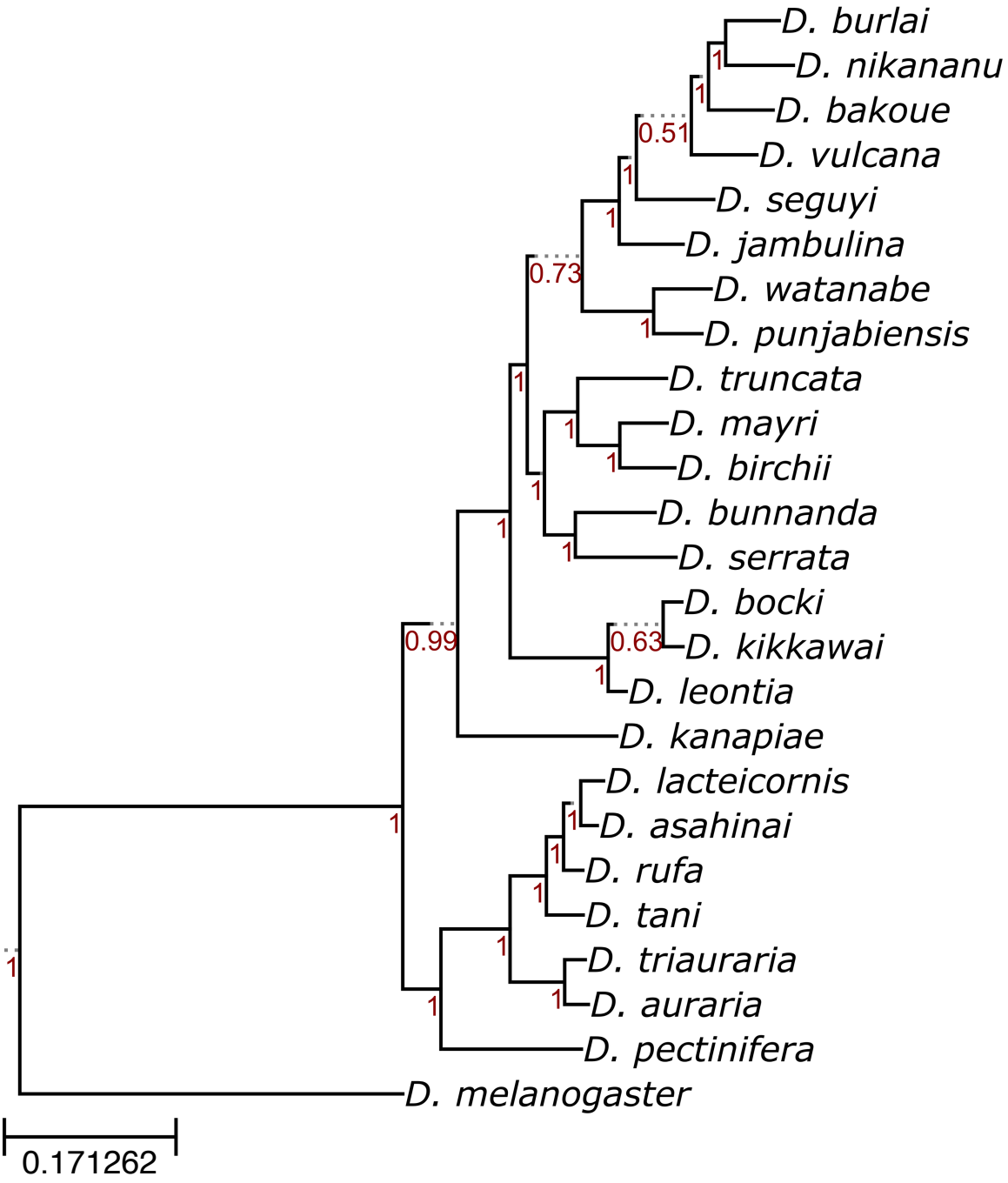


Figure 2.5. Maximum likelihood tree of the *Drosophila montium* subgroup based on 20 Bicoid-dependent enhancers.
Bicoid-dependent enhancers are from [57]. The phylogeny was reconstructed using RAxML [58] with the General Time Reversible (GTR) model of nucleotide substitution, the Gamma model of rate heterogeneity, and maximum likelihood estimates of base frequencies. All parameters were estimated independently for each of the 20 partitions. Branch support values (shown in red) are based on 100 searches using a rapid bootstrapping algorithm. Branch lengths are in substitutions per site.

Discussion

In this chapter, I described the creation of a novel genomic resource for the study of enhancer evolution. I sequenced and assembled 23 genomes from the *Drosophila montium* species subgroup, a large group of closely related species. I aligned each *montium* assembly to the extensively annotated *D. melanogaster* genome, and showed that my assemblies contain high percentages of known genes and enhancers.

The estimated genome sizes for my *montium* species range from 153 Mb to 225 Mb. This is consistent with the previously published *D. kikkawai* (164 Mb) [38] and *D. serrata* (198 Mb) [39] genomes.

To make this endeavor financially feasible, I sequenced a single, small-insert library for each species. This impacted the assemblies in a number of ways. The absence of long-distance information made the assemblies especially sensitive to repeats and high levels of heterozygosity. As a result, many of the assemblies are fragmented, and the scaffold NG50s vary widely based on genome / sample characteristics. The average scaffold NG50 is only 76 kb, and the total scaffold length of most assemblies is significantly shorter than the estimated genome size. Repeat content and heterozygosity act as opposing forces on the total scaffold length. In the absence of long-distance information, repeats form unresolvable structures in the de Bruijn graph. This leads to many breaks in the assembly, and an underestimate of the number of repeat copies in the genome. So as the number and size of repeats increases, so too does the gap between the estimated genome size and the total scaffold length. On the other hand, divergent haplotypes in a highly heterozygous sample might be assembled independently, and placed on different scaffolds. (Assemblers like Meraculous-2D [60] and Platanus [61] that are designed to handle high levels of heterozygosity typically require mate-pair libraries.) This increases the scaffold length, and artificially closes the gap between the estimated genome size and the total scaffold length. Finally, some assemblers can over-assemble the data and produce many small contigs / scaffolds, also known as “chaff” [62]. Unsurprisingly, the three assemblies that exceed their estimated genome size are all highly heterozygous. Even heterozygous samples can come up short though if the repeat content is large enough.

Just because most assemblies are fragmented, does not mean they are poor quality. Quite to the contrary, the BUSCO analysis showed that all assemblies, regardless of contiguity, contain at least 96 % of known single-copy Dipteran genes. I also used the whole-genome alignments to lift coordinates for thousands of *D. melanogaster* enhancers onto the *montium* assemblies. Pairwise alignments between *D. melanogaster* and *montium* sequences showed that this process correctly identified putative *montium* orthologs. Despite their modest scaffold NG50s, the *montium* assemblies are contiguous enough to study genes and enhancers.

All draft genomes contain assembly errors, and mine are no different. Most errors occur during scaffolding, or near the ends of contigs. I used REAPR [63] and Pilon [64] to

identify and correct as many errors as possible. These programs work best with large-insert libraries (which I didn't have), but they still made significant improvements. I also "phased" the assemblies so they represented the single majority haplotype - within the limits of a small-insert library. Small tandem alleles can also be a problem with highly heterozygous samples. Remaining tandem alleles in the *montium* assemblies can often be identified by the presence of single-N gaps.

In the future, efforts should focus on sequencing additional *montium* species. Five *montium* species were lost as a result of contamination in the laboratory (data not shown). Since my project began, several new species have also been collected. These species should be (re)acquired and sequenced. Despite the challenges, efforts should be made to intensively inbreed the lines for at least several generations prior to sequencing.

Any *montium* assembly can also be improved on an as-needed basis. For my purposes, the assemblies are generally contiguous and accurate enough to study enhancers. But if another researcher needs a high-quality draft genome, they could easily pair my short-read data with traditional mate-pair libraries or PacBio long-reads to generate a vastly more contiguous assembly that also includes most / all repeat copies.

Going forward, these genomes will enable me to study hundreds of enhancers / transcription factor bound regions across the *montium* subgroup. They are also a new and valuable resource for any researcher studying ecological, evolutionary, or behavioral questions using *montium* species.

Materials and Methods

Library Preparation and Sequencing

Fly lines for each *montium* species reported in Table 2.1 were gifts of Artyom Kopp and Michael Turelli, or were acquired from the *Drosophila* Species Stock Center.

For each species, DNA was extracted from three female flies using the QIAGEN QIAamp DNA Micro Kit. Sequencing libraries were constructed using the Illumina TruSeq DNA PCR-Free Kit for 350 bp inserts, and visualized on Agilent High Sensitivity DNA chips. Libraries were clustered on an Illumina HiSeq 2000 or HiSeq 2500 System, generating 100 bp paired-end reads. All sequencing was done at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. Multiple species were pooled on each lane in an effort to reach a sequencing depth of at least 30x coverage per species.

Read Exploration and Pre-Processing

Prior to assembly, read quality and genome / sample characteristics (e.g., genome size, repeat content, and heterozygosity) were explored using FastQC (v. 0.11.2) [65] and String Graph Assembler (SGA) Preqc (v. 0.10.15) [66].

Reads were adapter-trimmed for known Illumina adapters using BBDuk (BBMap v. 36.11) [67] with the options ktrim=r, k=23, mink=9, hdist=1, minlength=75, tpe=t, and tbo=t. The adapter-trimmed reads were then quality-trimmed to Q10 using BBDuk (which implements the Phred algorithm) with the options qtrim=rl, trimq=10, and minlength=51.

Read Decontamination

Two sequencing libraries were heavily contaminated with bacteria: *D. pectinifera* and *D. vulcana* (Figure 2.S1).

For *D. pectinifera*, I adopted a decontamination strategy similar to Kumar et al. [68]. The reads were first assembled using SOAPdenovo2 [69] with the options -K 49 and -R. Assembled scaffolds at least 1 kb in length were used to create a GC % vs. average *k*-mer coverage plot. Scaffolds with 35 <= GC % <= 66 and 40.5 <= average *k*-mer coverage <= 68 were considered candidate bacterial scaffolds. To avoid removing any *Drosophila* scaffolds, candidate bacterial scaffolds were aligned to sequences in NCBI's Nucleotide database using BLASTn (v. 2.2.31+) [59]. Candidate bacterial scaffolds that aligned to known bacterial sequences with bit scores greater than 214 were classified as genuine bacterial scaffolds. Finally, the original reads were aligned to the bacterial scaffolds using Bowtie 2 (v. 2.2.3) [70] with the option --local, and pairs of reads that aligned concordantly were removed prior to the subsequent primary assembly.

For *D. vulcana*, 10,000 reads were sampled from the R1 FASTQ file using seqtk sample (v. 1.0-r75-dirty) [71], and then converted to FASTA format using seqtk seq. High-GC % reads within the range 53 <= GC % <= 57 were identified, and a subset of 400 high-GC % reads were aligned to sequences in NCBI's Nucleotide database using BLASTn (v. 2.2.31+) [59]. This led to the identification of closely related bacterial (and yeast) genomes, which were combined into a single reference. Finally, the original reads were aligned to the reference using Bowtie 2 (v. 2.2.3) [70] with the option --local, and pairs of reads that aligned concordantly were removed prior to assembly.

Some libraries were also contaminated with highly abundant individual sequences, or groups of similar sequences. (For example, the per sequence GC % plot for *D. vulcana* also showed a low-GC % "spike" corresponding to an eight-bp simple sequence repeat (SSR) that was present in both the forward and reverse reads. The origin of the sequence was unclear.) Once the contaminating sequence was identified, matching sequences were removed from the reads using BBDuk (BBMap v. 36.11) [67] with the options k=75 and hdist=1.

Choosing an assembler

Exploration of the data using SGA Preqc (v. 0.10.15) [66] showed that the *montium* genomes / samples represented a diversity of genome sizes, heterozygosity levels, and sequencing error rates. Extensive tests were conducted to identify the assembler that performed the best across these diverse samples.

I tested the following assemblers: ABySS [72], MaSuRCA [40], Meraculous-2D [60], SOAPdenovo2 [69], SPAdes [73] / dipSPAdes [74], and Velvet [75]. The resulting assemblies were evaluated using a number of metrics, including contiguity statistics, REAPR [63], Feature Response Curves (FRC^{bam}) [76], BUSCO assessments [45, 46], and the scrutiny of individual enhancer sequences.

Primary assemblies

All genomes were assembled using MaSuRCA [40]. The assembler was provided with reads that had not been adapter-trimmed or quality-trimmed.

The configuration file for each species contained the insert-size mean and standard deviation for the corresponding sequencing library, as well as the following parameters: GRAPH_KMER_SIZE = auto, USE_LINKING_MATES = 1, CA_PARAMETERS = cgwErrorRate=0.15, KMER_COUNT_THRESHOLD = 1, and SOAP_ASSEMBLY=0. The Jellyfish hash size (JF_SIZE) was set to the product of the estimated genome size and estimated coverage.

Post-assembly pipeline

For each assembly, the MaSuRCA assembler created a small number of scaffolds with massive gaps (tens of kb in length). Given the insert-sizes of the sequencing libraries (~ 350 bp), these gaps had to be erroneous. Therefore, scaffolds were split on any gap that was unreasonably large relative to the insert-size of the library. (Typically around 300 - 600 bp, depending on the library.)

REAPR (v. 1.0.18) [63] was used to identify errors in the assemblies, and to generate new “broken” assemblies that were split on errors occurring over gaps. Errors within contigs were hard-masked with Ns. The command reapr smaltmap was used to align adapter-trimmed reads to the assemblies, and reapr pipeline generated the broken assemblies. Sequences starting with “REAPR_bin” (i.e., the original unmasked sequence) were later filtered from the broken assemblies.

Gaps in the assemblies were closed using a two-step process with adapter-trimmed and quality-trimmed reads. The first round of gap closing was performed using GapCloser (v. 1.12) [69]. This also helped to identify tandem alleles (a type of mis-assembly), which GapCloser left as single-N gaps. The second round was done using Sealer (abyss-sealer v. 2.0.2) [77], with the option -P 10. For each assembly, “*k* sweeps” typically ranged from *k*=80 to *k*=30 (in decrements of 10), but varied if Sealer became stuck on a

given *k*-mer size. After two rounds of gap closing, the *D. triauraria* assembly still contained more than 2,000 single-N gaps. The remaining single-N gaps (and associated flanking sequence) were hard-masked with 300 Ns, and Sealer was run a second time using the above settings. This potentially extended the flanking sequence extracted by Sealer beyond the boundaries of the original tandem allele, thereby making it possible to find a connecting path in the graph.

The assemblies were further improved using Pilon (v. 1.22) [64] with the options --fix all,amb, --diploid, and --mingap 1. This attempted to fix SNPs, indels, local misassemblies, and ambiguous bases, as well as fill remaining gaps.

A detailed inspection of aligned reads showed that many scaffolds were mosaics of multiple haplotypes present in the original sample. This was a significant problem for highly heterozygous samples, as it could create recombinant enhancer sequences not found in nature. My goal therefore was to create a "phased" assembly that reflected the majority haplotype at each variable locus.

Pilon (v. 1.22) [64] was run a second time on the improved assemblies, but this time it was used as a variant detection tool to generate VCF files (option --vcf). Variants in the VCF file were phased using the read-based phasing tool WhatsHap (v. 0.14.1) [78], with the options phase, --ignore-read-groups, --tag=PS, and --indels. BCFtools (v. 1.5) [79] with the options view, --phased or --exclude-phased was then used to create VCF files with only phased or un-phased variants. To facilitate parsing of the phased VCF file, a sequence dictionary was first created with the tool CreateSequenceDictionary from Picard (v. 2.12.1-SNAPSHOT) [80], and then VariantsToTable from the Genome Analysis Toolkit (GATK) (v. nightly-2017-09-13-g315c945) [81] was used to create a tab-delimited table of variants. For each phase set in the table, the majority haplotype was determined based on the cumulative read count of variants on each haplotype (A or B), with indels weighted half as much as SNPs (because of alignment issues with indels). Phased variants that were present on majority haplotypes were retained. For un-phased variants, the majority allele was retained. A new VCF file was then created using only the retained phased and un-phased variants. Finally, BCFtools consensus was used to create a new "phased" assembly by applying the variants in this VCF file to the original "un-phased" assembly.

Lastly, any remaining ambiguous bases (except N) were randomly assigned to a single base, and scaffolds shorter than 1 kb in length were removed.

BUSCO analysis

The assemblies were searched for known genes using BUSCO (v. 3.0.2) [45, 46] with the profile library diptera_odb9. The following options were specified in the configuration file: mode = genome, evalue = 1e-3, limit = 3, and long = False. The BUSCO plot was constructed using the included script generate_plot.py.

Whole-genome alignment pipeline

Each *montium* genome was individually aligned to the *D. melanogaster* genome (NCBI Assembly ID: 202931, Release 6 plus ISO1 MT / UCSC Genome Browser Assembly ID: dm6). Target and query genomes were soft-masked using RepeatMasker (v. open-4.0.7) [47] and Tandem Repeat Finder (TRF) (v. 4.04) [48], with the options -s, -species drosophila, -gccalc, -nocut, and -xsmall. Pairs of genomes were aligned using LASTZ (v. 1.03.73) [49], with the following options from Chen et al. [38]: target_genome[multiple], --masking=50, --hspthresh=2200, --ydrop=3400, --gappedthresh=4000, --inner=2000, and --format=axt. The LASTZ alignments were then processed using a series of programs described in detailed by Kent et al. [50]. Briefly, FASTA files for the target and query assemblies were converted to 2bit format using faToTwoBit. Files containing chromosome / scaffold lengths were created using faSize with the option -detailed. Chains were built using axtChain with the option -linearGap=medium. The chains were then filtered using chainPreNet, and ordered into nets using chainNet with the option -minSpace=1. Nets were annotated using netSyntenic. Finally, subsets of chains found in nets were extracted using netChainSubset, creating liftOver chain files.

Identification of orthologous *montium* enhancers

Kvon et al. [52] previously described a large set of DNA fragments (tiles) that drive expression in the *D. melanogaster* embryo. A total of 3,457 tiles were positive for enhancer activity and PCR-verified. *D. melanogaster* coordinates were lifted onto each *montium* assembly using liftOver [51] with the options -minMatch=0.1 and -multiple. (The liftOver program was originally written to remap coordinates between assemblies of the same species. However, it is routinely used for interspecies lifts, and in my experience, it performed well.) For each *montium* species, the output was filtered to only include sequences between 1.5 kb and 3.5 kb. Pairs of *D. melanogaster* and putatively orthologous *montium* sequences were aligned using BLASTn (v. 2.2.31+) [59] with the options -task blastn-short, -dust no, -evalue 0.00029, -reward 2, and -outfmt 6.

Phylogeny reconstruction

Chen et al. [57] previously described 66 Bicoid-dependent enhancers in *D. melanogaster*. I identified a subset of 20 enhancers that were present in all 23 *montium* assemblies and spread out across the genome. I extracted coordinates for 1.4 kb regions centered on each enhancer, and remapped them onto each *montium* assembly using liftOver [51], with the options -minMatch=0.1 and -multiple. Sequences for each enhancer were aligned using MAFFT (v. 7.407) [82-84] with the linsi preset, along with the options --anysymbol, --ep 0.123, and --nuc. The *Drosophila montium* species subgroup phylogeny was reconstructed using maximum likelihood, as implemented in RAxML (v. 8.2.12) [58] with the General Time Reversible (GTR) model of nucleotide substitution, the Gamma model of rate heterogeneity, and maximum likelihood estimates of base frequencies. All parameters were estimated independently for each of

the 20 partitions. Branch support values were calculated based on 100 searches using a rapid bootstrapping algorithm. The full list of options included -p 12345, -x 12345, -# 100, -m GTRGAMMA, --no-bfgs, and -f a. *D. melanogaster* was specified as the outgroup using the -o option.

Supporting Information

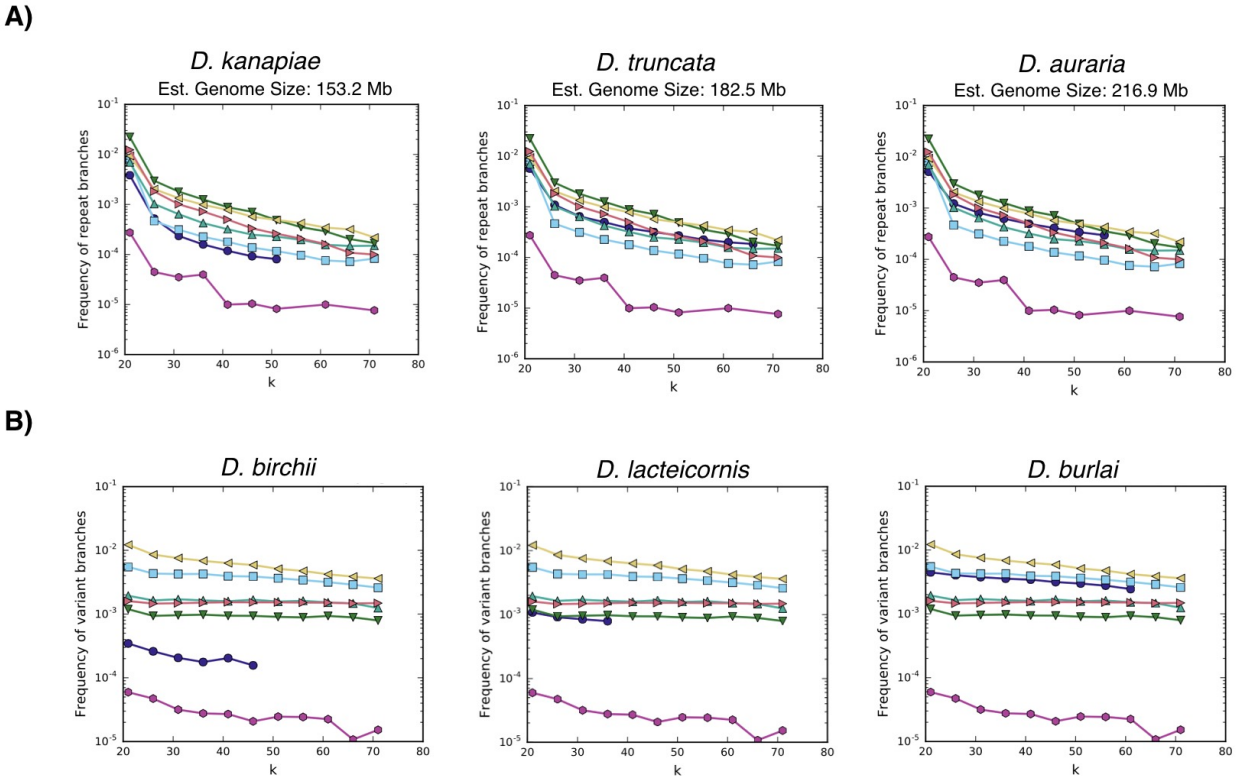


Figure 2.S1. Diverse genome sizes, repeat contents, and heterozygosity levels across *montium* species / samples.
Figures and genome size estimates are from SGA Preqc [66]. A) Repeat content and genome size estimates for three *montium* species. The frequency of repeat branches in the de Bruijn graph is shown as a function of *k*-mer size (*k*=21 to *k*=71). The fly genome is plotted in dark purple. Each figure also includes six diverse reference genomes for comparison: bird (light blue), fish (aqua), human (dark green), oyster (yellow), snake (light red), and yeast (pink). Genome size estimates are based on the *k*-mer frequency spectrum, using a method that also corrects for error *k*-mers. Note that larger genomes have more repeat branches. B) Heterozygosity levels for three *montium* species / samples. The frequency of variant branches in the de Bruijn graph is shown as a function of *k*-mer size. The fly genome is plotted in dark purple, along with the previously described references.

