2018-04-01

Genomic Structural Variation Across Five Continental Populations of Drosophila melanogaster

Evan Michael Long
Brigham Young University

Genomic Structural Variation Across Five Continental Populations of *Drosophila melanogaster*

Evan Michael Long

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

John Chaston, Chair
Joshua Udall
Eric Jellen
Stephen Piccolo

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

Genomic Structural Variation Across Five Continental Populations
of *Drosophila melanogaster*

Evan Michael Long
Department of Plant and Wildlife Sciences, BYU
Master of Science

Chromosomal structure variations (SV) including insertions, deletions, inversions, and translocations occur within the genome and can have a significant effect on organismal phenotype. Some of these effects are caused by structural variations containing genes. Modern sequencing using short reads makes the detection of large structural variations (> 1kb) very difficult. Large structural variations represent a significant amount of the genetic diversity within a population. We used a global sampling of *Drosophila melanogaster* (Ithaca, Zimbabwe, Beijing, Tasmania, and Netherlands) to represent diverse populations. We used long-read sequencing and optical mapping technologies to identify SVs in these genomes. Because the average read length used for these approaches are much longer than traditional short read sequencing, these maps facilitate the identification of chromosomal SVs of greater size and with more clarity. We found a wide diversity of structural variations in each of the five strains. These structural variations varied greatly in size and location, and significantly affected exonic regions of the genome. Structural variations accounted for a much larger difference in number of base pairs between strains than single nucleotide polymorphisms (SNPs).

Keywords: structural variation (SV), long-read sequencing, optical mapping

ACKNOWLEDGEMENTS

I'd like to thank my mentor Joshua Udall for all his support and guidance in pursuing this project as well as my education. I am also very grateful for the support from the other members of my committee and the department. Additionally, I must thank Carrie Evans for maintaining fly populations and refining DNA extraction protocols.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Genomic Structural Variation Across Five Continental Populations of *Drosophila Melanogaster*

Evan Michael Long, Carrie Evans, John Chaston, Joshua Udall
Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT

BACKGROUND

Comparative Genomics and Evolutionary Diversity

For the past twenty years, it has been the goal of geneticists to sequence and assemble species' genomes. It appears to some as the culminating achievement of the field to complete these genome sequences; however, as we have learned more about genetic diversity we have found that there are significant differences among populations within a species [1]. Genomes can vary in structure and sequence within a species through random mutations and alterations, being reinforced by evolutionary pressures [2]. To more fully understand genomic functions, it is necessary to compare diverse populations within a species.

The Characteristics and Difficulties of Structural Variations

A structural variation (SV) is commonly characterized as a change in a region of 50 bp or larger compared to another DNA sequence [3]. Genomic alterations categorized as structural variations include insertions, deletions, duplications, translocations, and inversions (Figure 1-1). Any structural alteration smaller than 50 bp is considered an indel [4, 5]. Although very similar, many studies will refer to SVs as copy number variants (CNVs), but this term applies to a subset of SVs including deletions, insertions, and duplications. Chromosomal rearrangments represent a large portion of the genetic diversity within a population, accounting for two- to four-fold greater locus-specific mutation frequency than single nucleotide polymorphisms [5, 6]. This means that on average more base pairs are changed through structural variation than by point mutations [3].

It is understood that SVs can have very pronounced effects on phenotype. On a chromosome level, nondisjunction disorders such as Down syndrome are very well characterized. While many SV-related diseases are known, few are well understood or characterized. Chromosomal rearrangements have been implicated in autism spectrum disorder, cancer, schizophrenia, epilepsy, and Parkinson's disease [5, 7]. Studies in *D. melanogaster* have implicated CNVs in wide array of phenotypic characteristics [8, 9]. The prominent mechanism by which SVs result in phenotypic effects is through gene dosage by interrupting the promoter or enhancer elements associated with the affected gene. Chromosomal SVs may also be the cause of a gene duplication, which would result in a similar effect [4, 5].

Although the significance of SVs is becoming more understood, there remains a large obstacle to their study. This obstacle resides in the low SV detection capabilities of current short-read, next generation sequencing techniques [3]. Because most sequencing methods involve mapping short reads to a reference genome assembly, the algorithms for assembly often fail to detect SVs of substantial size [10]. This is especially true with respect to insertions, as most algorithms favor calling deletions [4]. One method of overcoming this problem is to treat every genome assembly as a *de novo* assembly.