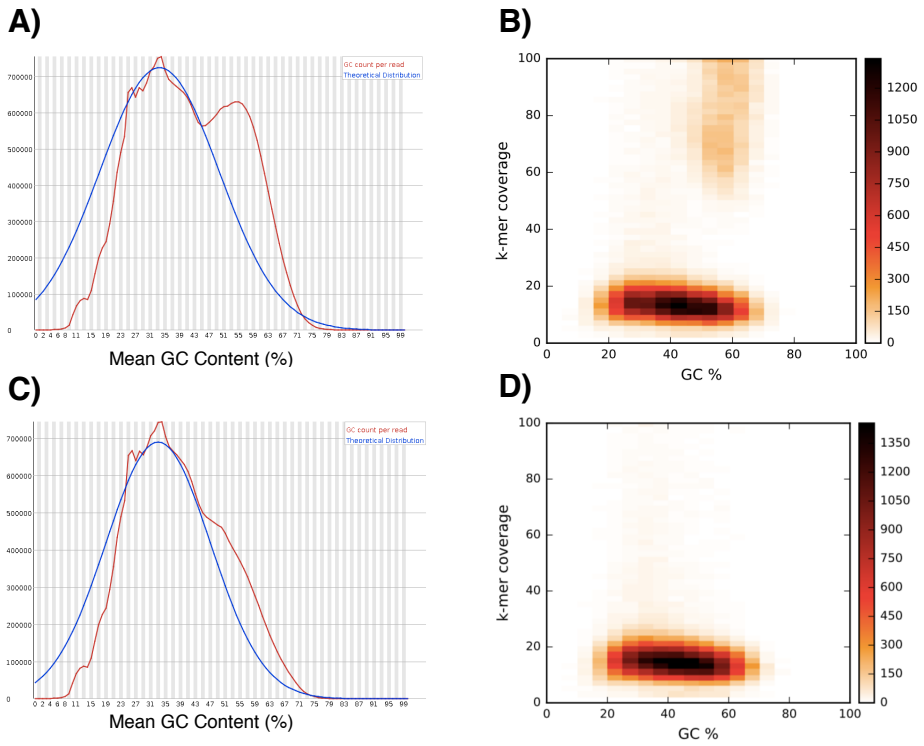


Figure 2.S2. The *D. pectinifera* library was heavily contaminated with high-GC % bacteria. Per sequence GC plots are from FastQC [65], and GC % bias plots were generated by SGA Preqc [66]. A) Per sequence GC plot for the contaminated reads shows a broad, high-GC % peak. Such secondary peaks are usually indicative of a contaminating genome. B) GC % bias plot of the contaminated reads shows the presence of high-GC %, high-coverage *k*-mers that are distinct from the fly genome *k*-mers. C) Per sequence GC plot for the decontaminated reads. In total, 17.3 % of the reads were removed. D) GC % bias plot for the decontaminated reads.

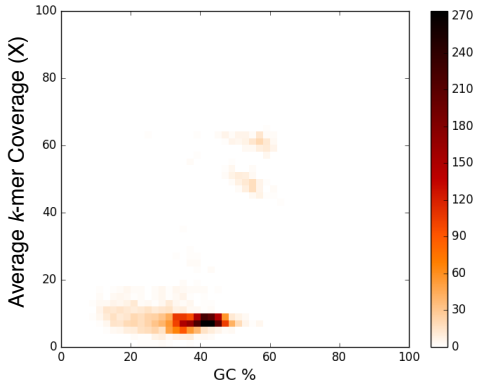


Figure 2.S3. A preliminary *D. pectinifera* assembly identifies high-GC %, high-average *k*-mer coverage bacterial scaffolds.
The contaminated reads were assembled using SOAPdenovo2 [69]. The candidate bacterial scaffolds form two distinct clusters of high-GC %, high-average *k*-mer coverage scaffolds that are distinct from the fly genome scaffolds. Reads that aligned to bacterial scaffolds were removed (see Figure 2.S2 Parts C and D) prior to the primary MaSuRCA assembly.

Table 2.S1. Remapping experimentally verified *D. melanogaster* enhancers [52] **onto *montium* assemblies.**

Species	Attempted Lifts	Successful Lifts	Number of Putative Orthologs (1.5 - 3.5 kb)	Percentage of Putative Orthologs (%)
D. kanapiae	3,457	3,449	3,117	90.2
D. birchii	3,457	3,449	3,134	90.7
D. truncata	3,457	3,449	3,095	89.5
D. punjabiensis	3,457	3,449	3,065	88.7
D. bunnanda	3,457	3,447	2,989	86.5
D. bocki	3,457	3,449	3,103	89.8
D. vulcana	3,457	3,448	3,058	88.5
D. asahinai	3,457	3,450	3,023	87.4
D. mayri	3,457	3,449	3,159	91.4
D. serrata	3,457	3,447	2,970	85.9
D. jambulina	3,457	3,450	3,032	87.7
D. lacteicornis	3,457	3,451	3,010	87.1
D. seguyi	3,457	3,444	3,057	88.4
D. pectinifera	3,457	3,449	3,055	88.4
D. tani	3,457	3,451	3,030	87.6
D. watanabe	3,457	3,449	2,969	85.9
D. auraria	3,457	3,448	3,019	87.3
D. rufa	3,457	3,452	3,020	87.4
D. bakoue	3,457	3,451	3,074	88.9
D. leontia	3,457	3,444	3,030	87.6
D. nikananu	3,457	3,451	3,012	87.1
D. burlai	3,457	3,450	3,057	88.4
D. triauraria	3,457	3,448	2,945	85.2

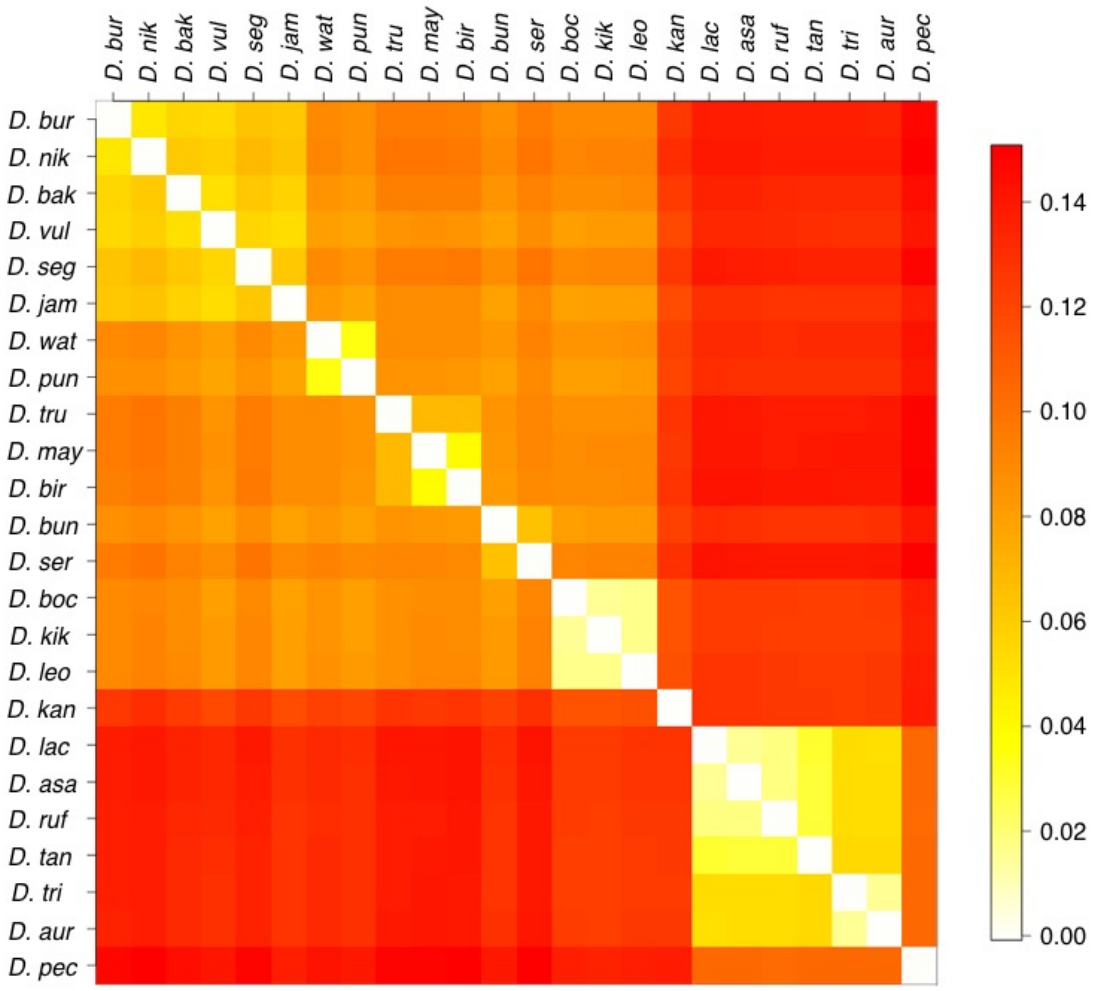


Figure 2.S4. Distance matrix for the *Drosophila montium* subgroup based on 20 Bicoid-dependent enhancers.
Bicoid-dependent enhancers are from [57]. Distances were calculated using the function dist.dna from the package ape (v. 5.2) [85], with the options model="TN93" and gamma=TRUE. This uses a model developed by Tamura and Nei [86].

Chapter 3: Studying binding site turnover in the *Drosophila montium* subgroup

Abstract

Enhancers often exhibit remarkable functional conservation despite extensive sequence divergence. Understanding the rules that allow enhancers to diverge, while simultaneously maintaining function, is a major challenge. If I can observe binding site turnover within the *montium* subgroup, then I have an unprecedented ability to identify key changes and events (along with their context) before they are obscured by additional mutations. I start by comparing the *eve* stripe 2 enhancer across 24 *montium* species and *D. melanogaster*. Importantly, I can observe one or two changes in transcription factor binding sites between closely related species. I also show that I can observe previously described patterns of binding site conservation and proximity. Next, I showed how patterns of (apparent) conservation and variation within the *montium* subgroup could be used to direct targeted mutagenesis experiments, and to inform models of enhancer grammar. To study binding site turnover on a large scale, I investigated the top ChIP peaks for the anterior morphogen Bicoid, the gap gene Krüppel, and the pioneer factor Zelda. I treated groups of orthologous binding site scores as continuous traits, reconstructed ancestral binding scores at each node of the tree, and then calculated score changes along each branch of the tree. For all three factors, gain events were more likely to occur along branches of the tree that also lost a binding site. This was true for both non-conserved and conserved sites, and most changes were statistically significant. However, when the analysis was repeated using shuffled matrices, the results were similar - though generally not as large - leaving me unable to conclude these were meaningful changes in transcription factor binding. The analysis was likely confounded by a significant fraction of non-functional *montium* sequences. Additional complications included the use of the Brownian motion model of evolution, and ancestral character estimation with a single species tree in the presence of widespread gene tree / species tree incongruence. Future analyses should focus on a set of high-quality bound regions, and ancestral states should be reconstructed using leap / jump models and individual gene trees.

Introduction

In the previous chapter, I described the sequencing and assembly of 23 genomes from the *Drosophila montium* subgroup, a large group of closely related species. I showed that my assemblies contain high percentages of known genes and enhancers. I also aligned each *montium* assembly to the *D. melanogaster* genome, and reconstructed the *montium* subgroup phylogeny using 20 enhancer regions.

In this chapter, I leverage the resources I created in the *montium* subgroup to address fundamental questions about enhancer evolution. Enhancers frequently show conservation of function despite extensive sequence divergence. This has been demonstrated in detailed analyses of individual enhancers, such as *eve* stripe 2 [17-19, 21] and *sparkling* [24], along with massively parallel assays like STARR-seq [27]. One explanation for this observation is binding site turnover: the gain and loss of sites for the same factor over evolutionary timescales [27, 87-91]. While such studies have provided dramatic examples - for example, functional conservation of the *eve* stripe 2 enhancer across 100 million years of evolution despite near-complete rearrangement of binding sites [21] - they also highlight key challenges. When divergence times are large, so many mutational events have occurred that it's difficult to identify the compensatory changes that actually matter, and to order key events. By working with a large number of closely related species in the *montium* subgroup, I can study enhancers at the earliest stages of divergence. If binding site turnover is observable within the *montium* subgroup, then I have an unprecedented ability to identify key changes and events (along with their context) before they are obscured by additional mutations.

In this chapter, I start by studying the well-described *eve* stripe 2 enhancer across 25 *Drosophila* species. Crucially, I show that individual changes in transcription factor binding sites are visible between closely related species. To study binding site turnover on a large scale, I investigated the top ChIP peaks for the factors Bicoid, Krüppel, and Zelda. These loci strongly overlap with known enhancers. I predicted binding sites within each bound-region, and looked for correlated changes in binding site scores along each branch of the species tree. While I did see a marked enrichment for binding score increases along branches that had also experienced loss events, these patterns were ultimately similar to those observed using shuffled matrices.

Results

Initial investigations of the *eve* stripe 2 enhancer

I start by investigating the extensively studied *eve* stripe 2 enhancer in the *montium* subgroup. Coordinates for the minimal enhancer in *D. melanogaster* [13] were remapped onto each *montium* species using liftOver [51], and binding sites for the known regulators Bicoid, Giant, Hunchback, Krüppel [9, 13], Sloppy Paired 1 [15], and Zelda [16] were predicted using PATSER [92], with ln(*p*-value) cutoffs from [21]. Predicted binding sites were then mapped onto a 25-species multiple sequence alignment (MSA) generated by MAFFT [82-84]. Figure 3.1 is a binding site plot for the *eve* stripe 2 enhancer across 24 *montium* species and *D. melanogaster*.

Detailed inspection of the plot shows that closely related species differ by only one or two binding sites. This is important, because it means that when key changes do occur, I can observe the context in which those changes happened, before they are obscured by additional mutational events.

The overall arrangement of the *eve* stripe 2 enhancer appears to be highly conserved across all species (i.e., the *montium* subgroup and *D. melanogaster*). This is not surprising given the modest divergence times [20]. Twenty two orthologous binding sites are present in *D. melanogaster* and at least 21/24 *montium* species, including overlapping Bicoid-Krüppel, Bicoid-Giant, and Hunchback-Giant sites thought to play an important role in the regulation of the enhancer [9, 12-14] (but see also [22]). Conversely, three binding sites - a Bicoid site at MSA position 433, and two Krüppel sites at positions 578 and 634 - appear to be highly conserved across the *montium* subgroup, but missing in *D. melanogaster*. In the case of the two Krüppel sites, they tightly flank a cluster of highly conserved Bicoid, Hunchback, and Sloppy Paired 1 sites. Interestingly, this cluster is also present in *D. melanogaster*, but instead of flanking Krüppel sites, it contains a single Krüppel site overlapping one of the Bicoid sites. This suggests that the location of a Krüppel site relative to these Bicoid, Hunchback, and Sloppy Paired 1 sites might be functionally important - at least in the context of the overall arrangement of sites present in *D. melanogaster* and the *montium* subgroup.

I turn now to investigating previously described patterns of binding site conservation and arrangement in the *eve* stripe 2 enhancer [21]. The results are shown in Figure 3.2. Each group of orthologous binding sites was assigned a simple conservation score based on the fraction of *montium* species it was observed in: high, moderate, low, or none. Within the *montium* subgroup, most sites show either high conservation or no conservation, with little in between (Part A). Given the relatively large number of species, there are many novel sites, hence the excess of sites showing no conservation. This is consistent with a core group of binding sites that are conserved across the entire subgroup, along with the acquisition of new sites within small subclades or individual species. Each group of orthologous binding sites was also classified based on its proximity to sites for different factors, using the nomenclature from [21]. Overlapping binding sites share at least one base pair; close binding sites are within ten base pairs, but do not overlap; and isolated binding sites are more than 10 base pairs apart. There is a marked enrichment for overlapping binding sites: nearly 70 % of all binding sites overlap a site for a different factor (Part B).

To better understand the relationship between conservation scores and binding site proximity, Figure 3.2 also shows proximity as a function of conservation (part C), and conservation as a function of proximity (part D). Highly conserved sites are almost exclusively overlapping, as are moderately conserved sites (but in the latter case, the sample size is very small). Sites with low or no conservation also tend to be overlapping, but also include a sizable fraction of sites that are close or isolated. Similarly, overlapping binding sites are either highly conserved or show no conservation. The fraction of highly conserved binding sites decreases in sites that are close and isolated; whereas the fraction of sites showing low or no conservation increases. Given these dynamics, I wanted to better understand proximity for novel sites, which by definition are present in only one species. Interestingly, when a new site arises, it is far more likely to overlap an existing site (for a different factor), than to be close or isolated (Part E).

Historically, laborious and time consuming structure-function experiments were used to dissect the *eve* stripe 2 and *sparkling* enhancers. While these approaches yielded key insights, they are impractical for large numbers of enhancers. One of the advantages of sequencing a large number of closely related species is that observed patterns of conservation and variation can be used to direct targeted mutagenesis experiments, and to inform models of enhancer grammar. For example, overlapping activator and repressor sites are thought to play an important role in the function of the *eve* stripe 2 enhancer [13]. In this context, selection acts directly to preserve the arrangement of overlapping sites. However, simulations have shown that long-lived pairs of overlapping activator and repressor sites can also arise when selection acts to preserve the overall number of sites - not their specific arrangement - in the context of a deletion bias [22].

The extreme 5' end of the minimal enhancer contains what appears to be a highly conserved cluster of overlapping Giant, Krüppel, and Zelda sites. In *D. watanabe*, a pair of adjacent substitutions decreased the strength of the Krüppel and Zelda sites. The predicted binding score for Krüppel dropped by more than seven points, while the score change for Zelda was much smaller (only 0.3 points), but enough to lower the ln(*p*-value) below the plotting cutoff. (This highlights a limitation of any binding site plot: a site can always be present just beneath the plotting threshold.) Intriguingly, *D. watanabe* also acquired new Krüppel and Zelda sites elsewhere in the enhancer. The new Krüppel site is located 40 bp away, and actually appears to have arisen in the ancestor of *D. watanabe* and *D. punjabiensis*. If this new Krüppel site compensates for the loss of another site in *D. watanabe*, then its targeted ablation should yield aberrant *eve* stripe 2 expression. Such a result would also illustrate an important point: while this looks like a highly conserved cluster of overlapping binding sites, the relative arrangement of sites might not be that important, so long as they are present somewhere in the enhancer [22].

In other cases, large changes are present without any obvious compensatory changes. For example, the extreme 3' end of the minimal enhancer contains a cluster of overlapping Hunchback-Giant-Sloppy Paired 1 and Bicoid-Krüppel sites. A single substitution in *D. pectinifera* dropped binding scores for Hunchback and Sloppy Paired 1 by 5.78 and 4.97 points, respectively. These sites are present in *D. melanogaster* and all other *montium* species, and overlapping activator / repressor sites are thought to play an important functional role in the enhancer [13]. However, *D. pectinifera* has not acquired any new Hunchback or Sloppy Paired 1 sites elsewhere in the enhancer. This raises several possibilities: 1) Despite the appearance of conservation, the Hunchback / Sloppy Paired 1 sites are unimportant, 2) Compensatory mutations involving Hunchback and Sloppy Paired 1 are present outside of the minimal enhancer, and 3) The loss of a (presumably functional) binding site is not always associated with the gain of a new site for the same factor. In the latter case, compensatory mutations might involve distance changes for existing sites, or the acquisition of new sites for different factors (as reported for the *sparkling* enhancer [24]).

Binding site turnover in Bicoid, Krüppel, and Zelda-bound regions

Next, I move on to studying binding site turnover across hundreds of enhancers / bound regions. To study binding site dynamics, I needed a large set of sequences with well-defined transcription factor binding. I investigated the top ChIP peaks from *D. melanogaster* for the anterior morphogen Bicoid (n=619) [93], the gap gene Krüppel (n=1,000) [93], and the pioneer factor Zelda (n=998) [94]. The top ChIP peaks are not strictly synonymous with enhancers, but there is strong overlap. Kvon et al. [52] identified roughly 3,500 embryonic enhancers by tiling across 15 % of the non-coding, non-repetitive *D. melanogaster* genome. The overlap between ChIP peaks and experimentally verified enhancers is 162/619 (26 %) for Bicoid; 245/1,000 (25 %) for Krüppel; and 190/998 (19 %) for Zelda. Given these results, it's likely the vast majority of top ChIP peaks intersect enhancers within the *D. melanogaster* genome.

Coordinates for bound regions in *D. melanogaster* were remapped onto each *montium* species using liftOver [51]. The *montium* sequences were then filtered in an attempt to remove sequences / regions that were not conserved relative to *D. melanogaster*. Significant changes in the size of an orthologous bound region might indicate changes in binding, so such sequences were removed. After removing individual short / long sequences, entire regions missing four or more *montium* species were removed. After filtering, there were 393 Bicoid-bound regions (9,528 sequences), 576 Krüppel-bound regions (13,865 sequences), and 528 Zelda-bound regions (12,872 sequences).

Mapping changes in binding site scores onto the species tree

One approach to studying binding site dynamics would be to predict sites using a *p*-value cutoff, and then treat the presence or absence of a site as a discrete trait. Characters could then be mapped onto the tree using parsimony or stochastic mapping. I adopted a different approach, in an attempt to include weaker sites, and to observe changes in binding strength in granular detail.