Long-Read Technologies

In recent years there have been developed long-read sequencing technologies as well as the single-molecule, nanochannel-based, genome optical mapping technology by BioNano. It begins with high molecular weight DNA that has been treated carefully to retain long segments. The long-read sequencing technology most prevalent today is produced by Pacific Biosciences and is called single molecule real time sequencing (SMRT). It has the ability of obtaining sequence reads with average lengths >10kb.

The optical mapping technologies by BioNano use modified enzymes to label endonuclease-nicked lesions with fluorescently tagged nucleotides. The DNA backbone is counterstained then flowed and imaged through nanochannels. These images are then overlapped analyzed and stitched together into an optical map of the genome. Among the benefits of using BioNano genomic maps is the ability to easily identify large structural variations that would otherwise be neglected, because its molecule lengths are a minimum of 150kb before assembly [4, 11].

The high molecular weight DNA used for these technologies also enables the spanning and proper detection of SVs that may fall in regions of repetitive elements, which are areas of some interest because of their propensity for unequal crossing over [12]. The combination of PacBio sequencing and BioNano mapping technologies is ideal for analyzing SVs because of their power and resolution.

Drosophila melanogaster: A Genetic Model Organism

Drosophila melanogaster is a species of fruit fly that has been a model organism for multicellular eukaryotic genetics since the beginning of the 20th century. Many important genes in human development such as "sonic hedgehog" and "Wnt" were first discovered *in D. melanogaster* [13]. Its genome was one of the first completed animal genomes, being sequenced and published in 2000 [14]. It exists in a large variety of ecosystems making it a species of interest for population genetic studiesdue to its global genetic diversity profile. Some of the characteristics that make it ideal as a genetic model include fast reproduction, simple containment and management, low chromosome number, and considerable gene homology to humans [13, 15].

Consequently *Drosophila melanogaster has* been a central model in studying genetic systems as well as evolutionary and population genetic processes [16]. The *D melanogaster* genome has also been well annotated, making varied analyses possible [14, 17–20]. To capture the diversity of the *D. melanogaster* population, we are using stable lines representative of a variety of geographical locations including Zimbabwe, Ithaca (New York), the Netherlands, Beijing, and Tasmania. These lines have been inbred for several generations to ensure homozygosity and purity for regional characterization. Previous studies have used strains from these areas to represent the diverse global population [16, 21]. These researchers previously performed low-coverage illumina sequencing on these strains to identify SNPs and small indels. Interestingly, they found that the strain from Zimbabwe was the most differentiated and diverse from the other strains- a situation mirroring the genetic diversity in human African populations (Figure 1-2).

Because past and current genomic methods have overlooked the prevalence of large SVs, we wanted to investigate the diversity of SVs among populations to determine their significance within a species. As mentioned, sequencing efforts previously performed could not properly evaluate the prevalence of SVs due to the shortness of the reads. The global diversity panel of *D. melanogaster* strains may present diverse, chromosomal architecture changes correlating to their evolutionary divergence from each other.

INTRODUCTION

Genome structural variations or rearrangements (SV) are thought to play a critical role in plant and animal diversity and speciation. Structural variations are characterized as differences larger than 50 bp between two aligned genomes [22]. Many structural variations can be found among different individuals within the same species [1]. These variants can include insertions, deletions, duplications, translocations, and inversions[4]. Given their size, they are more likely to disrupt gene function than single-nucleotide variants (SNVs), making them contribute significantly to phenotypes and pathology [5]. Although many studies refer to SVs as copy number variants (CNVs), the common usage of the term "CNVs" generally applies to a subset of SVs including deletions, insertions, and duplications discovered in short-read resequencing. Because of the short length of the reads, the exact nature of the duplications or deletions can remain ambiguous [10].