I predicted binding sites in each sequence using PATSER [92] with a score cutoff of zero. The binding site score is the sum of weights in the position weight matrix (PWM). Scores greater than zero indicate the sequence is more likely to be an instance of the motif than the background. This is generally more permissive than most *p*-value cutoffs. Predicted binding sites were then mapped onto multiple sequence alignments (MSAs) created by MAFFT [82-84], and clustered into groups of orthologous binding sites using a custom algorithm. Given the relatively large number of species (n=25), and the prevalence of low-complexity sequence in non-coding regions, alignment error was a significant challenge. Left uncorrected, such errors could create the false appearance of paired gain / loss events. I then treated each group of orthologous binding sites as a continuous trait, and reconstructed ancestral binding site scores at each internal node of the tree using maximum likelihood and the Brownian motion (BM) model of evolution, as implemented in the function anc.ML from the package phytools [95]. From there, I inferred increases / decreases in binding scores along each branch of the tree by taking

the difference between parent - child nodes. Overall, I think this was a more realistic approach than only looking at all-or-nothing gain / loss events on the tree. It also maximized my ability to observe binding site changes between closely related species. (Later on, for simplicity, I sometimes talk about “gains” and “losses”, but use the term to refer to both true gain / loss events, as well as changes in binding scores that do not completely eliminate a site.)

I started by looking at the magnitude and frequency of binding site changes across all branches of the tree. Figure 3.3 shows the distribution of binding score changes for Bicoid, Krüppel, and Zelda. The pattern is the same for all three factors. Most score changes are small, and the number of observed changes decreases exponentially as the magnitude of the change increases. Interestingly, there are twice as many gain events as loss events.

Correlated gain - loss changes along branches of the tree

If selection acts to preserve the total number of binding sites for a given factor in an enhancer / bound region, then compensatory gain - loss events should be correlated on the tree. (Simulations of the *eve* stripe 2 enhancer showed that such a simple rule could recapitulate well-described features such as clustered and overlapping binding sites [22].) When a binding site is lost (or undergoes a large decrease in binding score) along a given branch, are gain events (or increases in binding score) more likely to occur along the same branch, or its parent branch? For each transcription factor and branch on the tree, I divided bound regions into two groups: regions where the score of at least one binding site decreased by a factor-specific threshold, and regions with no such decrease. For each bound region, I then summed the total score increase along the original branch and its parent branch. Finally, I calculated the average score increase under both conditions (loss / no loss).

Figure 3.4 shows results for Bicoid, Krüppel, and Zelda-bound regions. Each branch on the tree is color-coded based on the ratio of the average score increase under the conditions loss / no loss. Darker reds indicate bigger differences. If a branch was too short for a meaningful comparison, it was colored black. The *melanogaster* branch, and the branch leading to the *montium* clade, were also excluded from the analysis and colored black. For Bicoid and Krüppel-bound regions, on average, cumulative binding scores increase by more than 40 % with a loss, compare to no loss. All branches (n=27) are statistically significant (Welch’s t-test, $p < 0.05$) for Bicoid, and 19/24 branches are significant for Krüppel. For Zelda-bound regions, binding scores increase by approximately 65 % with a loss, compare to no loss. Changes for 22/30 branches are statistically significant.

As a control, I shuffled the columns of the original matrices, and repeated the entire analysis. The results are similar, but the differences are generally smaller. When there’s a loss event, on average, cumulative binding site scores increase by 33 % within Bicoid-bound regions, 24 % within Krüppel-bound regions, and 40 % within Zelda-bound regions. Many of these differences are also statistically significant: 23/27 branches for

Bicoid-bound regions, 17/24 branches for Krüppel-bound regions, and 10/27 branches for Zelda-bound regions (Welch's t-test, $p < 0.05$).

Correlated gain - loss changes for conserved binding sites along branches of the tree

Next, I repeated the above analysis, but focused exclusively on the loss of (apparently) conserved binding sites. If the loss of a conserved site has a larger impact on function, then it might be more likely to be associated with compensatory changes. When a conserved site is lost (or undergoes a large decrease in binding score) along a given branch, are gain events (or increases in binding score) more likely to occur along the same branch, or its parent branch? Since changes in conserved binding sites are rare, I could not repeat the branch-level comparisons detailed above, while also maintaining adequate sample sizes. Instead, I controlled for branch length and created a single set of branches / bound regions that had lost a conserved site. I then created a control set by randomly sampling branches / bound regions that had not lost a conserved site, while maintaining the same composition of branches / bound regions. (For example, if the *D. asa* branch lost a conserved binding site in 15 bound regions, then the control set included a random sample of 15 bound regions where the *D. asa* branch did not lose a conserved site.) To control for sampling error, I created 100 randomly sampled control sets.

Figure 3.5 shows results for Bicoid, Krüppel, and Zelda-bound regions, for both the original matrix, and a shuffled matrix. When a conserved Bicoid or Krüppel site is lost along a branch, scores for other binding sites are more likely to increase along that branch as well, on average, compared to branches / bound regions that did not lose a conserved site. These differences are statistically significant, and do not depend on the control set used. However, from a practical standpoint, these differences are largely trivial, and shuffled matrices show similar patterns. For Zelda-bound regions, the relative difference is somewhat larger, but still small. Once again, the difference is statistically significant, and robust to various control sets. A shuffled Zelda matrix yields smaller differences compared to the original matrix, only half of which are statistically significant.

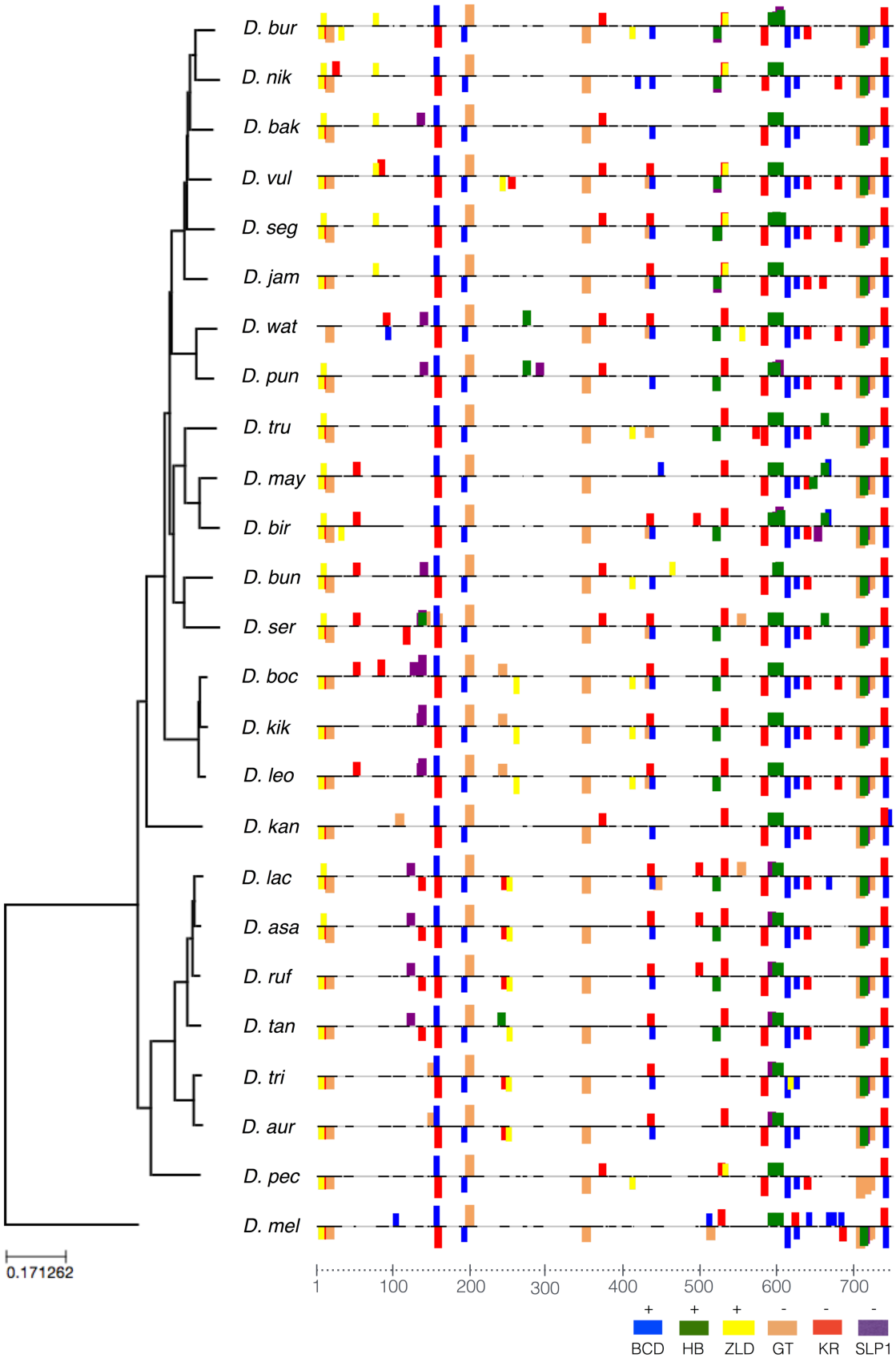
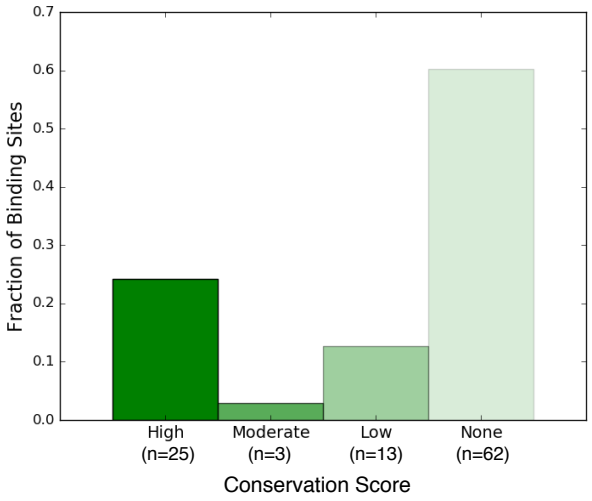
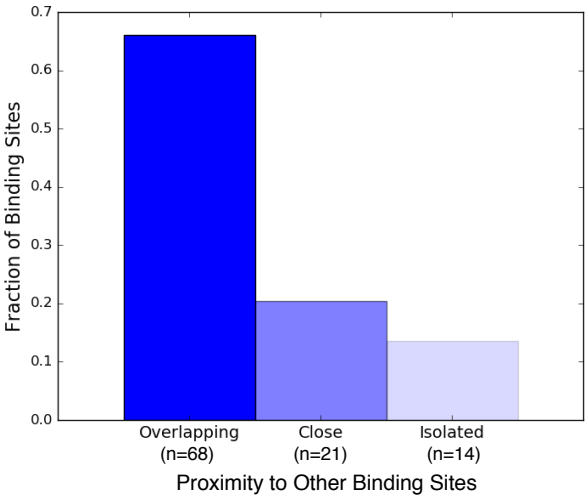

D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel
0.171262
1
100
200
300
400
500
600
700
+
+
+
-
-
-
BCD
HB
ZLD
GT
KR
SLP1

Figure 3.1. The arrangement of binding sites in the *eve* stripe 2 enhancer appears to be highly conserved across the *montium* subgroup and *D. melanogaster*, and individual changes are visible between closely related species.
Coordinates for the minimal *eve* stripe 2 enhancer in *D. melanogaster* [13] were remapped onto each *montium* assembly using liftOver [51]. Binding sites for the regulators Bicoid, Giant, Hunchback, Krüppel [13], Sloppy Paired 1 [15], and Zelda [16] were predicted in each *montium* species using PATSER [92], with the ln(*p*-value) cutoffs from [21]. Predicted binding sites were then mapped onto a 25-species multiple sequence alignment (MSA) generated by MAFFT [82-84]. In the plot, each binding site is represented by a color-coded rectangle, the height of which is proportional to the ln(*p*-value) of the predicted site. Stronger binding sites are represented by taller rectangles. Within each species, binding sites above the line were predicted on the positive strand, while sites below the line were predicted on the reverse strand. Aligned sequence in each species is represented by black lines; gaps are shown in gray. MSA coordinates are shown at the bottom of the alignment. A species tree based on 20 Bicoid-dependent enhancers from Chapter 2 is shown on the left.

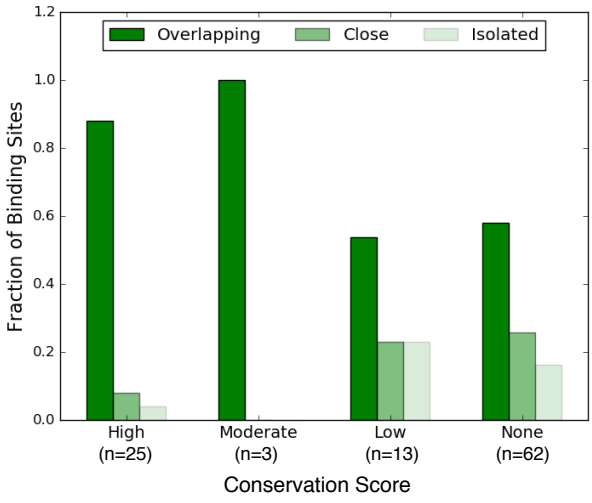
A)


Fraction of Binding Sites
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
High (n=25)
Moderate (n=3)
Low (n=13)
None (n=62)
Conservation Score

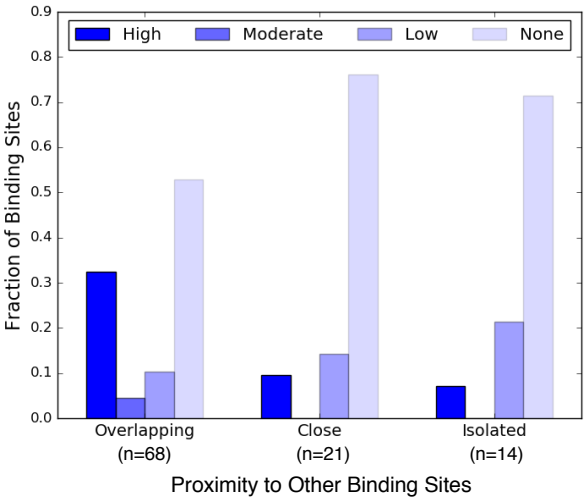
B)


Fraction of Binding Sites
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Overlapping (n=68)
Close (n=21)
Isolated (n=14)
Proximity to Other Binding Sites

C)


Overlapping
Close
Isolated
Fraction of Binding Sites
0.0
0.2
0.4
0.6
0.8
1.0
1.2
High (n=25)
Moderate (n=3)
Low (n=13)
None (n=62)
Conservation Score

D)


High
Moderate
Low
None
Fraction of Binding Sites
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Overlapping (n=68)
Close (n=21)
Isolated (n=14)
Proximity to Other Binding Sites

E)

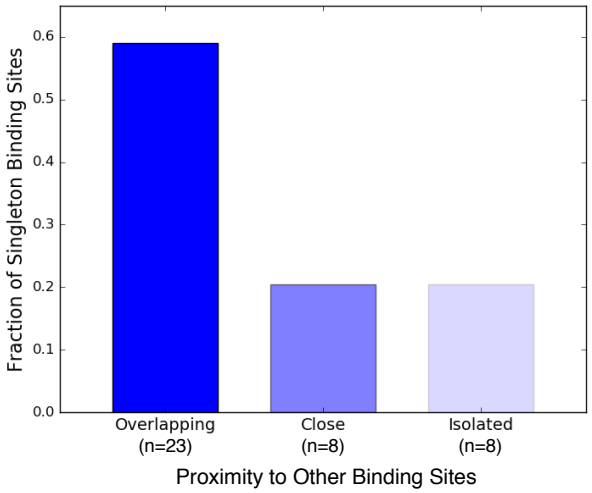
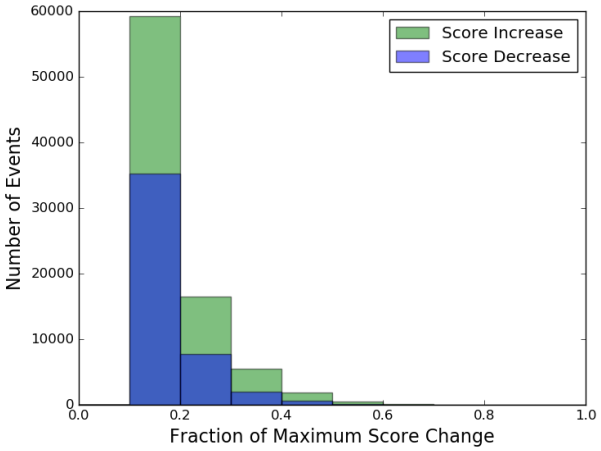
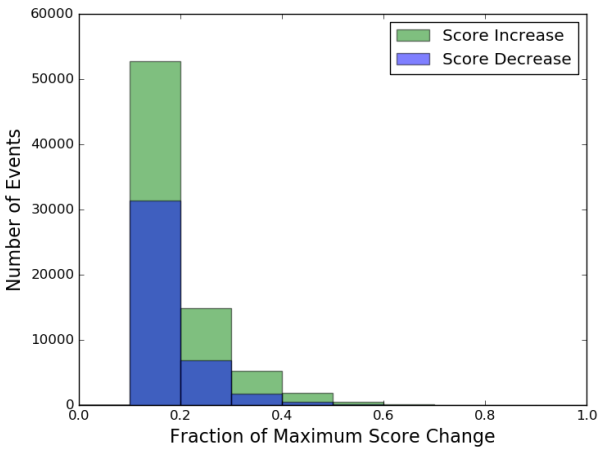

Fraction of Singleton Binding Sites
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Overlapping (n=23)
Close (n=8)
Isolated (n=8)
Proximity to Other Binding Sites

Figure 3.2. Binding site conservation and proximity in the *eve* stripe 2 enhancer across 24 *montium* species.
A) The distribution of conservation scores for all predicted binding sites. Each group of orthologous binding sites was assigned a simple conservation score based on the fraction of *montium* species it was observed in: high ≥ 0.8; 0.5 ≤ moderate < 0.8; 0.2 ≤ low < 0.5; and none < 0.2. B) Each group of orthologous binding sites was also classified based on its proximity to sites for different factors, using the nomenclature from [21]. Overlapping binding sites share at least one base pair; close binding sites are within ten base pairs, but do not overlap; and isolated binding sites are more than 10 base pairs apart. C) Binding site proximity as a function of conservation score. D) Conservation score as a function of binding site proximity. E) Binding site proximity for all singleton sites, which by definition are present in only one *montium* species.