Genome evolution and diversity is most often thought to act primarily through the occurrence of single nucleotide polymorphisms (SNPs). However, SVs have been found to account for two to four-fold greater locus-specific mutation frequency than single nucleotide polymorphisms [5, 6]. This implies that on average more base pairs are changed through structural variation than by point mutations [22]. Although researchers are finding an increased appreciation for SVs [7, 8], significant limitations remain for SV detection using short sequencing reads [10]. Because most SV detection methods involve mapping relatively short reads to a reference genome assembly, the algorithms for detection struggle to detect SVs larger than the read length [4]. This is especially true with respect to insertions, as the algorithms favor calling deletions [4]. Perhaps, SVs on a different scale also contribute to the genetic diversity between species.

A key element to understanding the nature of sequence SVs is having long-read data to span variations, especially in repetitive regions. Chromosomal rearrangements are more common in repetitive areas, which pose difficulties to short-read SV detection [22]. To overcome this limitation and to investigate SVs on a novel scale, we used PacBio long read sequencing paired with BioNano optical mapping to assess SVs across five different strains of *Drosophila melanogaster*.

D. melanogaster originated on the African continent ~5.4 million years ago and is now ubiquitous across the globe, enabling intraspecies comparisons of flies derived from diverse geographic locations [16, 23]. This makes it an ideal model for research in systems biology and population diversity. We used representatives of *D. melanogaster* collected on five different continents to represent diverse strains from around the globe (Table 1-1) [16]. These strains were selected from a previous study assessing SNP diversity [16].

Additionally, *Drosophila* has been a model in the study of chromosomal rearrangements for over a century [24]. Structural differentiation visible at the cytogenetic level has been extensively studied in *D. melanogaster* and other species of the genus Drosophila [25]. Translocations, inversions, duplications, and deficiencies are well documented in the literature. These rearrangements have been associated with ecological adaptation, fitness, divergence, and speciation [26, 27].

We present 5 high quality genome assemblies using long read sequencing paired with optical maps. We also assess the diversity of chromosomal structural variations and their potential impacts. Our study provides insights into the evolution of chromosomal architecture within *D. melanogaster* and offers insight into the nature of genome evolution.

METHODS

Optical Mapping DNA extraction

Before the extraction the flies were starved for 2 hours to reduce the number of contaminating reads that would be obtained from gut-associated bacteria. High molecular weight DNA was extracted from adult *D. melanogaster* by first grinding ~100-200 whole flies to a rough powder with a mortar and pestle in liquid nitrogen. The powder was suspended in homogenization buffer (10 mM Tris HCl pH 7.5, 60 mM NaCl, 10 mM EDTA, 5% sucrose) and disrupted with a 40 mL Dounce homogenizer before filtering through a 100 micron (VWR cat. # 21008-949) and 40 micron (VWR cat. # 21008-950) nylon mesh sequentially. The resulting pellet was resuspended in 200 uL of resuspension buffer (10 mM Tris HCl pH 7.5, 60 mM NaCl, 10 mM EDTA) and combined with 2% low melting agarose. The mixture was aliquoted into 80 uL plugs and placed in a 4°C fridge until solid. The agarose plugs were incubated with 200 µL proteinase K (QIAGEN, cat. # 158920) and 2.5 mL lysate solution (BioNano Prep Lysis Buffer, 20255) overnight and treated with RNase A (QIAGEN, cat. # 158924, 80 µL/mL) as described in BioNano protocol documentation (BioNano Prep Blood DNA Isolation Protocol, Document Number: 30033). DNA was extracted from the agarose plugs by melting and treating the plugs with agarase (Bio-Rad, cat. # 1703594).

SMRT DNA Extraction

We obtained high molecular weight DNA for single molecule real-time (SMRT) sequencing using a Qiagen genome-tip kit (Cat No./ID: 10243), because the previously explained method could not provide sufficient quantity. We used a modified a extraction protocol outlined in a previous study [28]. First, ~200 adult flies were ground in liquid nitrogen and transferred into 9.5 mL of buffer G2 with 38 µL of RNAse A (100 mg/ml) and 500 µL of proteinase K (QIAGEN, cat. # 158920). The solution was then incubated overnight at 50°C. It was then centrifuged at 5000 x g for 10 minutes at 4°C. The solution was then purified, washed, and eluted using the Qiagen genome-tip kit instructions. Sequencing libraries were created by shearing DNA to 35 kb on a Megaruptor (Diagenode) and selecting for 18-50 kb using a Blue-Pippin (Blue Pippin system, Sage Science, Beverly, MA, USA). DNA was then sequenced using a Sequel machine (Pacific Biosciences, Inc.) at the Brigham Young University DNA sequencing center.