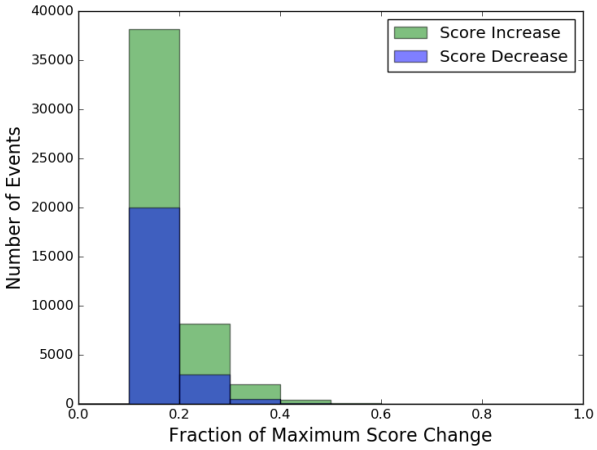
A) Bicoid



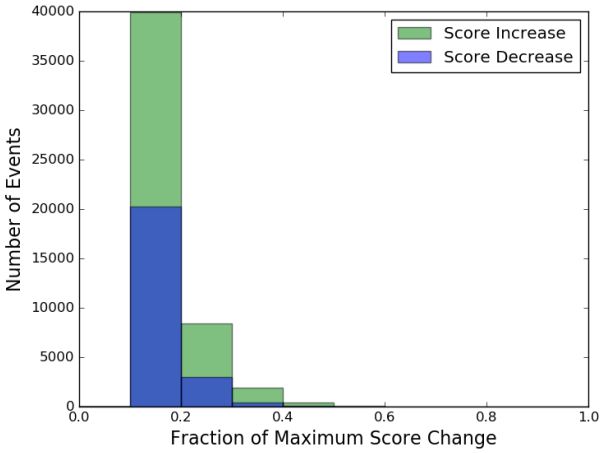
B) Bicoid Shuffled



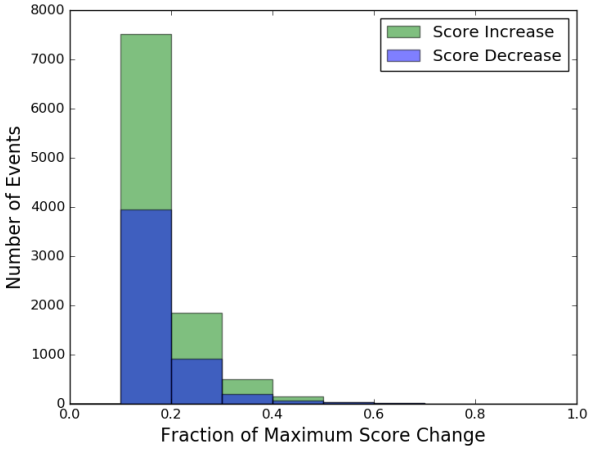
C) Krüppel



D) Krüppel Shuffled



E) Zelda



F) Zelda Shuffled

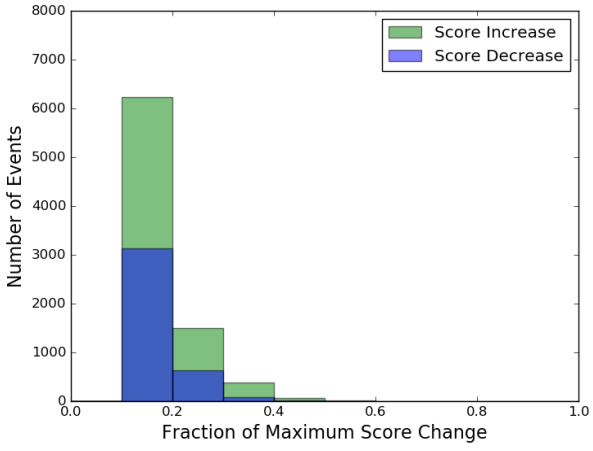
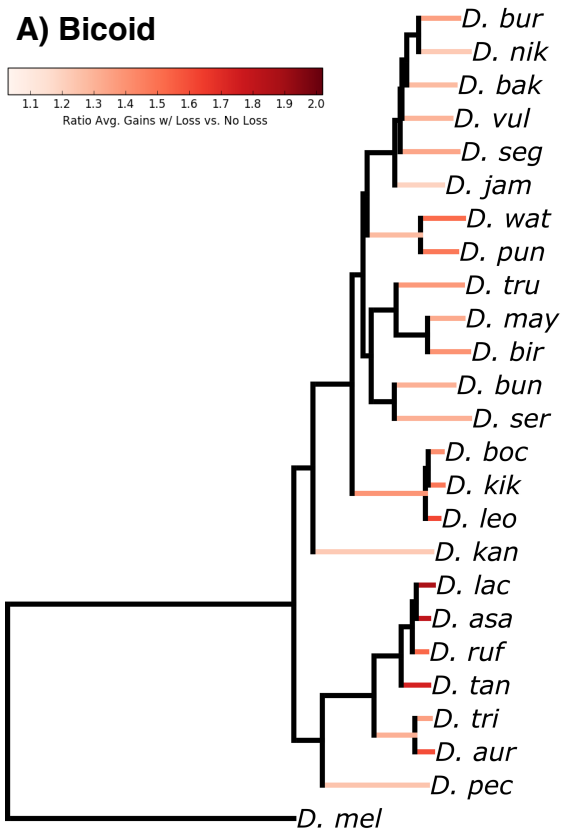
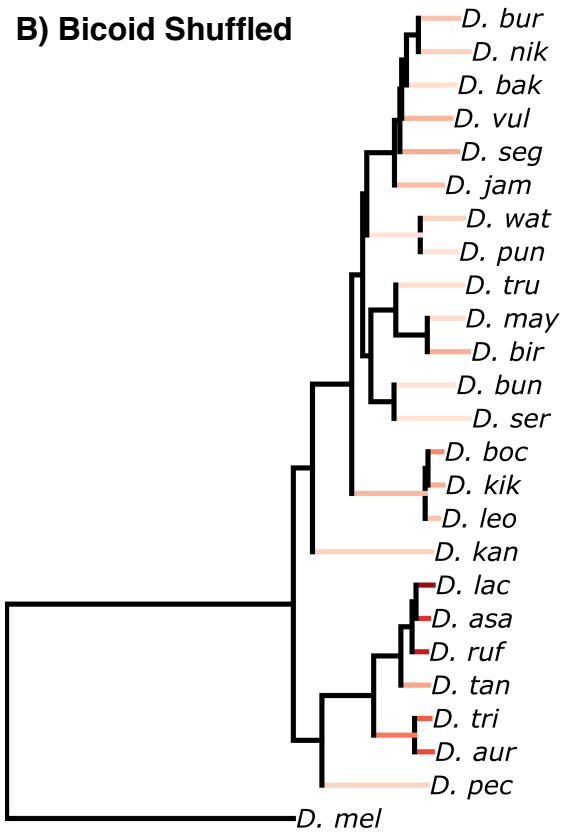
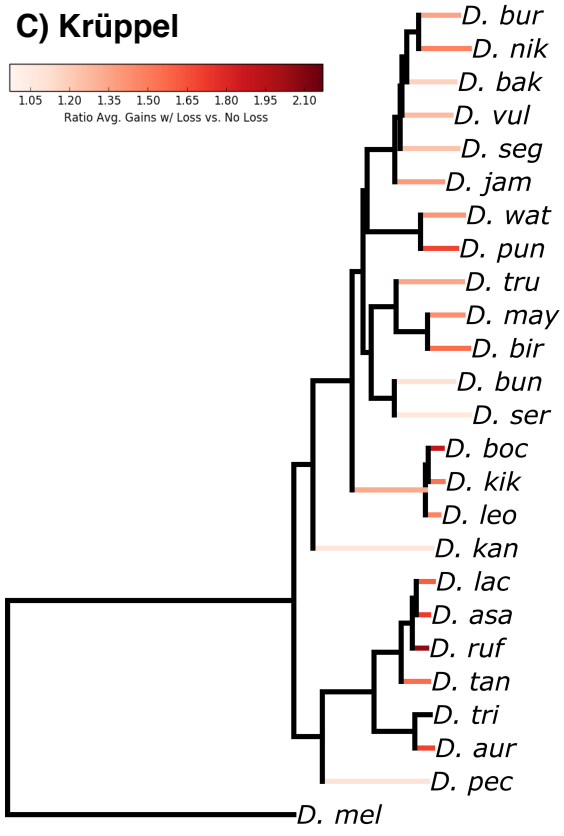
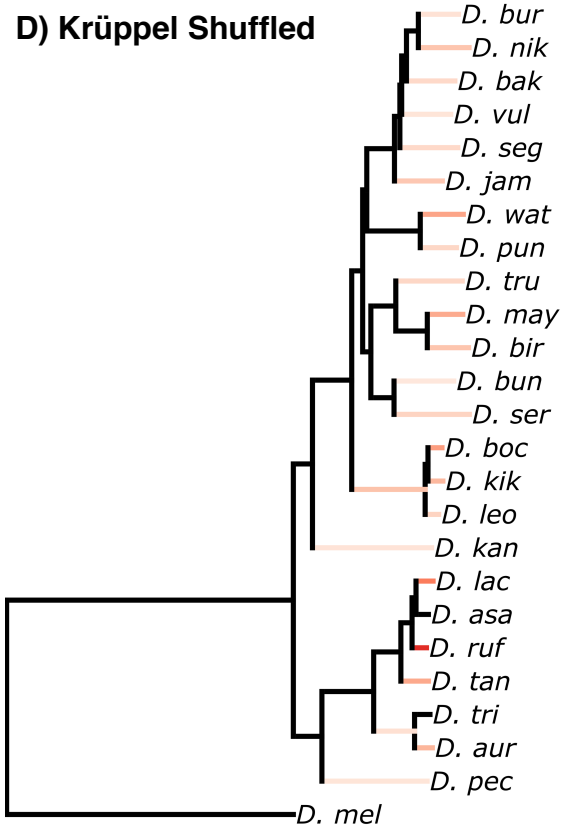


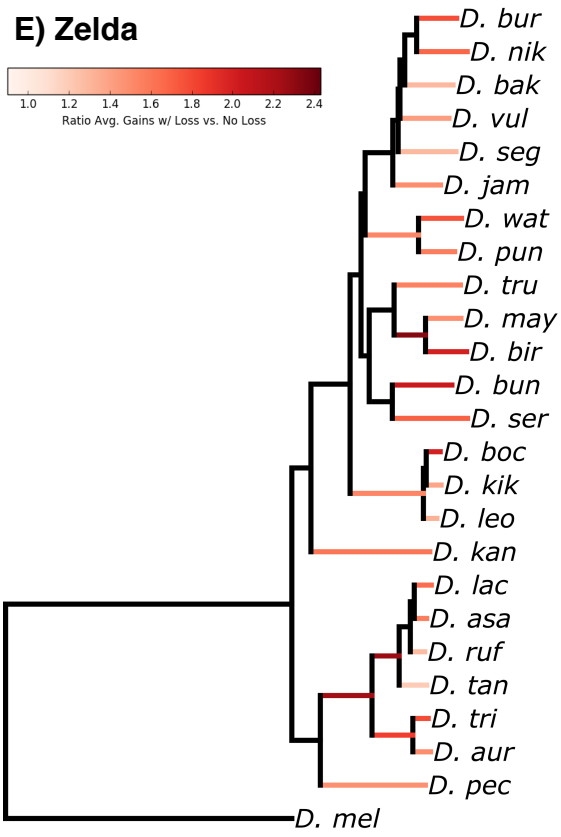
Figure 3.3. Distribution of binding score changes across all branches of the *montium* subgroup tree for Bicoid, Krüppel, and Zelda.
For each factor and bound region, binding sites were predicted in each species using PATSER [92], with a score cutoff of zero. Predicted binding sites were then mapped onto multiple sequence alignments (MSAs) generated by MAFFT [82-84], and clustered into groups of orthologous sites. For each group of sites, I reconstructed ancestral binding site scores at each internal node of the tree using maximum likelihood and the Brownian motion (BM) model of evolution, as implemented in the function anc.ML from the package phytools [95]. Changes along each branch of the tree were calculated by taking the score difference between parent - child nodes. Since binding sites were predicted using a cutoff of zero, the magnitude of the maximum score change is equal to the maximum binding score. This is 7.584 for Bicoid, 10.622 for Krüppel, and 10.515 for Zelda. In the figure, score changes are shown as fractions of the maximum score change, with a cutoff of 0.1. Changes were aggregated across all branches of the tree. The analysis was conducted using the original matrices, as well as matrices with shuffled columns. For comparison, note that there is a significant difference in the size of bound regions between factors. For Bicoid and Krüppel, the original bound regions in *D. melanogaster* were 1 kb, and between 300 - 400 bp for Zelda. A) Bicoid: 83,668 gain events, and 45,696 loss events. B) Bicoid Shuffled: 75,206 gain events, and 40,544 loss events. C) Krüppel: 48,823 gain events, and 23,618 loss events. D) Krüppel Shuffled: 50,765 gain events, and 23,787 loss events. E) Zelda: 10,072 gain events, and 5,183 loss events. F) Zelda Shuffled: 8,212 gain events, and 3,879 loss events.


A) Bicoid
1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
Ratio Avg. Gains w/ Loss vs. No Loss
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel


B) Bicoid Shuffled
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel


C) Krüppel
1.05 1.20 1.35 1.50 1.65 1.80 1.95 2.10
Ratio Avg. Gains w/ Loss vs. No Loss
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel


D) Krüppel Shuffled
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel


E) Zelda
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Ratio Avg. Gains w/ Loss vs. No Loss
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel

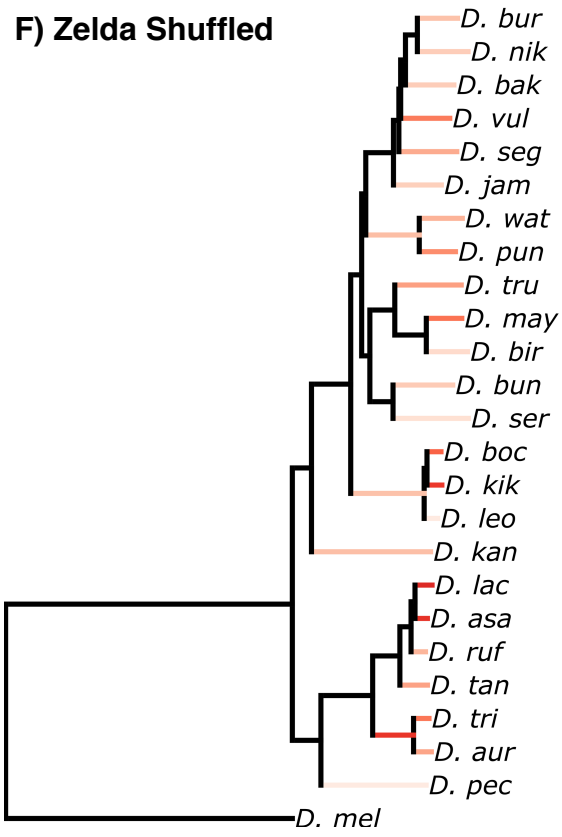
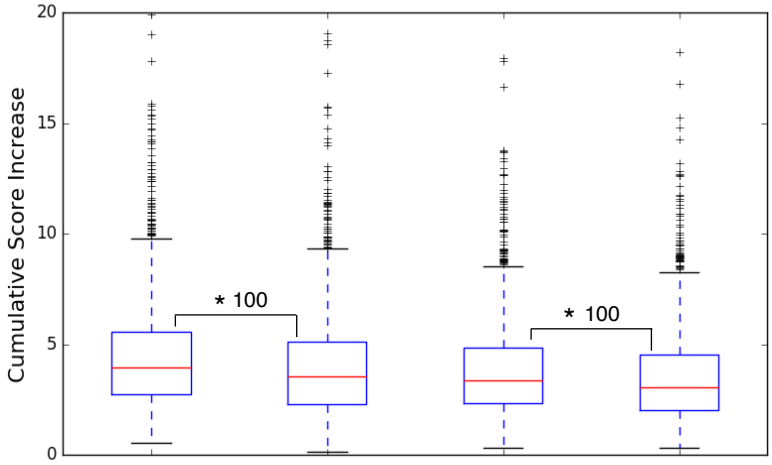
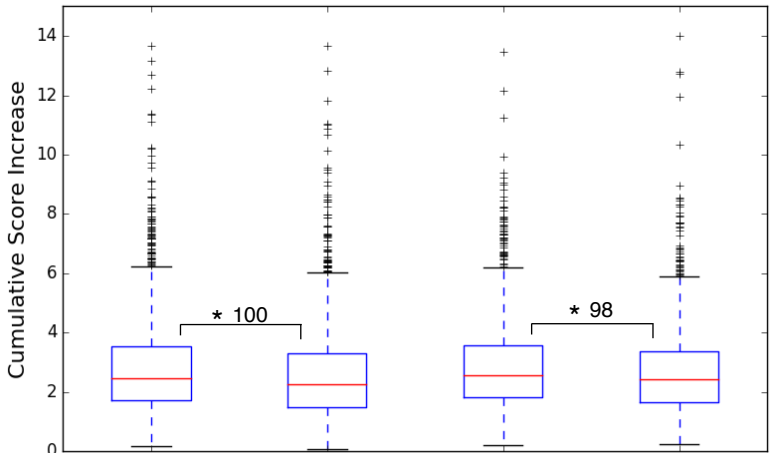

F) Zelda Shuffled
D. bur
D. nik
D. bak
D. vul
D. seg
D. jam
D. wat
D. pun
D. tru
D. may
D. bir
D. bun
D. ser
D. boc
D. kik
D. leo
D. kan
D. lac
D. asa
D. ruf
D. tan
D. tri
D. aur
D. pec
D. mel

Figure 3.4. Correlated changes in binding site scores across branches of the *montium* subgroup species tree.
For each factor and bound region, binding sites were predicted in each species using PATSER [92], with a score cutoff of zero. Predicted binding sites were then mapped onto multiple sequence alignments (MSAs) generated by MAFFT [82-84], and clustered into groups of orthologous sites. For each group of binding site scores, ancestral scores were reconstructed at each internal node of the tree using maximum likelihood and the Brownian motion (BM) model of evolution, as implemented in the function anc.ML from the package phytools [95]. Changes along each branch of the tree were calculated by taking the score difference between parent - child nodes. Bound regions were divided into two groups: regions where the score of one or more binding sites decreased by at least a factor-specific threshold, and regions with no such decrease. The cutoffs were -1.896 for Bicoid, -2 for Krüppel, and -1 for Zelda. For each bound region, I then summed the total score increase along the original branch and its parent branch. Finally, I calculated the average score increase under both conditions (loss / no loss). Each branch on the tree is color-coded based on the ratio of the average score increase under the conditions loss / no loss. Darker reds indicate bigger differences. If a branch was too short for a meaningful comparison, it was colored black. The *D. melanogaster* branch, and the branch leading to the *montium* clade, were also excluded from the analysis and colored black. The analysis was repeated using shuffled matrices for each factor.

A) Bicoid



B) Krüppel



C) Zelda

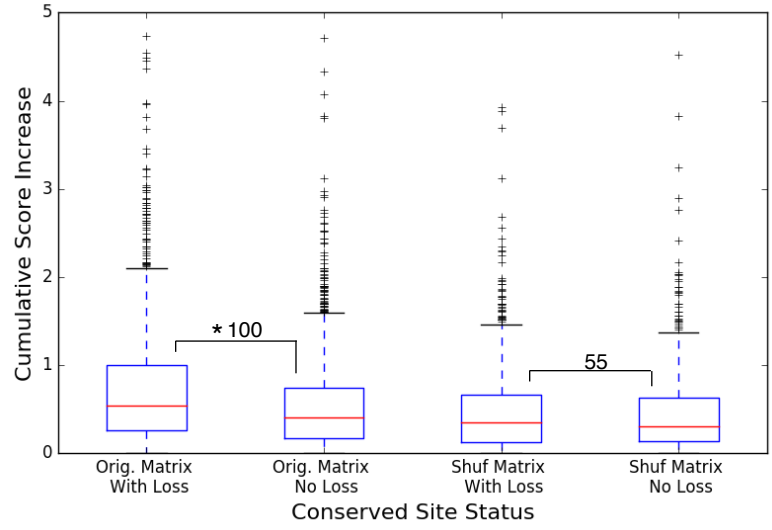


Figure 3.5. Correlated changes in binding scores for conserved sites across branches of the *montium* subgroup species tree.
For each factor (Bicoid, Krüppel, and Zelda) and condition (original or shuffled matrix), box plots compare the cumulative score increase with and without the loss of a conserved site. Cumulative score increases are per bound region, and adjusted for branch length differences on the tree. Results are shown for a single control set. The bracket linking adjacent box plots indicates whether or not the difference is statistically significant using Welch's t-test (*), and also shows the number of control sets that produce a statistically significant difference. A) Bicoid matrix: 2,028 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 4.50, and 4.04 without a loss. This difference is statistically significant ($p < 1.50e-8$). All control set comparisons are also statistically significant ($p < 0.05$). Shuffled Bicoid matrix: 1,736 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 3.89, and 3.59 without a loss. This difference is statistically significant ($p < 0.0002$). All control set comparisons are also statistically significant ($p < 0.05$). B) Krüppel matrix: 1,685 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 2.83, and 2.63 without a loss. This difference is statistically significant ($p < 0.0005$). All control set comparisons are also statistically significant ($p < 0.05$). Shuffled Krüppel matrix: 1,507 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 2.87, and 2.69 without a loss. This difference is statistically significant ($p < 0.002$). For the remaining control sets, 98/100 comparisons are also statistically significant ($p < 0.05$). C) Zelda matrix: 1,585 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 0.73, and 0.55 without a loss. This difference is statistically significant ($p < 2.96e-14$). All control set comparisons are also statistically significant ($p < 0.05$). Shuffled Zelda matrix: 832 branches / bound regions lost a conserved site. The average cumulative score increase with a loss is 0.49, and 0.45 without a loss. This difference is not statistically significant ($p = 0.11$). For the remaining control sets, 55/100 comparisons are statistically significant ($p < 0.05$).

Discussion

I started by looking at the extensively studied *eve* stripe 2 enhancer across 24 *montium* species and *D. melanogaster*. As expected, given the density of species and the recent divergence times, I observed one or two changes in transcription factor binding sites between closely related species. This level of resolution ensures that when key mutational events do occur, I can observe potential compensatory changes before they are obscured by additional mutations. I then showed that I could observe previously described patterns of binding site conservation and proximity for the known regulators of the *eve* stripe 2 enhancer [21]. Within the *montium* subgroup, a core group of sites appeared to be conserved across the radiation. Many new sites also appeared in subclades or individual species. The vast majority of binding sites were overlapping - whether they appeared to be highly conserved, or arose in a single species.

Next, I showed how patterns of (apparent) conservation and variation within the *montium* subgroup could be used to direct targeted mutagenesis experiments, and test models of enhancer grammar. Overlapping pairs of activators and repressors are thought to play an important functional role in the enhancer [9, 12-14]. However, simulations have shown that well-described enhancer features - such as overlapping and clustered sites - can also arise when selection acts on the overall number of sites, not their specific arrangement [22]. I highlighted two cases in which a seemingly highly conserved and overlapping binding site was lost in a single *montium* species. In the first case, a new, non-overlapping site arose elsewhere in the enhancer - suggesting that close proximity to other binding sites was not strictly necessary. Mutagenesis of the new site could help to answer this question. In the second case, no new binding site was visible in the minimal enhancer. This again challenges the potential importance of overlapping and conserved sites. It also shows that should compensatory changes exist, they don't necessarily have to involve the acquisition of sites for the same factor. Again, mutagenesis experiments could help to identify the compensatory changes in such cases.

After an initial exploration of the *eve* stripe 2 enhancer, I moved on to investigating hundreds of regions bound by the transcription factors Bicoid, Krüppel, and Zelda across 25 *Drosophila* species. I modeled groups of orthologous binding site scores as continuous traits, reconstructed ancestral binding scores at each node of the tree, and calculated score changes along each branch of tree. I think this was a more realistic methodology compared to all-or-nothing approaches to binding site dynamics. It allowed me to observe both true gain and loss events, as well smaller increases or decreases in binding strength.

Bound regions that lost a binding site (or saw a large decrease in binding score) along a given branch of the tree were more likely to gain sites along the same branch (or parent branch), compared to regions that did not lose a site. The average increase was about 40 % for Bicoid and Krüppel-bound regions, and 65 % for Zelda-bound regions. Most of these differences were statistically significant. Bound regions that lost a conserved site

also showed similar results. Taken in isolation, these results indicate that the loss of a binding site is often (but certainly not always) compensated by the gain of a new site for the same factor. However, shuffled matrices for Bicoid, Krüppel, and Zelda show similar patterns, making it difficult to conclude these are meaningful changes in transcription factor binding. That being said, the fact that the original matrices produced larger differences suggests that there may be real (but subtle) signals that are being obscured by a variety of confounding factors.

One likely possibility is the presence of large numbers of non-functional *montium* sequences. Multiple studies have shown that transcription factor binding can diverge between closely related species [27, 88-91]. Enhancers themselves can also turnover [27]. In this chapter, I remapped known ChIP peaks from *D. melanogaster* onto *montium* assemblies. It's possible that a significant fraction of these regions are unbound in one or all *montium* species. Problems related to remapping coordinates between species using liftOver [51] are also possible. If an orthologous enhancer / bound region was no longer under selection, then it would essentially drift, accumulating both gain and loss events at an increased rate. If such regions were mixed with a larger pool of constrained sequences, they could drive the correlated gain / loss signal seen with both the original and shuffled matrices - for reasons that have nothing to do with compensatory changes.

Accordingly, future efforts should focus on manually curating bound regions to identify individual sequences (or even entire regions) that no longer appear to be constrained, or are otherwise problematic. I attempted to do this with several algorithmic filtering steps, but manual inspection of each locus is likely required. After creating a set of high-quality regions, the entire analysis should be repeated. If non-functional sequences were obscuring a real signal, then the difference between the original and shuffled matrices should increase. In the future, it might also be advantageous to generate ChIP-seq and STARR-seq data within the *montium* subgroup itself. Data from only a handful of species would suffice to cover all major subclades.

Another confounding factor is that the suite of potential compensatory mutations might be large. In this chapter, I focused on gain - loss events for the same transcription factor. Many other changes are possible though. For example, the loss of a binding site for one factor might be compensated by the gain of a site for a different factor. Or instead of gaining a new site, the spacing between existing sites - for either the same factor, or different factors - could change so as to mitigate the effects of a loss. Evidence from the *sparkling* enhancer indicates that compensatory changes can involve different factors [24]. In my own analysis of the *eve* stripe 2 enhancer, I observed that *D. pectinifera* lost a strong Hunchback site that appeared to be conserved across the *montium* subgroup and *D. melanogaster*, suggesting functional importance. (Conservation can sometimes be illusionary though [22].) *D. pectinifera* did not acquire any new Hunchback sites. As the diversity of compensatory changes increase, it becomes harder to see a signal for any one type of change. But such processes could account for the flexibility seen in the *eve* stripe 2 and *sparkling* enhancers. This would also align with flexible models of enhancer evolution, such as the "billboard" [14, 30, 31] and "TF collective" [32] models. Understanding the full suite of compensatory changes,

their relative importance, and how / when they operate, will be a major challenge going forward.

In addition, this analysis posed a number of methodological challenges. Another limitation is using the Brownian motion (BM) model of evolution to estimate ancestral binding site scores. While this model is often used for estimating ancestral states of continuous characters [e.g., 85, 95], forcing changes to occur gradually can be misleading. For example, consider a single mutational event that causes a large and discontinuous increase in binding site score. This change should generally occur along a single branch of the tree. However, the forced gradualism of the BM model spreads this change out across multiple nodes, leading to the inference of multiple gain events across the tree. It can also create the erroneous appearance of loss events in closely related species. (This can also confound any conclusions about the relative number of gain / loss events for binding site scores.) Recently, models have been developed that allow for leaps / jumps in the evolution of continuous characters [e.g., 96, 97, 98]. However, current implementations of these models for ancestral character estimation are prohibitively time-consuming when tens of thousands of characters must be reconstructed for a single transcription factor / set of bound regions.

Finally, another potential problem is ancestral character estimation using a single species tree in the presence of widespread gene tree / species tree incongruence. Previous phylogenetic reconstructions of the *montium* subgroup - typically based on a small number of genes - have produced incongruent trees [34, 37, 53-56]. I also observed incongruence between individual gene trees when reconstructing a species tree based on 20 Bicoid-dependent enhancers (data not shown). Given the large number of closely related species, incomplete lineage sorting (ILS) and introgressive hybridization are likely major drivers of discordance. When characters are mapped onto a single species tree, incongruence creates the appearance of homoplasy, and leads to an overestimate of the number of gain / loss events on the tree. This phenomenon is known as hemiplasy [99-101]. This could confound my efforts to observe correlated gain / loss events along branches of the tree. Future efforts should focus on developing ancestral character estimation methods that can account for ILS. Short of that, the analysis could be repeated by mapping characters onto each gene tree. To aggregate the data, this would require combining results across all branches for a single bound region. So with the exception of the leaves of the tree, I would lose the ability to observe branch-level differences as in Figure 3.4.

Despite these limitations and challenges, the *montium* subgroup has proven itself to be a tractable system with which to study enhancer evolution.

Materials and Methods

montium eve stripe 2 enhancer analysis

Coordinates for the minimal *eve* stripe 2 enhancer in *D. melanogaster* [13] (RedFly ID: RFRC:0000000276.003) were remapped onto each *montium* assembly using liftOver [51], with the options -minMatch=0.1 and -multiple.

Binding sites for the factors Bicoid, Giant, Hunchback, Krüppel, Sloppy Paired 1, and Zelda were predicted in each species using PATSER (v. 3e) [92], with the options -c, -lp, -s, and a background file based on the average nucleotide frequencies across the top Bicoid-bound regions. Frequency matrices for Bicoid, Hunchback, and Krüppel were from [102]; Giant from [103, 104]; Sloppy Paired 1 from [105]; and Zelda from [104]. For all factors except Zelda, I used the ln(*p*-value) cutoffs from [21], which were chosen so that predicted binding sites coincided with experimentally verified sites (DNase I footprinting) in *D. melanogaster*. The cutoffs were as follows: Bicoid: -6, Giant: -5.5, Hunchback: -6, Krüppel: -6, and Sloppy Paired 1: -6. Separately, the ln(*p*-value) cutoff for Zelda was set to -6.

Orthologous sequences for each enhancer were aligned using MAFFT (v. 7.407) [82-84] with the linsi preset, along with the options --anysymbol, --ep 0.123, and --nuc. Predicted binding sites were then mapped onto the multiple sequence alignment (MSA). See Figure 3.1 for additional details about the binding site plot.

ChIP data

For Bicoid and Krüppel, I started with coordinates for the top ChIP-chip peaks (1 % FDR, symmetric-null test) from stage 5 embryos in *D. melanogaster* [93]. This included 619 Bicoid-bound regions, and 1,000 Krüppel-bound regions. For each bound region, I extracted coordinates for a 1 kb region centered on the peak. For Zelda, I started with coordinates for the top 1,000 ChIP-seq peaks from cycle 14 embryos in *D. melanogaster* [94]. These peaks were generally far more resolved than the broader Bicoid and Krüppel peaks. Bound regions shorter than 300 bp, or larger than 400 bp, were removed, leaving 796 Zelda-bound regions.

The overlap between the above ChIP peaks, and the *D. melanogaster* enhancers described in [52], were determined using BEDTools (v. 2.17.0) [106] intersect.

Coordinates for *D. melanogaster* ChIP peaks were remapped onto each *montium* assembly using liftOver [51], with the options -minMatch=0.1 and -multiple. For Bicoid and Krüppel-bound regions, individual *montium* sequences shorter than 750 bp, or longer than 1,300 bp, were removed. For Zelda-bound regions, *montium* sequences shorter than 250 bp, or longer than 400 bp, were removed. Finally, entire bound regions were removed if they contained less than 21 *montium* species.

Predicting binding sites and mapping them onto MSAs of ChIP regions

Binding sites for the factors Bicoid, Krüppel, and Zelda were predicted in each species using PATSER (v. 3e) [92], with the options -c, -ls 0, -s, and a background file based on

the average nucleotide frequencies across the top Bicoid-bound regions. This predicted sites on both strands with scores greater than or equal to zero. For a given species and factor, binding sites on the same strand were collapsed into a single site if they overlapped by four or more bases.

Orthologous sequences for each bound region were aligned using MAFFT (v. 7.407) [82-84] with the linsi preset, along with the options --anysymbol, --ep 0.123, and --nuc. Predicted binding sites were then mapped onto MSAs. Because of insertion events in other species, coordinates for mapped binding sites might be divided across one or more gaps. In such cases, the site was anchored onto the MSA at the position with the largest number of continuous bases.

Binding sites in different species that were on the same strand and position in the MSA were grouped into arrays of orthologous sites. Because of alignment error, orthologous binding sites might not always map to the same position in the MSA. This is a significant problem, since it can erroneously create the appearance of gain / loss events. I wrote an algorithm that merged nearby binding sites into arrays of putatively orthologous sites. First, all overlapping arrays on the same strand were grouped together. The algorithm then iteratively attempted to merge all arrays within the same group, starting with the closest and highest scoring arrays. If a binding site for at least one species was present in both arrays, and the scores were greater than or equal to one-tenth the maximum score, the merger was rejected. This equated to 0.75 for Bicoid, and 1.0 for Krüppel and Zelda. When a merger was accepted, the alignment position of the higher-scoring array was retained. If the alignment error was large enough, orthologous binding sites might not even overlap in the MSA. So after merging overlapping arrays, the algorithm went back and attempted to merge all adjacent arrays within six base pairs, using the method described above.

Reconstructing ancestral binding site scores

Ancestral binding site scores were estimated by the function anc.ML from the package phytools (v. 0.6.60) [95], using maximum likelihood, Brownian motion (model="BM"), and the species tree reported in Chapter 2. Score changes along each branch of the tree were calculated by taking the difference between parent - child nodes.

For each branch, bound regions were divided into two groups (loss / no loss) based on whether or not the score of at least one binding site decreased by a factor-specific threshold. The cutoffs were -1.896 for Bicoid, -2 for Krüppel, and -1 for Zelda. Cumulative score increases were calculated along each branch and parent branch by summing all score increases above a factor-specific threshold (one-tenth the maximum score increase). This equated to 0.7584 for Bicoid, 1.0622 for Krüppel, and 1.0515 for Zelda.

Chapter 4: Sequencing the genome of *E. muscae* ‘Berkeley’, a parasite that manipulates *D. melanogaster* behavior

The work detailed in this chapter was a joint effort between Michael Bronski, Carolyn Elya, and Michael Eisen. The chapter itself was written by Michael Bronski.

Abstract

Elya et al. [107] recently discovered a strain of *Entomophthora muscae* infecting wild *Drosophila*, and developed methods to maintain infected *D. melanogaster* in the laboratory. This is an exceptional system with which to study the molecular basis of parasite-induced behavioral manipulation. *E. muscae* ‘Berkeley’ produces a suite of behavioral changes, including summit disease, proboscis extension / attachment, and raised / spread wings. In this chapter, we describe the sequencing and assembly of the *E. muscae* ‘Berkeley’ genome. Based on a *k*-mer frequency spectrum, the estimated genome size is 1.3 Gb. We assembled the genome from a 10X Chromium library using the Supernova assembler. This yielded a 1.24 Gb assembly. Given that large genome sizes and polyploidy appear to be common among related entomopathogenic fungi, estimating the haploid genome size in the absence of additional experimental data is challenging - but it might be around 650 Mb. The *E. muscae* ‘Berkeley’ genome is highly repetitive, with a repeat content of roughly 85 %. The BUSCO assessment of the 10X assembly showed that 40 % of known single-copy fungal genes are missing or fragmented. The significance of this is unclear. In a separate analysis using single-isoform BUSCOs (SIBs), we showed that the average coverage of single-copy SIBs is twice that of duplicated SIBs; and that single-copy SIBs contain large numbers of unfiltered biallelic SNPs. These data indicate that single-copy SIBs represent genomic regions where only one of two closely related haplotypes was assembled. In a testament to the high repeat content, the alignment of scaffolds containing the same duplicated SIB produced characteristic X-alignments, where the forward strand of one scaffold also aligned to the reverse complement of the second scaffold. Finally, to look for mis-assemblies, we aligned PacBio long-reads to the 10X assembly. Nearly one quarter of the alignments include flaps of at least 4 kb, nearly half the average read length. The majority of these flaps cannot be explained by intersections with scaffold ends or large gaps, suggesting numerous-misassemblies. Future efforts will focus on improving and annotating the genome. Going forward, the *E. muscae* ‘Berkeley’ genome will support our efforts to understand the mechanistic basis of fungal-induced behavioral manipulations in *D. melanogaster*.

Introduction

Parasitic manipulation of host behavior is ubiquitous

Many parasites have evolved the ability to manipulate the behavior of their host, often in dramatic fashion. Notable examples include *Ophiocordyceps*-infected “zombie ants” that bite down on north-facing leaves with precise temperature and humidity conditions [e.g., 108], and rats that lose their natural aversion to cat urine when infected by *Toxoplasma gondii* [e.g., 109, 110]. The sheer number of known examples indicates that such ecological interactions are ubiquitous. Despite widespread interest in extended phenotypes [111], we know little about the molecular basis of these manipulations. This is driven in large part by the dearth of resources available in most host-parasite systems. What’s needed is a resource-rich and tractable host-parasite system.

The insect destroyer (and manipulator) *Entomophthora muscae*

The entomopathogenic fungus *Entomophthora muscae*, first described by Cohn [112], infects flies from several Dipteran families, including Muscidae and Drosophilidae [e.g., 113]. In the late afternoon and early evening, critically ill flies climb to elevated positions to die, a phenomenon known as summit disease. Once in position, the fly extends its proboscis to the substrate, where it is anchored in place by rhizoids and fungal-produced secretions [114]. The fly then spreads it wings, raises them up and off the abdomen, and finally dies in place [115]. Throughout the night and early morning, the sporulating cadaver showers the surrounding area (and any exposed flies) with infectious conidia [116, 117]. What’s more, healthy male house flies appear to be attracted to mycosed female cadavers, and become infected when they attempt to mount the cadavers [e.g., 118]. The molecular basis of these manipulations is entirely unknown. An extensive literature describes *E. muscae* infection in the house fly, *Musca domestica*. This stems in large part from the desire to use *E. muscae* to control filth fly populations on farms. Elya et al. [107] recently discovered an *E. muscae* strain infecting wild *Drosophila* in Berkeley, CA, and developed methods to maintain infected *D. melanogaster* in the laboratory. We start by reviewing the *M. domestica* literature, and then move on to recent advances in *D. melanogaster*.

Typical *E. muscae* symptoms in infected house flies

Today, *Entomophthora muscae* (Cohn) Fresenius [112] is known to be a species-complex, its members collectively referred to as *Entomophthora muscae sensu lato*. The complex is currently divided into six species based on the number of nuclei per conidium, the nuclear diameter, the dimensions of the primary conidia, and the host species [119-122].

Infection begins when a conidium (an infectious spore) penetrates the host cuticle, almost always on the abdomen. Protoplasts first invade the hemocoel, and then colonize the fat-rich abdomen. Fungal growth is concentrated in the abdomen for the

majority of the four to five-day incubation period. However, after consuming the fat body, the pathogen begins to invade tissues and organs in the abdomen, thorax, and head (including the brain). At this point, the fly is near death [123].

Throughout the course of the incubation period, infected house flies show varying responses to temperature. On days two and three of a five-day incubation period, flies actively seek out warm temperatures, a phenomenon known as behavioral fever. When placed in a temperature gradient ranging from 26^{o}C to 42^{o}C, most infected flies quickly move to 42^{o}C, where they remain for the next several hours. At this point in the infection, incubation at 35^{o}C - 42^{o}C for 4 - 6 hours is sufficient to cure most flies of the pathogen. As the infection progresses though, behavioral fever is replaced by heat avoidance, a behavioral manipulation thought to benefit the sporulation / germination of the fungus. When placed in the same temperature gradient on day five, most infected flies quickly move to 26^{o}C, and remain there (or at 28^{o}C) until death [124, 125]. As death approaches, additional behavioral changes appear.

In the late afternoon and early evening, critically ill flies crawl to elevated positions to die, a phenomenon known as summit disease. Flies usually come to rest on vertical surfaces, but in those rare instances when they die on horizontal surfaces, they are always found on the underside of the object [117]. After coming to rest in an elevated location, the fly extends its proboscis to the substrate, where it is anchored in place by rhizoids and a sticky secretion [114]. The fly then spreads its wings latero-dorsally, raises them up and off the abdomen, and then finally dies in place. This sequence – from the last locomotory movement to the complete lifting of the wings – usually takes 75 minutes [115, 122].

Several hours after host death, conidiophores erupt from intersegmental membranes along the abdomen, forming prominent white bands. Throughout the night and early morning, the cadaver releases showers of forcibly ejected primary conidia, which in turn give rise to secondary conidia. Peak primary discharge occurs 10 - 12 hours post mortem, while the lagged secondary discharge peaks 16 -18 hours post mortem. From its elevated location, the cadaver is well positioned to shower the surrounding area (and any exposed flies) with infectious conidia. Interestingly, the discharge of highly infectious secondary conidia peaks between the hours of 7 and 10 a.m., a time when flies are active and aggregating [116, 117].

The observation that infected flies always die in the late afternoon / early evening prompted detailed ethological experiments in the laboratory. By altering light-dark cycles in the laboratory, Krasnoff et al. [115] showed that infected house flies always die 0 - 5 hours before the onset of darkness. This pattern is consistent with a gated phenomenon controlled by a biological clock. Killing flies in the late afternoon / early evening likely ensures a cool and humid environment for the sporulation of the cadaver and the germination of conidia. But, as Mullens and co-authors speculated, host death may be timed so that highly infectious secondary conidia are released at a time when they can infect large numbers of exposed flies.

In addition to conidia showers, the pathogen appears to spread through house fly populations by manipulating the attractiveness of mycosed females. Reports from both the field and the laboratory indicate that healthy male house flies frequently attempt to copulate with mycosed female cadavers [117, 118, 126, 127]. What's more, when presented with a choice between a mycosed female and an uninfected female cadaver, males almost always choose the mycosed female [118]. The visual and / or chemical cues that mediate this attraction are unknown, but it may account for the higher incidence of infection seen in males [117, 126]. As they attempt to mount the cadaver, males come into direct contact with the sporulating abdomen, all but ensuring infection and death. Before succumbing to the mycosis though, they continue to transfer spores to numerous uninfected females via mechanical transmission [128].

E. muscae infections in *Drosophila*

Historically, three members of the *E. muscae s. l.* species complex were reported to infect *Drosophila*: *E. ferdinandii*, *E. muscae sensu stricto*, and *E. schizophorae* [113, 122, 129, 130]. Furthermore, two of these species were known to infect *D. melanogaster* – both in the field and in the laboratory – and produce typical infections (e.g., summit disease, proboscis extension / attachment, and white bands of conidiophores). Goldstein [129] described naturally occurring epizootics in New York that produced typical *E. muscae s. l.* symptoms in *D. melanogaster* and *D. repleta*. Based on sketches of primary conidia, these outbreaks appear to have been caused by *E. ferdinandii*. In the laboratory, Steinkraus and Kramer [113] exposed flies from eight Dipteran families – including *D. melanogaster* – to conidia showers from an isolate of *E. schizophorae*. This isolate infected 11 % of exposed *D. melanogaster*, and produced a typical mycosis. This study also highlights the impressive host range of some isolates.

Elya et al. [107] recently discovered a strain of *E. muscae s. l.* infecting wild *Drosophila* in Berkeley, CA, and developed methods to maintain infected *D. melanogaster* in the laboratory. Under optimized laboratory conditions, *E. muscae* 'Berkeley' kills approximately 80 % of CantonS flies after a 4 - 5 day incubation period. Based on imaging of primary conidia, this isolate appears most closely related to *E. muscae sensu stricto*. *E. muscae* 'Berkeley' infection produces characteristic symptoms, including summit disease, proboscis extension / attachment, raised / spread wings, and white bands of conidiophores around the abdomen. Elya et al. [107] also assembled the *E. muscae* 'Berkeley' transcriptome, and profiled host and parasite gene expression at key time points throughout the infection. Intriguingly, histological experiments showed that the parasite invades the nervous system early, and is visible in the brain within 48 hours of infection. However, the significance of this finding as it relates to behavioral manipulation is still unclear.

Little is known about the *E. muscae* genome. The closest reference genome is *Conidiobolus coronatus*, which at 40 Mb is typical for most fungi [131]. However, it is a distantly related member of the phylum *Entomophthoramycota* [132, 133]. In contrast, experimental evidence from the far more closely related fungus *Entomophaga aulicae* indicates its genome might be as large as 8 Gb [134]. Evidence also suggests that *E.*

muscae could be polyploid. Ultrastructural studies of mitosis in species from the genera *Erynia* and *Strongwellsea* reported chromosome counts of 8, 12+, 16, and 32 [reviewed in 135]. These genera belong to the subfamily *Erynioideae*, one of two subfamilies in the larger *Entomophthoraceae*. The other subfamily, the *Entomophthoroideae*, includes *E. muscae* [132, 133]. Based on this limited survey, polyploidy might be common in the *Entomophthoraceae*, with a basal chromosome number of 8 [135, Richard Humber, personal communication].

In this chapter, we describe our efforts to assemble the *E. muscae* 'Berkeley' genome. A draft genome will aid our current and future efforts to understand the mechanistic basis of *E. muscae* 'Berkeley'-induced behavioral manipulations.

Results

Assembling the *E. muscae* 'Berkeley' genome using Chromium linked-reads

Early on, when the size of the *E. muscae* 'Berkeley' genome was still unclear, we sequenced 350 bp and 550 bp Illumina PCR-Free libraries, along with two PacBio SMRT Cells. The Illumina libraries clustered poorly, and much of what did cluster was adapter dimer. But we were able to create a low-coverage *k*-mer frequency spectrum (Figure 4.S1). This indicated that the *E. muscae* 'Berkeley' genome is around 1.3 Gb, and initial exploration of the PacBio reads showed that it is also highly repetitive.

To contend with a large and highly repetitive genome, we assembled the *E. muscae* 'Berkeley' genome with the SuperNova assembler [136] using Chromium linked-reads from 10X Genomics [137, 138]. While this technology uses Illumina short-reads, the library making process preserves long-distance information that can be used to assemble large phase blocks. Table 4.1 reports summary statistics for the 10X assembly. The scaffold N50 is 435,293 bp, and the longest scaffold is 2,251,750 bp. The total scaffold length (assembly size) is 1,235,972,964 bp. The assembly contains approximately 110 Mb of gaps. Contig sizes are significantly shorter, with a contig N50 of 34,915 bp. When the standard N50 calculation is repeated for all integers from 1 to 100, the result is an "Nx plot" [139]. Figure 4.1 includes an Nx plot showing the distribution of scaffold lengths across the assembly, along with a cumulative scaffold length plot. The GC content of the assembly is 41.33 %. We discuss the *E. muscae* genome size, which is not necessarily synonymous with the assembly size, at the end of the Results.

The *E. muscae* 'Berkeley' genome is highly repetitive

We annotated repeats in the *E. muscae* 10X assembly using RepeatModeler [140] and RepeatMasker [47]. Commensurate with its large size, the *E. muscae* genome is highly repetitive. Our initial annotation shows that approximately 83 % of the genome is

repeats (Table 4.2). The majority of repeats fall into various classes of retrotransposons, but there is also a small but significant fraction of DNA elements. Repeats vary in size and age, and based on the transcriptome, some are still active. Subsequent curation suggests that the repeat content is at least 85 % (Michael Eisen, personal communication).

The 10X assembly is missing many known genes

One way to assess the quality of an assembly is by annotation: a good assembly should contain a high percentage of known genes. Benchmarking Universal Single-Copy Orthologs (BUSCOs) are single-copy genes present in more than 90 % of surveyed species [45, 46]. The fungal BUSCO (v. 3.0.2) set contains 290 genes, and is based on a survey of 85 species. Figure 4.2 shows the BUSCO (v. 3.0.2) assessment results for the 10X assembly. Forty percent of BUSCOs are missing or fragmented. For the 60 % of BUSCOs that are complete, roughly half are duplicated.

The BUSCO results are difficult to interpret. Typically, large numbers of missing and fragmented genes indicate problems with the assembly. However, given that *E. muscae* is only distantly related to other sequenced fungal genomes, the BUSCO set might be a poor measure of genes present in *E. muscae* and related species. Many genes also appear to be duplicated. We return to this observation later in the Results.

PacBio long-reads suggest many mis-assemblies in the 10X genome

Early on, when the size of the *E. muscae* genome was still uncertain, we sequenced two PacBio SMRT Cells. Later, we sequenced five more, bringing the total to seven. The relatively high cost of PacBio data precluded sequencing additional SMRT Cells.

After processing all seven SMRT Cells, there were 527,523 circular consensus sequence (CCS) reads totaling 5,396,500,791 bp of sequence. This equates to approximately 4.4x coverage of the 1.24 Gb assembly. Figure 4.3 shows the length distribution of CCS reads. The average read length is 10,230 bp (median = 9,231 bp; maximum = 59,863 bp).