Assembly and Scaffolding

PacBio reads were assembled using CANU assembler V1.4. Assemblies were then scaffolded using optical maps with the Solve-hybrid-scaffold pipeline created by BioNano Genomics. Scaffolded assemblies were uploaded to "Assemblytics" [29] for alignment to the reference genome and detection of structural variants. For whole genome collinearity analyses genomes were scaffolded into whole chromosome arms using the reference genome and the Solve-hybrid-scaffold pipeline previously mentioned.

Analysis of structural variants

To evaluate the structural evolution in the populations of *D. melanogaster* we analyzed the coincidence of structural variations with other genomic features. This was done by primarily using bedtools and the function "IntersectBed" (Supplemental Methods 1) [30]. Whole genome alignments were created using minimap2 and minidot (Supplemental Methods 1) [31]. We evaluated the evolutionary distance between the strains using coincidence using a short R script with the "pvclust" package (Supplemental Methods 1) [32].

Optical Mapping

To visualize the DNA molecules each sample underwent a labeling process that marks a specific hexameric sequence recognized by the restriction enzyme *BssS*I, along each DNA strand. Each molecule was nicked by *BssS*I, labeled with fluorescently labeled nucleotides, repaired to prevent breakage, and counterstained. The process is described in detail in BioNano protocol documentation (BioNano Prep™ Labeling - NLRS Protocol, Document Number: 30024). The samples were then loaded into flow cells where each individual DNA molecule was moved through nanochannels using electrophoresis and their fluorescence was imaged. We completed an average of four complete cycles for each strain, each cycle containing several thousand images.

The data from each DNA molecule were compiled using BioNano software. Based on distances between fluorescent labels of each DNA molecule assemblies of each genome were created. Each BioNano assembly was created using over 100X coverage of molecules with minimum length of 150 kb. The assemblies were aligned with the published *Drosophila* *melanogaster* version 5 reference genome for identification of structural variations using BioNano SVdetect [11].

RESULTS

Sequence and Optical Assemblies

We created high-quality sequence and optical map assemblies for each of the five global strains of *D. melanogaster.* The estimated genome size of *D. melanogaster* is around 180 Mb, with one-third composed of highly repetitive, heterochromatic sequence. The reference genome of *D. melanogaster* has correctly assembled two-thirds of the genome representing the euchromatic regions. The variance in genome assembly size could be derived from the varied success in assembling these repetitive, heterochromatic regions or the assembly of some residual heterozygosity. The sequence assemblies vary in quality however all have a contig N50 great than 1 Mb (Table 1-2). The assembly of strain T29A represents the most contiguous sequence assembly. High quality of each optical map was assured by a > 90% rate of mapping to the assembled pseudomolecules of *D. melanogaster* ISO1 release 5 genome. Each genome has a high BUSCO score validating completeness of the genomes by detecting the presence of widely conserved orthologous genes. Optical maps were aligned to each genome assembly to scaffold and improve assembly contiguity (Table 1-2). Whole genome alignments between each strain display the collinearity and completion of each assembly (Figure 1-3). The large amount of collinearity across the chromosome arms confirms the likelihood of correctly assembled genomes. The one exception being strain T29A, where we see a few large areas of possible translocation and inversion, however its sequence assembly is the most continuous of the five assemblies, and we believe its differences to be biological rather than mis-assembly. The high-quality of these assemblies, shown by their high contiguity and alignment, allowed us to confidently perform further analyses into the depth of variation between the genomes.