Given the highly repetitive nature of the *E. muscae* genome, long-reads could be advantageous during the initial *de novo* assembly. The Maryland Super Read Cabog Assembler (MaSuRCA) can perform hybrid *de novo* assemblies using Illumina short-reads and PacBio/MinION long-reads [40, 141]. The lower limit for PacBio coverage is generally around 10x, but we were curious if 4.4x coverage could produce a working assembly. However, we were unable to secure a high-memory machine (512 Gb - 1 Tb) that could run an assembly for 3 - 4 weeks. Low-coverage PacBio data can also be used to identify structural variants (SVs) in existing assemblies, using tools like Parliament [142] and PBHoney [143]. Unfortunately, these tools were not designed for large and highly repetitive draft genomes (Andrew Carroll, personal communication).

Given these limitations, we did the next best thing: use the alignment of PacBio long-reads to look for errors in the 10X assembly. Long-reads that fail to align end-to-end could indicate misassemblies.

CCS reads were aligned to a "long" form of the 10X assembly (scaffolds >= 40 kb) using BLASR [144], with settings optimized for the human genome (another large and repetitive genome). Nearly 98 % of the reads aligned to the 10X assembly. Reads that failed to align tended to be short (Figure 4.S2). For reads that did align, BLASR assigned a mapping quality (MAPQ) to each alignment that ranged from 0 to 254 (the best); however, in practice, these qualities were "binary". For example, 481,515 reads aligned with MAPQ=254, while 33,104 reads aligned with MAPQ=0. (The remaining 889 reads fell somewhere in between.) Unlike reads that didn't align at all, MAPQ=0 reads were not short. In fact, there were discernible differences in the read length distributions of MAPQ=0 and MAPQ=254 reads, suggesting problems with the assembly, not the reads themselves (Figure 4.S2).

To evaluate the 10X assembly in granular detail, we parsed the cigar strings of the MAPQ=254 reads. The cigar string is a detailed record of every match, mismatch, insertion, deletion, and soft clip in the alignment. (If the 5' or 3' end of a read doesn't align to the reference, those bases are soft clipped in the alignment, creating a "flap".) Despite the high mapping quality, many of the alignments contain large insertions, deletions, and flaps. For example, as reported in Table 4.3, 198,135 alignments (41 %) have at least a 1 kb flap; and 107,515 alignments (22 %) have at least a 4 kb flap - nearly half the average read length.

There are several reasons why a long-read might not align end-to-end. The "long" 10X assembly contains approximately 108 Mb of gaps, some as long as 60 kb. Gaps of that size are difficult for even the longest of PacBio reads to span, leading to soft clipping at the margins of the gap. Similarly, the fragmented assembly contains 3,677 scaffolds. Long-reads will be soft clipped when an alignment reaches the end of a scaffold. We repeated the above the calculations, but this time ignored flaps that were near ends of scaffolds or large gaps. At most, gaps and scaffold ends account for only 20 % of flaps (Table 4.3). This suggests that the 10X assembly is missing large chunks of the *E. muscae* genome. Given the highly repetitive nature of the genome, these regions could include collapsed repeats and other misassemblies.

To visualize all of the alignments in aggregate, we compared the predicted and calculated accuracies for all MAPQ=254 reads (Figure 4.4). Based on several quality metrics, each PacBio CCS read has a predicted accuracy. The average predicted accuracy is 0.86. For every aligned read, we compared the predicted accuracy to an adjusted calculated accuracy, where we accounted for mismatches and deletions that intersected gaps in the assembly, as well as flaps that occurred near scaffold ends or large gaps. While the calculated accuracies for many reads were similar to their predicted accuracies, the "smear" down the middle of the plot indicates that many reads aligned poorly to the 10X assembly. Taken as a whole, this suggests that the 10X assembly contains many misassemblies.

Read Depth and SNPs within Genes in the 10X Assembly

An initial analysis of the transcriptome assembly showed that there were two closely related haplotypes for many genes. We next wanted to see how these haplotypes were present in the 10X assembly. For simplicity, we initially focused on a small subset of 311 genes that met the following criteria: 1) The Trinity transcriptome assembly produced a single isoform, and 2) Based on a BUSCO (v. 1.1) [45] analysis of the transcriptome, the gene was complete and single-copy. We called such genes single-isoform BUSCOs (SIBs).

We aligned single-isoform BUSCOs to the 10X assembly and identified SIBs that are complete on one scaffold (single-copy), or two different scaffolds (duplicated). Across the 10X assembly, 87 SIBs are single-copy, and 117 are duplicated. (Other SIBs fall outside these categories for a variety of reasons, including being fragmented across one or more scaffolds; complete on one scaffold but fragmented on at least one other scaffold; or duplicated on the same scaffold.) We then mapped short-read Illumina data to the 10X assembly, and compared the average coverage across single-copy and duplicated SIBs (Figure 4.5). Coverage levels within both SIB classes are variable, but the mean coverage for single-copy SIBs is twice that of duplicated SIBs ($p < 5.1e\text{-}55$, Welch's t-test). The duplicated SIBs are also generally clustered around the average coverage across the entire 10X assembly (1X in Figure 4.5). In contrast, single-copy SIBs approach twice the assembly-wide average (2X in Figure 4.5).

Next, we compared the number of unfiltered single nucleotide polymorphisms (SNPs) in single-copy and duplicated SIBs. The difference is striking. There are 2,260 biallelic SNPs spread across 76/87 single-copy SIBs, but only 88 SNPs divided between 16/117 duplicated SIBs (Figure 4.5). Taken together, the coverage and SNP results indicate that single-copy SIBs represent genomic regions where only one of two closely related haplotypes was assembled. When short-reads are aligned to the genome, reads from both haplotypes pileup in the same region, doubling the coverage and vastly increasing the number of SNPs. For duplicated SIBs, both haplotypes were assembled independently and placed on different scaffolds, suggesting that such sequences are not present on the same physical chromosome. Small numbers of SNPs can plausibly arise within duplicated SIBs. For example, the assembler might collapse a repeat present on one haplotype.

To better understand large-scale structures in the genome, we aligned pairs of scaffolds using LASTZ [49]. Figure 4.6 shows the alignment of three pairs of scaffolds that share at least one single-isoform BUSCO, along with two random scaffolds that do not share any SIBs. All scaffold pairs produce a characteristic X-alignment, indicating that the forward strand of the first scaffold also aligns to the reverse complement of the second scaffold. X-alignments were first described in bacteria, where the most likely mechanism involves large chromosomal inversions about the origin of replication or terminus [145]. In *E. muscae*, we think such alignments are likely a product of the genome's dense repeat structure. Due to extensive retrotransposition, any two regions of the genome

share a large number of repetitive elements, inserted in various orientations. In this way, elements on the forward strand of one scaffold can align to similar elements on the reverse complement of another scaffold - even if the scaffolds share no non-repetitive sequence. The density of repeats makes it difficult to compare the similarity of non-coding regions. For the scaffold pairs in Parts A and B, the alignments appear to be driven by more than just random repeats, with relatively long alignment blocks. There are also significant gaps. The scaffold pair in Part C looks no different than random scaffolds, indicating that the alignment is driven largely by repetitive elements.

How Large is the *E. muscae* Genome?

Our initial *k*-mer frequency spectrum suggested the *E. muscae* 'Berkeley' genome might be around 1.3 Gb. The cumulative length of the 10X assembly is also 1.24 Gb. A *de novo* transcriptome assembly showed that many genes contain two distinct haplotypes, and based on our analysis of single-isoform BUSCOS (SIBs) in the 10X genome assembly, many of these haplotypes were assembled independently on different scaffolds. Absent additional biological data, accurately estimating the genome size is difficult - especially considering that polyploidy appears to be common in closely related species [135]. Taken as a whole though, the evidence suggests that the haploid genome size might be around 650 Mb.

Table 4.1. *E. muscae* 10X Assembly Statistics.

Number of Scaffolds	43,011
Total Scaffold Length (bp)	1,235,972,964
Scaffold N50 (bp)	435,293
Longest Scaffold (bp)	2,251,750
Number of Gaps	44,558
Total Gap Length (bp)	110,478,770
Contig N50 (bp)	34,915
Longest Contig (bp)	323,750

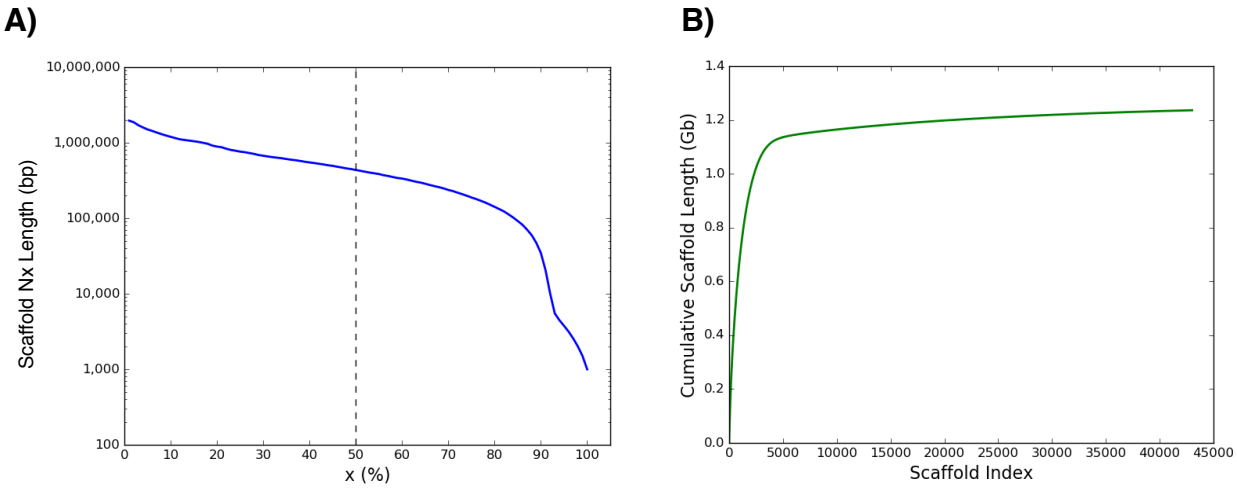


Figure 4.1. Nx plot and cumulative scaffold length plot for the *E. muscae* 10X assembly.
A) To calculate the scaffold N50, scaffold lengths are ordered from longest to shortest, and then summed. The N50 is the scaffold length that brings the sum above 50 % of the total scaffold length (assembly size). When this calculation is repeated for all integers from 1 to 100, the result is an Nx graph [139]. B) Scaffold lengths are ordered from longest to shortest, and summed. The cumulative scaffold length is plotted as a function of the number of summed scaffolds (scaffold index).

Table 4.2. RepeatModeler [140] **/ RepeatMasker** [47] **results for the *E. muscae* 10X assembly.**

Element Type	Number of Elements	Length Occupied (bp)	Percentage of Assembly (%)
SINEs	0	0	0
ALUs	0	0	0
MIRs	0	0	0
LINEs	5,667	5,521,628	0.45
LINE1	2,430	1,590,366	0.13
LINE2	0	0	0
L3/CR1	0	0	0
LTR Elements	244,926	380,724,694	30.80
ERVL	419	173,082	0.01
ERVL-MaLRs	0	0	0
ERV_classI	11,657	12,807,371	1.04
ERV_classII	2,602	943,225	0.08
DNA Elements	150,480	200,904,537	16.25
hAT-Charlie	0	0	0
TcMar-Tigger	0	0	0
Unclassified	554,590	433,490,114	35.07
Total Interspersed Repeats		1,020,640,973	82.58

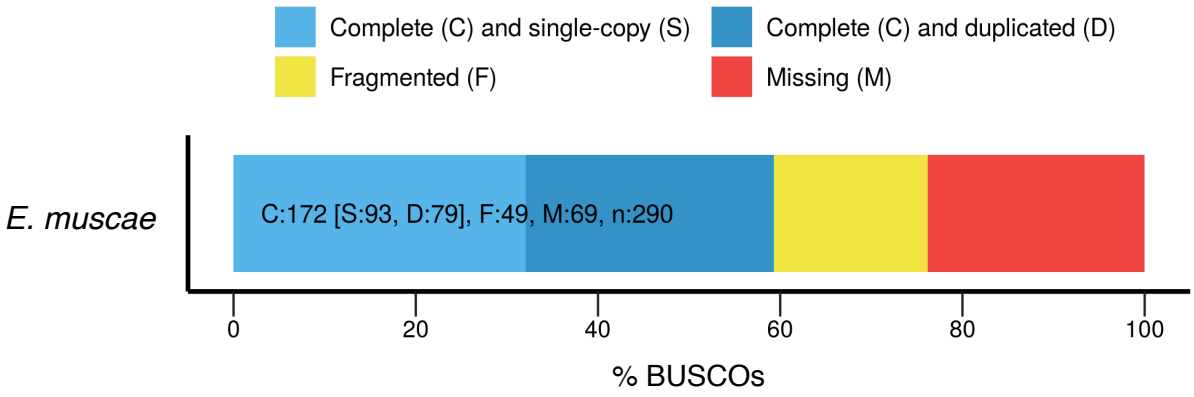


Figure 4.2. Many genes in the *E. muscae* 10X assembly are missing, fragmented, and duplicated.
BUSCO [45, 46] assessment results for the *E. muscae* 10X assembly. The fungal BUSCO set contains 290 genes. The bar graph shows the number of BUSCOs that are complete and single-copy, complete and duplicated, fragmented, and missing.

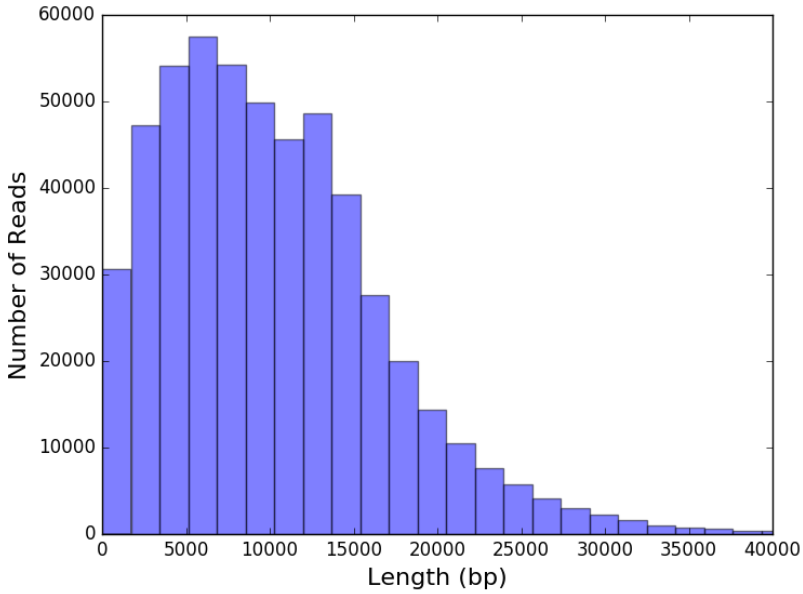


Figure 4.3. Length distribution of *E. muscae* PacBio CCS reads.
Length distribution for 527,523 CCS reads from seven SMRT Cells. The average CCS read length is 10,230 bp (median = 9,231 bp), and the longest is 59,863 bp.

Table 4.3. Many PacBio CCS reads align to the *E. muscae* 10X assembly with large flaps, the vast majority of which cannot be explained by gaps or scaffold ends.
In total, 481,515 PacBio CCS reads aligned with a mapping quality of 254. The numbers below include only MAPQ=254 reads.

Length of 5' or 3' Flap (bp)	Number of Reads with Flap	Number of Reads with Flap Near Scaffold End or Gap	Percentage of Reads with Flap Near Scaffold End or Gap (%)
500	223,331	42,076	18.8
1,000	198,135	39,181	19.8
2,000	160,809	32,490	20.2
4,000	107,515	20,830	19.4

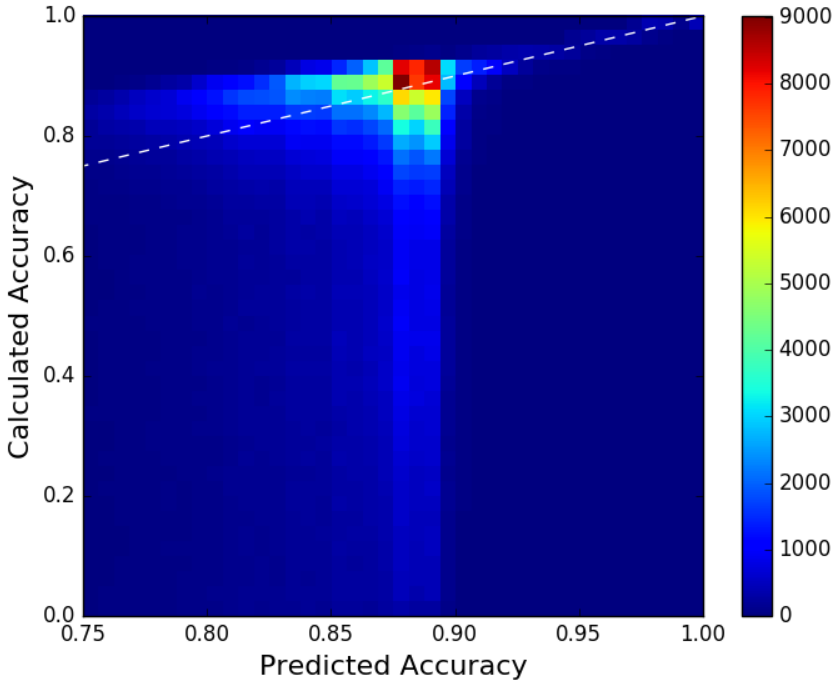


Figure 4.4. The calculated accuracy for many PacBio CCS reads aligned to the *E. muscae* 10X assembly is significantly lower than the predicted accuracy.
For each PacBio CCS read with MAPQ=254, we plotted the predicted accuracy and an adjusted calculated accuracy. The calculated accuracy was equal to the number of matches divided by the sum of matches, mismatches, insertions, deletions, and flaps. Mismatches or deletions that intersected gaps in the 10X assembly were excluded from the calculation, as were flaps that occurred near scaffold ends or large gaps. The dashed line is y = x.

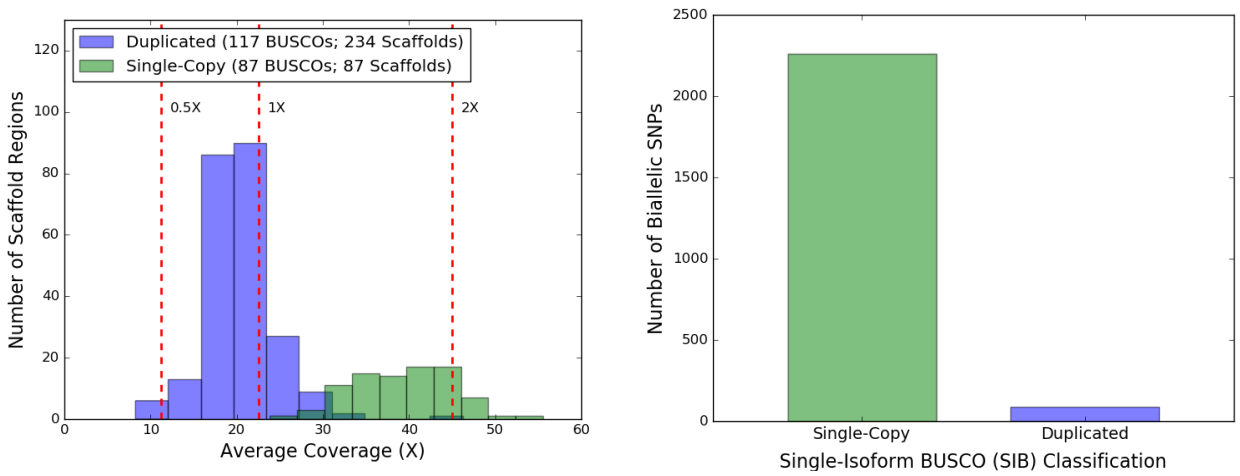


Figure 4.5. Single-copy SIBs have twice the average coverage, and vastly more SNPs, than duplicated SIBs.
Single-isoform BUSCOs (SIBs) (n=311) were aligned to the 10X assembly using BLASTn [59], revealing 87 SIBs that are complete on one scaffold (single-copy) and 117 SIBs that are complete on two different scaffolds (duplicated). A) Illumina short-reads were aligned to the 10X assembly, and used to calculate the average coverage within single-copy and duplicated SIBs. These regions include introns and exons. B) The number of unfiltered biallelic SNPs within single-copy and duplicated SIBs in the 10X assembly.

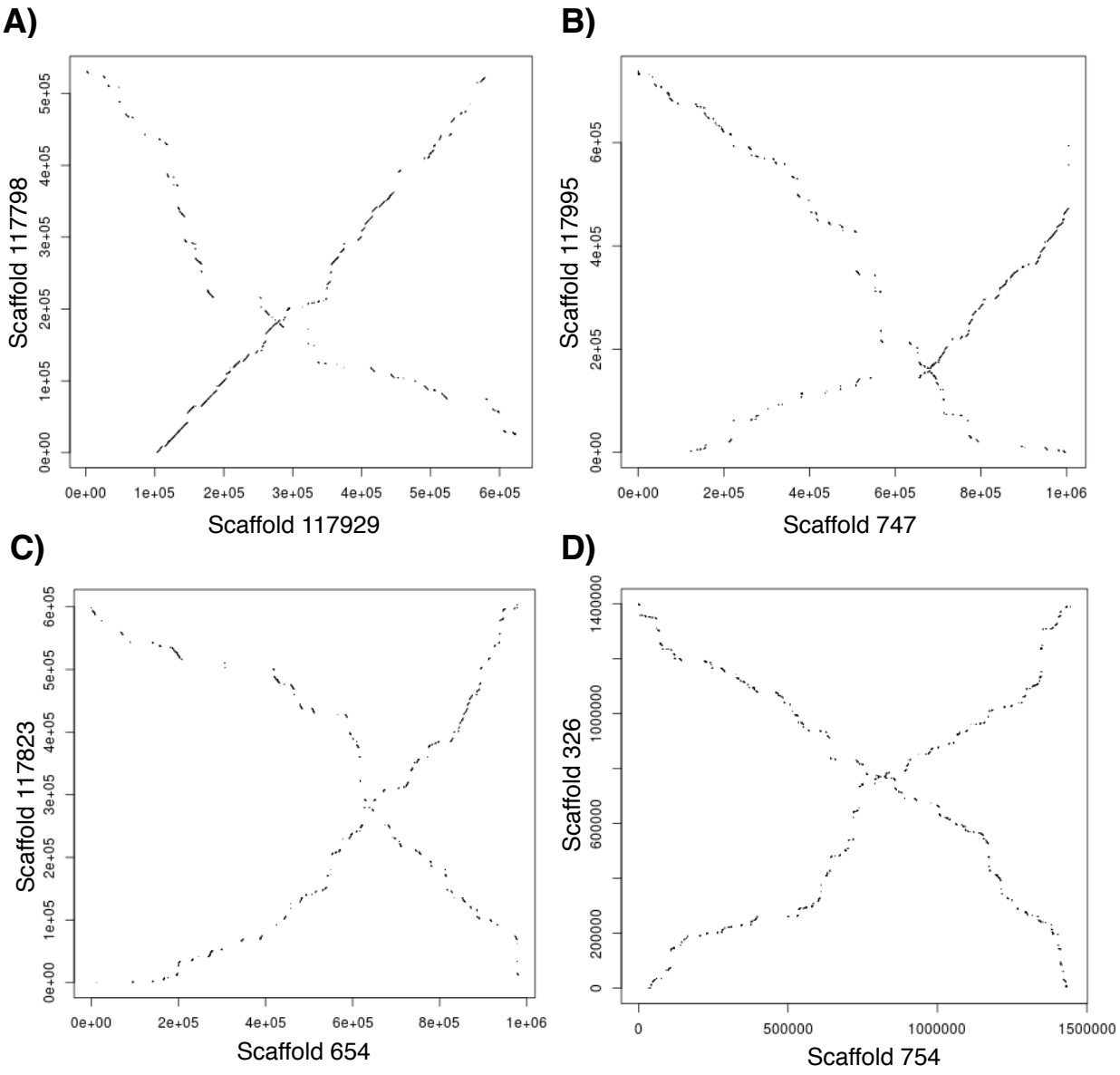


Figure 4.6. The alignment of pairs of *E. muscae* scaffolds with at least one shared single-isoform BUSCO produces X-alignments.
Pairs of scaffolds were aligned using LASTZ [49], and dot plots representing alignment blocks (gap-free segments) were constructed in R. Parts A), B), and C) show alignments between pairs of scaffolds that share at least one duplicated SIB. Efforts were made to select pairs of scaffolds with similar sizes; however, the SIB was not always centrally located on both scaffolds. D) The alignment of random scaffolds that do not share any genes from our small set of SIBs. All scaffold pairs produce a characteristic X-alignment, indicating that the forward strand of the first scaffold also aligns to the reverse complement of the second scaffold.

Discussion

In this chapter, we described the sequencing and assembly of the challenging *E. muscae* 'Berkeley' genome. It is exceptionally large compared to previously sequenced fungi, and highly repetitive (at least 85 %). Our best assembly was generated using technology from 10X Genomics.

The alignment of PacBio long-reads to the 10X assembly suggests that it contains many misassemblies. Nearly one quarter of the alignments include flaps of at least 4 kb, nearly half the average read length. This indicates that scaffolds are missing large pieces of the genome. (It's unlikely flaps arise because of systematic problems with the PacBio reads themselves.) Given the highly repetitive nature of the genome, many of these sequences likely involve repeats. But other lines of evidence suggest that the 10X assembly may also be missing genes.

The BUSCO assessment of the 10X assembly showed that 40 % of known single-copy fungal genes are missing or fragmented. This typically indicates problems with the assembly. However, given that *Entomophthora* is distantly related to previously sequenced fungal species, the BUSCO set might not be a good way to assess the genic content of the assembly. In a separate analysis using single-isoform BUSCOs (SIBs), we found that the average coverage of single-copy SIBs is twice that of duplicated SIBs. Single-copy SIBs also contain large numbers of unfiltered biallelic SNPs. These data indicate that single-copy SIBs represent genomic regions where only one of two closely related haplotypes was assembled. So the 10X assembly does appear to be missing some genic regions.

At this point, generating an improved assembly requires additional sequencing data. If we can secure a high-memory server for 3 - 4 weeks, a hybrid *de novo* assembly using MaSuRCA is an attractive option. This assembler can combine Illumina short-reads with either PacBio long-reads or Oxford Nanopore MinION ultra-long reads (but not both) [40, 141]. So one option is to double our existing PacBio coverage so that it approaches the requisite 10x coverage. Or, we could generate MinION ultra-long reads (which can reach hundreds of kb), and then evaluate the resulting assembly using our low-coverage PacBio data. Given the size and density of repeats, the latter approach might be the most useful.

Going forward, it will also be helpful to annotate the genome - using either our existing 10X assembly, or a new *de novo* assembly. Given that many of the annotated genes will likely be novel, understanding their function will be a challenge.

Future efforts to characterize the *E. muscae* 'Berkeley' genome should also include experimental work. Accurately estimating the genome size is difficult using only sequencing data. The haploid genome size might be around 650 Mb. Flow cytometry could be used to estimate the total amount of DNA in a single nucleus, and ultrastructural studies of mitosis could provide chromosome counts. Previous work

indicates that polyploidy might be common in the *Entomophthoraceae*, with a basal chromosome number of 8 [135, Richard Humber, personal communication]. Additionally, all cell types (e.g., protoplasts, conidiophores, and conidia) appear to be multinucleate. One intriguing possibility is that different nuclei within the same cell differ genetically, a phenomenon known as heterokaryon. Given that the number of nuclei per cell type varies within a somewhat broad range, the partitioning of nuclei does not appear to be strictly controlled. This probably makes heterokaryon unlikely. To answer this question experimentally, future work should focus on isolating and sequencing individual nuclei from the same cell.

Going forward, the *E. muscae* 'Berkeley' genome will support our ongoing efforts to understand the mechanistic and molecular basis of behavioral manipulation. It will also aid future comparative studies aimed at understanding the evolution of genome size and repeat content within this ancient fungal lineage.

Materials and Methods

Initial Illumina data and 10X assembly

We ran into several practical difficulties when we tried to sequence this genome. We first created 350 bp and 550 bp Illumina TruSeq DNA PCR-Free libraries, and attempted to cluster them on a HiSeq2500 System. The libraries clustered poorly, and up to 40 % of what did cluster was adapter dimer. We created a *k*-mer frequency spectrum using the remaining data. *k*-mers were counted using Jellyfish (v. 2.2.6) [146] with *k*=31. Based on a method from [147], the estimated genome size is roughly 1.3 Gb. Next, a PCR-amplified library was created, and run a HiSeq4000 System. This library appeared to be significantly biased, probably because of the PCR amplification.

A 10X Chromium [137, 138] library was prepared by the UC Davis Genomics Core. The library was first clustered on an Illumina HiSeq4000 System, generating 150 bp paired-end reads. Later, the same library was clustered on an Illumina HiSeq2500 System, generating 150 bp paired-end reads. The second round of sequencing was done at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. All results in this chapter describe assembly and analysis from the first round of sequencing. Reads were assembled using Supernova [136], with the output option --style=pseudohap. This creates a single record for each scaffold.

Repeat Annotation and BUSCO Analysis of the 10X Genome Assembly

Repeats in the 10X assembly were annotated using RepeatModeler [140] and RepeatMasker [47].

The 10X assembly was searched for known genes using BUSCO (v. 3.0.2) [45, 46], with the profile library fungi_odb9. This BUSCO set contains 290 genes. The following

options were specified in the configuration file: mode = genome, evalue = 1e-3, limit = 3, and long = False. The BUSCO plot was constructed using the included script generate_plot.py.

PacBio Data

We sequenced seven SMRT Cells on a PacBio RSII Instrument using P6v2/C4v2 chemistry. The resulting SMRT Cell movies (.bax.h5 files) were converted to unaligned PacBio BAM files using bax2bam (v. 0.0.8) with the options --subread and --pulsefeatures=DeletionQV,DeletionTag,InsertionQV,IPD,MergeQV,SubstitutionQV, PulseWidth,SubstitutionTag. Subreads from the same ZMW were merged into a single circular consensus sequence (CCS) using the ccs program (v. 2.1.0 (commit a256e12-dirty)) with the options --maxLength=60000, --minPasses=0, --minPredictedAccuracy=0.75, --polish, and --richQVs.

CCS reads were aligned to the "long" version of our *E. muscae* 10X assembly using BLASR (v. 5.3.9c6f0a5) [144], with the following options from Chaisson et al. [148]: --bestn 1, --maxAnchorsPerPosition 100, --affineAlign, --affineOpen 100, --affineExtend 0, --insertion 5, --deletion 5, --extend, --maxExtendDropoff 20, --clipping soft, and --bam.

The adjusted calculated accuracy of an alignment was equal to the number of matches divided by the sum of matches, mismatches, insertions, deletions, and flaps. Mismatches or deletions that intersected gaps in the 10X assembly were excluded from the calculation. Soft clipping (flaps) were also ignored if they occurred within 150 bp of a scaffold end, or within 50 bp of the boundary of a large gap (on either side).

Single-isoform BUSCO Analysis

BUSCO (v. 1.1b1) [45] was run on the Trinity transcriptome assembly with the options -l fungi and -m trans. The fungi BUSCO set contained 1,438 genes. If the Trinity transcriptome assembly produced a single isoform, and the BUSCO analysis showed the gene to be complete and single-copy, we called it a single-isoform BUSCO (SIB). (Genes with multiple splice isoforms are sometimes marked as duplicated in a BUSCO analysis.) There were 311 SIBs.

Single-isoform BUSCOs (SIBs) were aligned to the 10X assembly using BLASTn (v. 2.2.31+) [59] with an Expect value (E) cutoff of 1e-9. The output was parsed to identify SIBs that were complete on one scaffold, or on two different scaffolds. Reads from a 10X library run on an Illumina HiSeq2500 System were adapter trimmed, and then aligned to the 10X assembly using Bowtie2 (v. 2.2.3) [70] with the options --local and -X 800. Per-base coverage values were calculated using pileup (BBMap v. 36.11) [67] with the options 32bit=t, delcoverage=f, secondary=f, and basecov. Unfiltered SNPs were called with samtools (v. 0.1.19-96b5f2294a) [79] mpileup with the options -u and -g, followed by bcftools (v. 0.1.19-96b5f2294a) with the options -v, -m, and -O z.

Pairs of scaffolds containing the same SIB were aligned using LASTZ (v. 1.03.73) [49], with the options --chain and --format=rdotplot. Alignment blocks (gap-free segments) were plotted in R.

Supporting Information

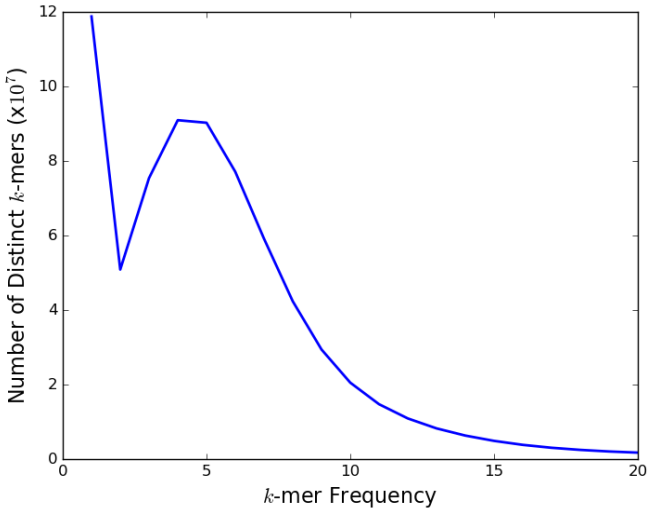


Figure 4.S1. *k*-mer frequency spectrum (*k*=31) for *E. muscae* 'Berkeley'.
The *k*-mer frequency spectrum was generated using data from Illumina TruSeq DNA PCR-Free libraries (350 bp and 550 bp) clustered on a HiSeq2500 System. The libraries clustered poorly, hence the low coverage. *k*-mers were counted using Jellyfish [146] with *k*=31. Using a genome estimation method from [147], the *E. muscae* 'Berkeley' genome is approximately 1.3 Gb.

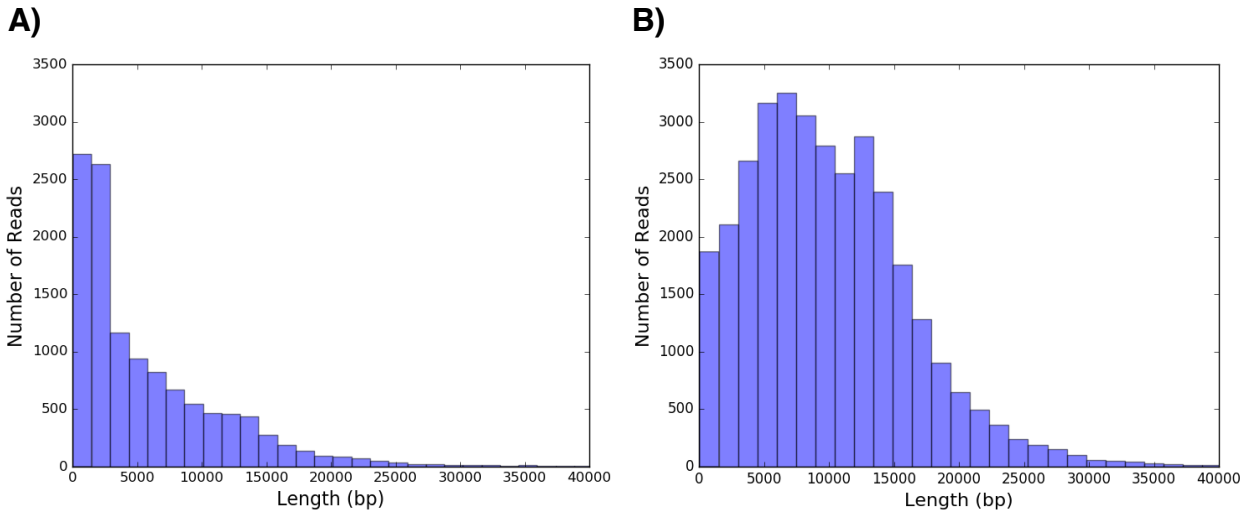


Figure 4.S2. PacBio CCS reads that did not align to the *E. muscae* 10X assembly tend to be short, whereas aligned reads with MAPQ=0 are full-length reads.
A) The length distribution of 11,965 unaligned PacBio CCS reads. B) The length distribution of 33,104 reads that aligned with a mapping quality (MAPQ) of zero. Mapping qualities were assigned by BLASR [144].

Bibliography

1. Levine, M., Transcriptional enhancers in animal development and evolution. Current biology : CB, 2010. **20**(17): p. R754-R763.
2. Rubinstein, M. and F.S.J. de Souza, *Evolution of transcriptional enhancers and animal diversity.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences. **368**(1632): p. 20130017-20130017.
3. Sakabe, N.J., D. Savic, and M.A. Nobrega, Transcriptional enhancers in development and disease. Genome Biology, 2012. **13**(1): p. 238.
4. Banerji, J., S. Rusconi, and W. Schaffner, Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell, 1981. **27**(2 Pt 1): p. 299-308.
5. Benoist, C. and P. Chambon, In vivo sequence requirements of the SV40 early promotor region. Nature, 1981. **290**(5804): p. 304-10.
6. Bulger, M. and M. Groudine, Looping versus linking: toward a model for long-distance gene activation. Genes Dev, 1999. **13**(19): p. 2465-77.
7. Amano, T., T. Sagai, H. Tanabe, Y. Mizushina, H. Nakazawa, and T. Shiroishi, Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev Cell, 2009. **16**(1): p. 47-57.
8. Calhoun, V.C. and M. Levine, Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9878-83.
9. Stanojevic, D., S. Small, and M. Levine, Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. Science, 1991. **254**(5036): p. 1385-7.
10. Goto, T., P. Macdonald, and T. Maniatis, Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell, 1989. **57**(3): p. 413-22.
11. Štanojević, D., T. Hoey, and M. Levine, Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Krüppel in Drosophila. Nature, 1989. **341**: p. 331.
12. Small, S., R. Kraut, T. Hoey, R. Warrior, and M. Levine, Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev, 1991. **5**(5): p. 827-39.
13. Small, S., A. Blair, and M. Levine, Regulation of even-skipped stripe 2 in the Drosophila embryo. Embo j, 1992. **11**(11): p. 4047-57.
14. Arnosti, D.N., S. Barolo, M. Levine, and S. Small, The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development, 1996. **122**(1): p. 205-14.
15. Andrioli, L.P., V. Vasisht, E. Theodosopoulou, A. Oberstein, and S. Small, Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. Development, 2002. **129**(21): p. 4931-40.
16. Liang, H.-L., C.-Y. Nien, H.-Y. Liu, M.M. Metzstein, N. Kirov, and C. Rushlow, The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature, 2008. **456**: p. 400.
17. Ludwig, M.Z. and M. Kreitman, Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Mol Biol Evol, 1995. **12**(6): p. 1002-11.
18. Ludwig, M.Z., N.H. Patel, and M. Kreitman, Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development, 1998. **125**(5): p. 949-58.
19. Ludwig, M.Z., C. Bergman, N.H. Patel, and M. Kreitman, Evidence for stabilizing selection in a eukaryotic enhancer element. Nature, 2000. **403**(6769): p. 564-7.
20. Tamura, K., S. Subramanian, and S. Kumar, Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol, 2004. **21**(1): p. 36-44.

21. Hare, E.E., B.K. Peterson, V.N. Iyer, R. Meier, and M.B. Eisen, Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet, 2008. **4**(6): p. e1000106.
22. Lusk, R.W. and M.B. Eisen, Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. PLoS Genet, 2010. **6**(1): p. e1000829.
23. Swanson, C.I., N.C. Evans, and S. Barolo, Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev Cell, 2010. **18**(3): p. 359-70.
24. Swanson, C.I., D.B. Schwimmer, and S. Barolo, Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol, 2011. **21**(14): p. 1186-96.
25. Farley, E.K., K.M. Olson, W. Zhang, D.S. Rokhsar, and M.S. Levine, Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proceedings of the National Academy of Sciences, 2016. **113**(23): p. 6508.
26. Arnold, C.D., D. Gerlach, C. Stelzer, L.M. Boryn, M. Rath, and A. Stark, *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
27. Arnold, C.D., D. Gerlach, D. Spies, J.A. Matts, Y.A. Sytnikova, M. Pagani, N.C. Lau, and A. Stark, Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet, 2014. **46**(7): p. 685-92.
28. Thanos, D. and T. Maniatis, Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell, 1995. **83**(7): p. 1091-100.
29. Panne, D., T. Maniatis, and S.C. Harrison, An atomic model of the interferon-beta enhanceosome. Cell, 2007. **129**(6): p. 1111-23.
30. Kulkarni, M.M. and D.N. Arnosti, Information display by transcriptional enhancers. Development, 2003. **130**(26): p. 6569-75.
31. Arnosti, D.N. and M.M. Kulkarni, Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem, 2005. **94**(5): p. 890-8.
32. Junion, G., M. Spivakov, C. Girardot, M. Braun, E.H. Gustafson, E. Birney, and E.E. Furlong, *A transcription factor collective defines cardiac cell fate and reflects lineage history.* Cell, 2012. **148**(3): p. 473-86.
33. Brake, I. and G. Bachli, *World catalogue of insects. 9, 9*. 2008, Stenstrup: Apollo Books.
34. Yassin, A., E.K. Delaney, A.J. Reddiex, T.D. Seher, H. Bastide, N.C. Appleton, J.B. Lack, J.R. David, S.F. Chenoweth, J.E. Pool, and A. Kopp, *The pdm3 Locus Is a Hotspot for Recurrent Evolution of Female-Limited Color Dimorphism in Drosophila.* Curr Biol, 2016. **26**(18): p. 2412-2422.
35. Kellermann, V., B. van Heerwaarden, C.M. Sgro, and A.A. Hoffmann, Fundamental evolutionary limits in ecological traits drive Drosophila species distributions. Science, 2009. **325**(5945): p. 1244-6.
36. Ramniwas, S. and B. Kajla, Divergent strategy for adaptation to drought stress in two sibling species of montium species subgroup: Drosophila kikkawai and Drosophila leontia. J Insect Physiol, 2012. **58**(12): p. 1525-33.
37. Chen, C.C., M. Watada, H. Miyake, T.K. Katoh, Z. Sun, Y.F. Li, M.G. Ritchie, and S.Y. Wen, Courtship patterns in the Drosophila montium species subgroup: repeated loss of precopulatory courtship? Zoolog Sci, 2013. **30**(12): p. 1056-62.
38. Chen, Z.X., D. Sturgill, J. Qu, H. Jiang, S. Park, N. Boley, A.M. Suzuki, A.R. Fletcher, D.C. Plachetzki, P.C. FitzGerald, C.G. Artieri, J. Atallah, O. Barmina, J.B. Brown, K.P. Blankenburg, E. Clough, A. Dasgupta, S. Gubbala, Y. Han, J.C. Jayaseelan, D. Kalra, Y.A. Kim,

C.L. Kovar, S.L. Lee, M. Li, J.D. Malley, J.H. Malone, T. Mathew, N.R. Mattiuzzo, M. Munidasa, D.M. Muzny, F. Ongeri, L. Perales, T.M. Przytycka, L.L. Pu, G. Robinson, R.L. Thornton, N. Saada, S.E. Scherer, H.E. Smith, C. Vinson, C.B. Warner, K.C. Worley, Y.Q. Wu, X. Zou, P. Cherbas, M. Kellis, M.B. Eisen, F. Piano, K. Kionte, D.H. Fitch, P.W. Sternberg, A.D. Cutter, M.O. Duff, R.A. Hoskins, B.R. Graveley, R.A. Gibbs, P.J. Bickel, A. Kopp, P. Carninci, S.E. Celniker, B. Oliver, and S. Richards, *Comparative validation of the D. melanogaster modENCODE transcriptome annotation.* Genome Res, 2014. **24**(7): p. 1209-23.
39. Allen, S.L., E.K. Delaney, A. Kopp, and S.F. Chenoweth, *Single-Molecule Sequencing of the Drosophila serrata Genome.* G3 (Bethesda), 2017. **7**(3): p. 781-788.
40. Zimin, A.V., G. Marcais, D. Puiu, M. Roberts, S.L. Salzberg, and J.A. Yorke, *The MaSuRCA genome assembler.* Bioinformatics, 2013. **29**(21): p. 2669-77.
41. Earl, D., K. Bradnam, J. St John, A. Darling, D. Lin, J. Fass, H.O. Yu, V. Buffalo, D.R. Zerbino, M. Diekhans, N. Nguyen, P.N. Ariyaratne, W.K. Sung, Z. Ning, M. Haimel, J.T. Simpson, N.A. Fonseca, I. Birol, T.R. Docking, I.Y. Ho, D.S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M.C. Schatz, D.R. Kelley, A.M. Phillippy, S. Koren, S.P. Yang, W. Wu, W.C. Chou, A. Srivastava, T.I. Shaw, J.G. Ruby, P. Skewes-Cox, M. Betegon, M.T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F.J. Ribeiro, S. Yin, T. Sharpe, G. Hall, P.J. Kersey, R. Durbin, S.D. Jackman, J.A. Chapman, X. Huang, J.L. DeRisi, M. Caccamo, Y. Li, D.B. Jaffe, R.E. Green, D. Haussler, I. Korf, and B. Paten, *Assemblathon 1: a competitive assessment of de novo short read assembly methods.* Genome Res, 2011. **21**(12): p. 2224-41.
42. Bradnam, K.R., J.N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J.A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.C. Chou, J. Corbeil, C. Del Fabbro, T.R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N.A. Fonseca, G. Ganapathy, R.A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J.B. Hiatt, I.Y. Ho, J. Howard, M. Hunt, S.D. Jackman, D.B. Jaffe, E.D. Jarvis, H. Jiang, S. Kazakov, P.J. Kersey, J.O. Kitzman, J.R. Knight, S. Koren, T.W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M.D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T.D. Otto, B. Paten, O.S. Paulo, A.M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F.J. Ribeiro, S. Richards, D.S. Rokhsar, J.G. Ruby, S. Scalabrin, M.C. Schatz, D.C. Schwartz, A. Sergushichev, T. Sharpe, T.I. Shaw, J. Shendure, Y. Shi, J.T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B.M. Vieira, J. Wang, K.C. Worley, S. Yin, S.M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I.F. Korf, *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* Gigascience, 2013. **2**(1): p. 10.
43. Yandell, M. and D. Ence, A beginner's guide to eukaryotic genome annotation. Nat Rev Genet, 2012. **13**(5): p. 329-42.
44. *NCBI Drosophila serrata Annotation Release 100*. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Drosophila_serrata/100/.
45. Simao, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 2015. **31**(19): p. 3210-2.
46. Waterhouse, R.M., M. Seppey, F.A. Simao, M. Manni, P. Ioannidis, G. Klioutchnikov, E.V. Kriventseva, and E.M. Zdobnov, *BUSCO applications from quality assessments to gene prediction and phylogenomics.* Mol Biol Evol, 2017.
47. Smit, A.F.A., R. Hubley, and P. Green. *RepeatMasker Open-4.0*. 2013-2015; Available from: <http://www.repeatmasker.org>.
48. Benson, G., Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 1999. **27**(2): p. 573-80.