Chromosome Structural Variation

Both sequence assemblies and optical maps were aligned to the *D. melanogaster* release 5 reference genome for the detection of structural variations. We used this older version of the reference to make reference points consistent with previous work in these strains of *D. melanogaster*. In all our analyses structural variations are defined as discrepancies >50 bp between the assembly and the reference. Structural rearrangements were detected using both sequence alignment and optical map alignment methods (Figure 1-4). The "Assemblytics" software classified SVs into insertions, deletions, tandem expansions, tandem contractions, repeat contraction, and repeat expansion [29]. The optical map alignment detected insertions and deletions independently from the sequence alignment. The low resolution of optical mapping only allows for the detection of very large SVs (>1000 bp) (Figure 1-5) [4]. We see a higher frequency of insertions and deletions not associated with tandem or repetitive elements. Both SV detection methods display a balanced frequency between insertions and deletions, a feat rendered difficult by short read sequencing which favors calling deletions [10]. We also compared the long-read sequence SVs to the previously performed short-read SV detection [16], and found that many of the long-read SVs were undetected using short-read sequencing, especially in regards to insertions (Figure 1-5). Whole genome alignments revealed larger variations including inversions and translocations (Figure 1-3). Most visible are the large inversions and translocations located on chromosome 3 of T29A. By examining the coincidence of SVs between each strain, we were able to build an evolutionary tree. Like previously published studies, this tree places ZH26 as the most differentiated from the other strains.

Genome Evolution

We next evaluated the extent to which these structural variations affected the exonic regions of genome. To ensure the accurate calling of these SVs, we only used SVs that were validated by the independent optical map SV detection (Figure 1-6). This may result in an underrepresentation of exonic SVs, because the optical maps only detect SVs >1 kb. We calculated the number of base pairs affected by these SVs and concluded that there are many genes impacted by structural rearrangements (Figure 1-7). This total length of exonic sequence affected by SVs is greater than SNPs found in these strains [16]. Although there were originally more insertions and deletions without repetitive or tandem elements, we see exons to be much more likely to contain a repeat contraction or expansion SV than the other types.

By examining the relative density of each SV type we were able to display their relative patterns of occurrence (Figure 1-8). From this distribution we can visually detect some trends in SV location. To build the evolutionary relationships between these strains we evaluated the coincidence of SVs between each of these strains (Figure 1-9). This tree demonstrates the structural evolution between each of the five strains.

DISCUSSION

In this study we used a powerful combination of assessing the genetic diversity of structural variations in global populations of *D. melanogaster* using long-read PacBio sequencing paired with optical mapping. This allowed us to find large structural variations, previously invisible to short-read through sequencing. Other studies have been done which describe the relative visibility of these SVs to short read sequencing [33]. We claim a high confidence in the observed structural variation, because of the independent identification derived from the high molecular weight DNA sequencing and optical mapping methods.

Many sequencing projects today are acknowledging the importance of obtaining more than one high quality genome to understand a species' diversity. The construction of a panel of genomes, known as a pan-genome, renders a more complete image of the genome. Although this concept began in bacteria, it is being applied to eukaryotic organisms including plants such as rice [34]. In this study, Zhao et al. found multiple previously unknown domestication events by creating a large panel of *De novo* rice accession genomes. For *D. melanogaster,* there has been one additional reference level genome assembled by Chakraborty et al. in addition to the original ISO1 reference [33]. Their assembly of the A4 strain allowed for the detection of previously hidden genetic variation, including the discovery of multiple genes with varied copy numbers. Our five assemblies of the globally diverse strains of *D. melanogaster* builds upon these resources to enhance our understanding of its genome.

We report a high frequency and variability of chromosomal structural rearrangements within the *D. melanogaster* species across five continental populations. This panel of genomes assembled with long-read technologies is a resource for the investigation into the nature of the evolution of chromosome structure. Among the high variability in structural rearrangements we found a significant amount coinciding with gene coding regions of the genome.

There has been serious investigation into the impact of SVs on the divergence and evolution of species [35]. Previous work has shown the retention of SVs to be due to either genetic drift or positive selection. Although we expect that some of the SVs presented here could be the product of positive selection, it remains a task for the future to obtain evidence for such events [36]. Chromosomal rearrangements that impact genes provide testable hypotheses with respect to mechanisms of positive selection, and direct functional tests of gene expression level and consequence phenotypic impact can be relatively straightforward. The alteration of gene number by SVs has been associated with speciation in Drosophila [37]. Although inversions are less likely to have a genic effect, they can influence the recombination between species, creating reproductive isolation [27]. The global setting for these strains gives important adaptive context to these SVs. It is more likely for species undergoing migration to contain few variants of large effect [38]. The consistency between the evolutionary relationships found in out SV coincidence data (Figure 1-9) and previous work suggests a regular frequency of SVs[16] . Using expected mutation frequencies [23], previously produced SNP data[16], and our SV data, we postulate that SVs in *D. melanogaster* occur at a rate of ~50/MY/Mb.

We propose that these SVs have a substantial impact on species evolution and divergence. The large size and diversity of these SVs within a single species leads us to predict these features lead to diversity of a species. Further studies into the patterns of structural variation could serve to discover the extent of this evolutionary impact.

CONCLUSION

The populations of *D. melanogaster* used in this study were sampled from five continents around the globe. They represent the high diversity that can be found within a species. As we sequenced, assembled, and analyzed one aspect of the variation between these strains, we have begun to capture a more accurate image of the genome. Today, researchers are beginning to appreciate the concept of a ‘Pangenome’ or the analysis of more than one individual to understand a genome.

Much of our current understanding of genetic diversity and evolution within a species has relied upon the detection of single nucleotide polymorphisms (SNPs) and small indels (insertions and deletions <50bp in length). The financial barrier to *De novo* sequencing and assembling of genomes limited us to only mapping short reads against reference genomes to assess genetic differences. The advent of long-read technologies such as PacBio sequencing and BioNano optical mapping has increased our ability to capture genetic variation due to large structural variations (SVs) including insertions, deletions, duplications, translocations, and inversions. These SVs are >50bp in length and have been shown to account for more variation in base pairs than SNPs.

We report a high frequency and variability of chromosomal structural rearrangements within the *D. melanogaster* species across the five global populations. This panel of genomes assembled with long-read technologies is a resource for the investigation into the nature of the evolution of chromosome structure. We found that there were some patterns of occurrence based on location and SV identity across the *D. melanogaster genome* (Figure 1-8). Although the strains can differ greatly, there are noticeable, retained trends across the chromosome arms.

Among the high variability in structural rearrangements we found a significant amount coinciding with gene coding regions of the genome (Figure 1-7). Many of the SVs affecting exonic regions of the genome were repeat contractions and expansions. These repeat type SVs suggest the prevalence of gene copy number variations. We also found that these SVs accounted for a very large number of base pairs compared to the total gene coding portion of the genome.

Whole genome alignment of our assemblies also allowed us to detect large inversions and translocation present between the strains (Figure 1-3). Most visible are the large inversions and translocations located on chromosome 3 of T29A. The large SVs found between genomes has been hypothesized to play a strong selective role in speciation. This is due to the difficulty in chromosome pairing during meiosis between highly varied genomes.

Our investigation into the frequency, diversity, and implications of large SVs gives a powerful perspective on species' genomic variation. This panel of diverse populations of *D. melanogaster* can be a valuable resource for further investigation into the nature and effects chromosome evolution.

LITERATURE CITED

1. Hardison, R. C., Sharp, P., Li, W., Guigo, R., and Beckstrom-Sternberg, S. "Comparative Genomics" *PLoS Biology* 1, no. 2 (2003): e58. doi:10.1371/journal.pbio.0000058, Available at http://dx.plos.org/10.1371/journal.pbio.0000058

2. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor, G. L., Miklos, Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S. B., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. "Comparative Genomics of the Eukaryotes" *Science* 287, no. 5461 (2000): Available at http://science.sciencemag.org/content/287/5461/2204.full

3. Alkan, C., Coe, B. P., and Eichler, E. E. "Genome Structural Variation Discovery and Genotyping" *Nature Reviews Genetics* 12, no. 5 (2011): 363–376. doi:10.1038/nrg2958, Available at http://www.nature.com/doifinder/10.1038/nrg2958

4. Cao, H., Hastie, A. R., Cao, D., Lam, E. T., Sun, Y., Huang, H., Liu, X., Lin, L., Andrews, W., Chan, S., Huang, S., Tong, X., Requa, M., Anantharaman, T., Krogh, A., Yang, H., Cao, H., and Xu, X. "Rapid Detection of Structural Variation in a Human Genome Using Nanochannel-Based Genome Mapping Technology" *GigaScience* 3, no. 1 (2014): 34. doi:10.1186/2047-217X-3-34, Available at https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-3-34