49. Harris, R.S., *Improved pairwise alignment of genomic DNA*. 2007, The Pennsylvania State University.
50. Kent, W.J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11484-9.
51. Hinrichs, A.S., D. Karolchik, R. Baertsch, G.P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T.S. Furey, R.A. Harte, F. Hsu, J. Hillman-Jackson, R.M. Kuhn, J.S. Pedersen, A. Pohl, B.J. Raney, K.R. Rosenbloom, A. Siepel, K.E. Smith, C.W. Sugnet, A. Sultan-Qurraie, D.J. Thomas, H. Trumbower, R.J. Weber, M. Weirauch, A.S. Zweig, D. Haussler, and W.J. Kent, *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D590-8.
52. Kvon, E.Z., T. Kazmar, G. Stampfel, J.O. Yanez-Cuna, M. Pagani, K. Schernhuber, B.J. Dickson, and A. Stark, *Genome-scale functional characterization of Drosophila developmental enhancers in vivo.* Nature, 2014. **512**(7512): p. 91-5.
53. Zhang, Z., N. Inomata, M.L. Cariou, J.L. Da Lage, and T. Yamazaki, Phylogeny and the evolution of the Amylase multigenes in the Drosophila montium species subgroup. J Mol Evol, 2003. **56**(2): p. 121-30.
54. Da Lage, J.-L., G.J. Kergoat, F. Maczkowiak, J.-F. Silvain, M.-L. Cariou, and D. Lachaise, A phylogeny of Drosophilidae using the Amyrel gene: questioning the Drosophila melanogaster species group boundaries. 2007. **45**(1): p. 47-63.
55. Miyake, H. and M. Watada, Molecular phylogeny of the Drosophila auraria species complex and allied species of Japan based on nuclear and mitochondrial DNA sequences. Genes Genet Syst, 2007. **82**(1): p. 77-88.
56. Yang, Y., Z.C. Hou, Y.H. Qian, H. Kang, and Q.T. Zeng, Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the Drosophila melanogaster species group (Drosophilidae, Diptera). Mol Phylogenet Evol, 2012. **62**(1): p. 214-23.
57. Chen, H., Z. Xu, C. Mei, D. Yu, and S. Small, A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes. Cell, 2012. **149**(3): p. 618-29.
58. Stamatakis, A., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 2014. **30**(9): p. 1312-3.
59. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden, *BLAST+: architecture and applications.* BMC bioinformatics, 2009. **10**: p. 421-421.
60. Goltsman, E., I. Ho, and D. Rokhsar, Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes. arXiv, 2017.
61. Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh, *Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.* Genome Res, 2014. **24**(8): p. 1384-95.
62. Salzberg, S.L., A.M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T.J. Treangen, M.C. Schatz, A.L. Delcher, M. Roberts, G. Marcais, M. Pop, and J.A. Yorke, *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome Res, 2012. **22**(3): p. 557-67.
63. Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T.D. Otto, *REAPR: a universal tool for genome assembly evaluation.* Genome Biol, 2013. **14**(5): p. R47.
64. Walker, B.J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C.A. Cuomo, Q. Zeng, J. Wortman, S.K. Young, and A.M. Earl, *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PLoS One, 2014. **9**(11): p. e112963.

65. Andrews, S. *FastQC: A quality control tool for high throughput sequence data*. 2010; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
66. Simpson, J.T., Exploring genome characteristics and sequence quality without a reference. Bioinformatics, 2014. **30**(9): p. 1228-35.
67. Bushnell, B. *BBMap*. Available from: sourceforge.net/projects/bbmap/.
68. Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. 2013. **4**(237).
69. Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, and J. Wang, *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.* Gigascience, 2012. **1**(1): p. 18.
70. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012. **9**(4): p. 357-9.
71. Li, H. *seqtk*. 2013; Available from: https://github.com/lh3/seqtk.
72. Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, and I. Birol, *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.
73. Bankevich, A., S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P.A. Pevzner, *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
74. Safonova, Y., A. Bankevich, and P.A. Pevzner, dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. J Comput Biol, 2015. **22**(6): p. 528-45.
75. Zerbino, D.R. and E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008. **18**(5): p. 821-9.
76. Vezzi, F., G. Narzisi, and B. Mishra, Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. PLoS One, 2012. **7**(12): p. e52210.
77. Paulino, D., R.L. Warren, B.P. Vandervalk, A. Raymond, S.D. Jackman, and I. Birol, *Sealer: a scalable gap-closing application for finishing draft genomes.* BMC Bioinformatics, 2015. **16**: p. 230.
78. Martin, M., M. Patterson, S. Garg, S.O. Fischer, N. Pisanti, G.W. Klau, A. Schoenhuth, and T. Marschall, *WhatsHap: fast and accurate read-based phasing.* bioRxiv, 2016.
79. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
80. *Picard*. Available from: http://broadinstitute.github.io/picard.
81. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo, *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
82. Katoh, K., K. Misawa, K. Kuma, and T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res, 2002. **30**(14): p. 3059-66.
83. Katoh, K., K. Kuma, H. Toh, and T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res, 2005. **33**(2): p. 511-8.
84. Katoh, K. and D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol, 2013. **30**(4): p. 772-80.

85. Paradis, E., J. Claude, and K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics, 2004. **20**(2): p. 289-90.
86. Tamura, K. and M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol, 1993. **10**(3): p. 512-26.
87. Dermitzakis, E.T. and A.G. Clark, Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol, 2002. **19**(7): p. 1114-21.
88. Moses, A.M., D.A. Pollard, D.A. Nix, V.N. Iyer, X.Y. Li, M.D. Biggin, and M.B. Eisen, *Large-scale turnover of functional transcription factor binding sites in Drosophila.* PLoS Comput Biol, 2006. **2**(10): p. e130.
89. Doniger, S.W. and J.C. Fay, Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol, 2007. **3**(5): p. e99.
90. Bradley, R.K., X.-Y. Li, C. Trapnell, S. Davidson, L. Pachter, H.C. Chu, L.A. Tonkin, M.D. Biggin, and M.B. Eisen, Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. PLOS Biology, 2010. **8**(3): p. e1000343.
91. Paris, M., T. Kaplan, X.Y. Li, J.E. Villalta, S.E. Lott, and M.B. Eisen, Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression. PLOS Genetics, 2013. **9**(9): p. e1003748.
92. Hertz, G.Z. and G.D. Stormo, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 1999. **15**(7-8): p. 563-77.
93. Li, X.-y., S. MacArthur, R. Bourgon, D. Nix, D.A. Pollard, V.N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C.L.L. Hendriks, H.C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S.E. Celniker, D.W. Knowles, T. Gingeras, T.P. Speed, M.B. Eisen, and M.D. Biggin, *Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm.* PLOS Biology, 2008. **6**(2): p. e27.
94. Harrison, M.M., X.-Y. Li, T. Kaplan, M.R. Botchan, and M.B. Eisen, Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. PLOS Genetics, 2011. **7**(10): p. e1002266.
95. Revell, L.J., phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol Evol, 2012. **3**: p. 217-223.
96. Landis, M.J., J.G. Schraiber, and M. Liang, Phylogenetic analysis using Levy processes: finding jumps in the evolution of continuous traits. Syst Biol, 2013. **62**(2): p. 193-204.
97. Elliot, M.G. and A.O. Mooers, Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC Evol Biol, 2014. **14**: p. 226.
98. Landis, M.J. and J.G. Schraiber, Pulsed evolution shaped modern vertebrate body sizes. Proc Natl Acad Sci U S A, 2017. **114**(50): p. 13224-13229.
99. Avise, J.C. and T.J. Robinson, Hemiplasy: a new term in the lexicon of phylogenetics. Syst Biol, 2008. **57**(3): p. 503-7.
100. Hahn, M.W. and L. Nakhleh, Irrational exuberance for resolved species trees. Evolution, 2016. **70**(1): p. 7-17.
101. Mendes, F.K. and M.W. Hahn, Gene Tree Discordance Causes Apparent Substitution Rate Variation. Syst Biol, 2016. **65**(4): p. 711-21.
102. Berman, B.P., Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, and M.B. Eisen, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A, 2002. **99**(2): p. 757-62.

103. Bergman, C.M., J.W. Carlson, and S.E. Celniker, Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics, 2005. **21**(8): p. 1747-9.
104. Zhu, L.J., R.G. Christensen, M. Kazemian, C.J. Hull, M.S. Enuameh, M.D. Basciotta, J.A. Brasefield, C. Zhu, Y. Asriyan, D.S. Lapointe, S. Sinha, S.A. Wolfe, and M.H. Brodsky, *FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system.* Nucleic acids research, 2011. **39**(Database issue): p. D111-D117.
105. Noyes, M.B., X. Meng, A. Wakabayashi, S. Sinha, M.H. Brodsky, and S.A. Wolfe, A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res, 2008. **36**(8): p. 2547-60.
106. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 2010. **26**(6): p. 841-842.
107. Elya, C., T.C. Lok, Q.E. Spencer, H. McCausland, C.C. Martinez, and M. Eisen, Robust manipulation of the behavior of Drosophila melanogaster by a fungal pathogen in the laboratory. eLife, 2018. **7**: p. e34414.
108. Andersen, S.B., S. Gerritsma, K.M. Yusah, D. Mayntz, N.L. Hywel-Jones, J. Billen, J.J. Boomsma, and D.P. Hughes, *The life of a dead ant: the expression of an adaptive extended phenotype.* Am Nat, 2009. **174**(3): p. 424-33.
109. Berdoy, M., J.P. Webster, and D.W. Macdonald, Fatal attraction in rats infected with Toxoplasma gondii. Proc Biol Sci, 2000. **267**(1452): p. 1591-4.
110. Ingram, W.M., L.M. Goodrich, E.A. Robey, and M.B. Eisen, Mice infected with low-virulence strains of Toxoplasma gondii lose their innate aversion to cat urine, even after extensive parasite clearance. PLoS One, 2013. **8**(9): p. e75246.
111. Dawkins, R., *The extended phenotype : the gene as the unit of selection*. 1982, Oxford Oxfordshire ; San Francisco: Freeman. viii, 307 p.
112. Cohn, F., Empusa muscae und die Krankheit der Stubenfliegen. Hedwigia, 1855. **1**: p. 57-61.
113. Steinkraus, D.C. and J.P. Kramer, Susceptibility of sixteen species of Diptera to the fungal pathogen Entomophthora muscae (Zygomycetes: Entomophthoraceae). Mycopathologia, 1987. **100**: p. 55-63.
114. Balazy, S., On rhizoids of Entomophthora muscae (Cohn) Fresenius (Entomophthorales: Entomophthoraceae). Mycotaxon, 1984. **19**: p. 397-407.
115. Krasnoff, S.B., D.W. Watson, D.M. Gibson, and E.C. Kwan, Behavioral effects of the entomopathogenic fungus, Entomophthora muscae on its host Musca domestica: postural changes in dying hosts and gated pattern of mortality. Journal of Insect Physiology, 1995. **41**(10): p. 895-903.
116. Mullens, B.A. and J.L. Rodriguez, Dynamics of Entomophthora muscae (Entomophthorales: Entomophthoraceae) conidial discharge from Musca domestica (Diptera: Muscidae) cadavers. Environmental Entomology, 1985. **14**: p. 317-322.
117. Mullens, B.A., J.L. Rodriguez, and J.A. Meyer, An epizootiological study of Entomophthora muscae in Muscoid fly populations on southern California poultry facilities, with emphasis on Musca domestica. Hilgardia, 1987. **55**(3): p. 1-41.
118. Møller, A.P., A fungus infecting domestic flies manipulates sexual behavior of its host. Behavioral Ecology and Sociobiology, 1993. **33**: p. 403-407.
119. Keller, S., *Entomophthora muscae als Artenkomplex.* Mitteilungen der Schweizerischen Entomologischen Gesellschaft, 1984. **57**: p. 131-132.
120. Keller, S., Arthropod-pathogenic Entomophthorales of Switzerland. I. Conidiobolus, Entomophaga and Entomophthora. Sydowia, 1987. **40**: p. 122-167.

121. Keller, S., V. Kalsbeek, and J. Eilenberg, Redescription of Entomophthora muscae (Cohn) Fresenius. Sydowia, 1999. **51**: p. 197-209.
122. Keller, S., The genus Entomophthora (Zygomycetes, Entomophthorales) with a description of five new species. Sydowia, 2002. **54**: p. 157-197.
123. Brobyn, P.J. and N. Wilding, Invasive and developmental processes of Entomophthora muscae infecting houseflies (Musca domestica). Transactions of the British Mycological Society, 1983. **80**(1): p. 1-8.
124. Watson, D.W., B.A. Mullens, and J.J. Petersen, Behavioral fever response of Musca domestica (Diptera: Muscidae) to infection by Entomophthora muscae (Zygomycetes: Entomophthorales). Journal of Invertebrate Pathology, 1993. **61**: p. 10-16.
125. Kalsbeek, V., B.A. Mullens, and J.B. Jespersen, Field studies of Entomophthora (Zygomycetes: Entomophthorales) — induced behavioral fever in Musca domestica (Diptera: Muscidae) in Denmark. Biological Control, 2001. **21**: p. 264-273.
126. Steinkraus, D.C., C.J. Geden, and D.A. Rutz, Prevalence of Entomophthora muscae (Cohn) Fresenius (Zygomycetes: Entomophthoraceae) in house flies (Diptera: Muscidae) on dairy farms in New York and induction of epizootics. Biological Control, 1993. **3**: p. 93-100.
127. Zurek, L., D. Wes Watson, S.B. Krasnoff, and C. Schal, Effect of the entomopathogenic fungus, Entomophthora muscae (Zygomycetes: Entomophthoraceae), on sex pheromone and other cuticular hydrocarbons of the house fly, Musca domestica. J Invertebr Pathol, 2002. **80**(3): p. 171-6.
128. Watson, D.W. and J.J. Petersen, Sexual activity of male Musca domestica (Diptera: Muscidae) infected with Entomophthora muscae (Entomophthoraceae: Entomophthorales). Biological Control, 1993. **3**: p. 22-26.
129. Goldstein, B., An Empusa disease of Drosophila. Mycologia, 1927. **19**(3): p. 97-109.
130. Turian, G. and J. Wuest, Mycoses ä Entomophthoracees frappant des populations de Fourmis et de Drosophiles. Mitteilungen der Schweizerischen Entomologischen Gesellschaft, 1969. **42**: p. 197-201.
131. Chang, Y., S. Wang, S. Sekimoto, A.L. Aerts, C. Choi, A. Clum, K.M. LaButti, E.A. Lindquist, C. Yee Ngan, R.A. Ohm, A.A. Salamov, I.V. Grigoriev, J.W. Spatafora, and M.L. Berbee, *Phylogenomic Analyses Indicate that Early Fungi Evolved Digesting Cell Walls of Algal Ancestors of Land Plants.* Genome Biology and Evolution, 2015. **7**(6): p. 1590-1601.
132. Gryganskyi, A.P., R.A. Humber, M.E. Smith, J. Miadlikowska, S. Wu, K. Voigt, G. Walther, I.M. Anishchenko, and R. Vilgalys, *Molecular phylogeny of the Entomophthoromycota.* Mol Phylogenet Evol, 2012. **65**(2): p. 682-94.
133. Gryganskyi, A.P., R.A. Humber, M.E. Smith, K. Hodge, B. Huang, K. Voigt, and R. Vilgalys, *Phylogenetic lineages in Entomophthoromycota.* Persoonia, 2013. **30**: p. 94-105.
134. Murrin, F., J. Holtby, R.A. Noland, and W.S. Davidson, The genome of Entomophaga aulicae (Entomophthorales, Zygomycetes): Base composition and size. Experimental Mycology, 1986. **10**(1): p. 67-75.
135. Humber, R.A., Strongwellsea vs. Erynia: the case for a phylogenetic classification of the Entomophthorales (Zygomycetes). Mycotaxon, 1982. **15**: p. 167-184.
136. Weisenfeld, N.I., V. Kumar, P. Shah, D.M. Church, and D.B. Jaffe, *Direct determination of diploid genome sequences.* Genome Res, 2017. **27**(5): p. 757-767.
137. Zheng, G.X.Y., B.T. Lau, M. Schnall-Levin, M. Jarosz, J.M. Bell, C.M. Hindson, S. Kyriazopoulou-Panagiotopoulou, D.A. Masquelier, L. Merrill, J.M. Terry, P.A. Mudivarti, P.W. Wyatt, R. Bharadwaj, A.J. Makarewicz, Y. Li, P. Belgrader, A.D. Price, A.J. Lowe, P. Marks, G.M. Vurens, P. Hardenbol, L. Montesclaros, M. Luo, L. Greenfield, A. Wong, D.E. Birch, S.W. Short, K.P. Bjornson, P. Patel, E.S. Hopmans, C. Wood, S. Kaur, G.K. Lockwood, D. Stafford, J.P. Delaney, I. Wu, H.S. Ordonez, S.M. Grimes, S. Greer, J.Y. Lee, K. Belhocine, K.M. Giorda, W.H.

Heaton, G.P. McDermott, Z.W. Bent, F. Meschi, N.O. Kondov, R. Wilson, J.A. Bernate, S. Gauby, A. Kindwall, C. Bermejo, A.N. Fehr, A. Chan, S. Saxonov, K.D. Ness, B.J. Hindson, and H.P. Ji, *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.* Nature biotechnology, 2016. **34**(3): p. 303-311.
138. Marks, P., S. Garcia, A. Martinez Barrio, K. Belhocine, J. Bernate, R. Bharadwaj, K. Bjornson, C. Catalanotti, J. Delaney, A. Fehr, B. Galvin, H. Heaton, J. Herschleb, C. Hindson, E. Holt, C.B. Jabara, S. Jett, N. Keivanfar, S. Kyriazopoulou-Panagiotopoulou, M. Lek, B. Lin, A. Lowe, S. Mahamdallie, S. Maheshwari, T. Makarewicz, J. Marshall, F. Meschi, C. keefe, H. Ordonez, P. Patel, A. Price, A. Royall, E. Ruark, S. Seal, M. Schnall-Levin, P. Shah, S. Williams, I. Wu, A. Wei Xu, N. Rahman, D. MacArthur, and D.M. Church, *Resolving the Full Spectrum of Human Genome Variation using Linked-Reads.* bioRxiv, 2017.
139. Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler, *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-5.
140. Smit, A.F.A. and R. Hubley. *RepeatModeler Open-1.0*. 2008-2015; Available from: <http://www.repeatmasker.org>.
141. Zimin, A.V., D. Puiu, M.C. Luo, T. Zhu, S. Koren, G. Marcais, J.A. Yorke, J. Dvorak, and S.L. Salzberg, Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res, 2017. **27**(5): p. 787-792.
142. English, A.C., W.J. Salerno, O.A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D.I. Ritter, C.R. Beck, C.F. Davis, M. Dahdouli, S. Ma, A. Carroll, N. Veeraraghavan, J. Bruestle, B. Drees, A. Hastie, E.T. Lam, S. White, P. Mishra, M. Wang, Y. Han, F. Zhang, P. Stankiewicz, D.A. Wheeler, J.G. Reid, D.M. Muzny, J. Rogers, A. Sabo, K.C. Worley, J.R. Lupski, E. Boerwinkle, and R.A.J.B.G. Gibbs, *Assessing structural variation in a personal genome—towards a human reference diploid genome.* 2015. **16**(1): p. 286.
143. English, A.C., W.J. Salerno, and J.G.J.B.B. Reid, PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. 2014. **15**(1): p. 180.
144. Chaisson, M.J. and G.J.B.B. Tesler, Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. 2012. **13**(1): p. 238.
145. Eisen, J.A., J.F. Heidelberg, O. White, and S.L. Salzberg, Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome biology, 2000. **1**(6): p. RESEARCH0011-RESEARCH0011.
146. Marcais, G. and C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 2011. **27**(6): p. 764-70.
147. Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O.A. Ryder, F.C.-C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C.C. Steiner, T.T.-Y. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M.W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T.-W. Lam, S.-M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G.K.-S. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang and J. Wang, *The sequence and de novo assembly of the giant panda genome.* Nature, 2009. **463**: p. 311.

148. Chaisson, M.J.P., J. Huddleston, M.Y. Dennis, P.H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J.M. Landolin, J.A. Stamatoyannopoulos, M.W. Hunkapiller, J. Korlach, and E.E. Eichler, *Resolving the complexity of the human genome using single-molecule sequencing.* Nature, 2015. **517**(7536): p. 608-611.