5. Lupski, J. R. "Genomic Rearrangements and Sporadic Disease" *Nature Genetics* 39, no. 7s (2007): S43–S47. doi:10.1038/ng2084, Available at

http://www.nature.com/doifinder/10.1038/ng2084

6. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., Macdonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. "ARTICLES Global Variation in Copy Number in the Human Genome" doi:10.1038/nature05329, Available at http://www.nature.com/nature/journal/v444/n7118/pdf/nature05329.pdf

7. Stankiewicz, P. and Lupski, J. R. "Structural Variation in the Human Genome and Its Role in Disease" *Annual Review of Medicine* 61, no. 1 (2010): 437–455. doi:10.1146/annurev-med-100708-204735, Available at http://www.annualreviews.org/doi/10.1146/annurev-med-100708-204735

8. Massouras, A., Waszak, S. M., Albarca-Aguilera, M., Hens, K., Holcombe, W., Ayroles, J. F., Dermitzakis, E. T., Stone, E. A., Jensen, J. D., Mackay, T. F. C., and Deplancke, B. "Genomic Variation and Its Impact on Gene Expression in Drosophila Melanogaster." *PLoS genetics* 8, no. 11 (2012): e1003055. doi:10.1371/journal.pgen.1003055, Available at http://www.ncbi.nlm.nih.gov/pubmed/23189034

9. Zhou, J., Lemos, B., Dopman, E. B., and Hartl, D. L. "Copy-Number Variation: The Balance between Gene Dosage and Expression in Drosophila Melanogaster." *Genome biology and evolution* 3, (2011): 1014–24. doi:10.1093/gbe/evr023, Available at http://www.ncbi.nlm.nih.gov/pubmed/21979154

10. Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. "Computational Tools for Copy Number

Variation (CNV) Detection Using next-Generation Sequencing Data: Features and Perspectives" *BMC Bioinformatics* 14, no. Suppl 11 (2013): S1. doi:10.1186/1471-2105-14-S11-S1, Available at http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1

11. Mak, A. C. Y., Lai, Y. Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K. Y., Poon, A., Mostovoy, Y., Hastie, A. R., Stedman, W., Anantharaman, T., Andrews, W., Zhou, X., Pang, A. W. C., Dai, H., Chu, C., Lin, C., Wu, J. J. K., Li, C. M. L., Li, J.-W., Yim, A. K. Y., Chan, S., Sibert, J., Džakula, Ž., Cao, H., Yiu, S.-M., Chan, T.-F., Yip, K. Y., Xiao, M., and Kwok, P.-Y. "Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays" *Genetics* 202, no. 1 (2016):

12. Russell, P. J. "IGenetics : A Molecular Approach" (2010):

13. Pierce, B. A. and Preceded by: Pierce, B. A. "Genetics : A Conceptual Approach"

14. Adams, M. D. "The Genome Sequence of Drosophila Melanogaster" *Science* 287, no. 5461 (2000): 2185–2195. doi:10.1126/science.287.5461.2185, Available at http://www.sciencemag.org/cgi/doi/10.1126/science.287.5461.2185

15. Beckingham, K. M., Armstrong, J. D., Texada, M. J., Munjaal, R., and Baker, D. A. "Drosophila Melanogaster--the Model Organism of Choice for the Complex Biology of Multi-Cellular Organisms." *Gravitational and space biology bulletin : publication of the American Society for Gravitational and Space Biology* 18, no. 2 (2005): 17–29. Available at http://www.ncbi.nlm.nih.gov/pubmed/16038090

16. Grenier, J. K., Arguello, J. R., Moreira, M. C., Gottipati, S., Mohammed, J., Hackett, S. R., Boughton, R., Greenberg, A. J., and Clark, A. G. "Global Diversity Lines–A Five-Continent Reference Panel of Sequenced Drosophila Melanogaster Strains" *G3: Genes, Genomes, Genetics* 5, no. 4 (2015):

17. Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley,

S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., and Begun, D. J. “Genomic Variation in Natural Populations of Drosophila Melanogaster” *Genetics* 192, no. 2 (2012):

18. Cardoso-Moreira, M., Arguello, J., and Clark, A. G. “Mutation Spectrum of Drosophila CNVs Revealed by Breakpoint Sequencing” *Genome Biology* 13, no. 12 (2012): R119. doi:10.1186/gb-2012-13-12-r119, Available at http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-12-r119

19. Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lyman, R. F., Magwire, M. M., Rollmann, S. M., Duncan, L. H., Lawrence, F., Anholt, R. R. H., and Mackay, T. F. C. “Systems Genetics of Complex Traits in Drosophila Melanogaster” *Nature Genetics* 41, no. 3 (2009): 299–307. doi:10.1038/ng.332, Available at http://www.nature.com/doifinder/10.1038/ng.332

20. Ometto, L., Glinka, S., Lorenzo, D. De, and Stephan, W. “Inferring the Effects of Demography and Selection on Drosophila Melanogaster Populations from a Chromosome-Wide Scan of DNA Variation” *Molecular Biology and Evolution* 22, no. 10 (2005): 2119–2130. doi:10.1093/molbev/msi207, Available at https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msi207

21. Corbett-Detig, R. B. and Hartl, D. L. “Population Genomics of Inversion Polymorphisms in Drosophila Melanogaster” *PLoS Genetics* 8, no. 12 (2012): e1003056. doi:10.1371/journal.pgen.1003056, Available at http://dx.plos.org/10.1371/journal.pgen.1003056

22. Alkan, C., Coe, B. P., and Eichler, E. E. “Genome Structural Variation Discovery and Genotyping” *Nature Reviews Genetics* 12, no. 5 (2011): 363–376. doi:10.1038/nrg2958, Available at http://www.nature.com/doifinder/10.1038/nrg2958

23. Tamura, K., Subramanian, S., and Kumar, S. “Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks” *Molecular Biology and Evolution* 21, no. 1 (2003): 36–

44. doi:10.1093/molbev/msg236, Available at https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msg236

24. Om, H. A. “DOBZHANSKY, BATESON, and the Genetics of Speciation” Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207686/pdf/ge14441331.pdf

25. Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O’Grady, P. M., Rohde, C., Valente, V. L. S., Aguadé, M., Anderson, W. W., Edwards, K., Garcia, A. C. L., Goodman, J., Hartigan, J., Kataoka, E., Lapoint, R. T., Lozovsky, E. R., Machado, C. A., Noor, M. A. F., Papaceit, M., Reed, L. K., Richards, S., Rieger, T. T., Russo, S. M., Sato, H., Segarra, C., Smith, D. R., Smith, T. F., Strelets, V., Tobari, Y. N., Tomimura, Y., Wasserman, M., Watts, T., Wilson, R., Yoshida, K., Markow, T. A., Gelbart, W. M., and Kaufman, T. C. “Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred from Genetic and Physical Maps.” *Genetics* 179, no. 3 (2008): 1601–55. doi:10.1534/genetics.107.086074, Available at http://www.ncbi.nlm.nih.gov/pubmed/8056307

26. Rieseberg, L. H. “Chromosomal Rearrangements and Speciation” *Trends in Ecology & Evolution* 16, no. 7 (2001): 351–358. doi:10.1016/S0169-5347(01)02187-5, Available at https://www.sciencedirect.com/science/article/pii/S0169534701021875

27. Noor, M. A., Grams, K. L., Bertucci, L. A., and Reiland, J. “Chromosomal Inversions and the Reproductive Isolation of Species.” *Proceedings of the National Academy of Sciences of the United States of America* 98, no. 21 (2001): 12084–8. doi:10.1073/pnas.221274498, Available at http://www.ncbi.nlm.nih.gov/pubmed/11593019

28. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., and Emerson, J. J. “Contiguous and Accurate *de Novo* Assembly of Metazoan Genomes with Modest Long Read Coverage” *Nucleic Acids Research* 44, no. 19 (2016): gkw654. doi:10.1093/nar/gkw654, Available at https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw654

29. Nattestad, M. and Schatz, M. C. “Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly” *Bioinformatics* 32, no. 19 (2016): 3021–3023. doi:10.1093/bioinformatics/btw369, Available at http://www.ncbi.nlm.nih.gov/pubmed/27318204

30. Quinlan, A. R. and Hall, I. M. “Tools: A Flexible Suite of Utilities for Comparing Genomic Features” *Bioinformatics* 26, no. 6 (2010): 841–842. doi:10.1093/bioinformatics/btq033, Available at https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033

31. Li, H. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences” *Bioinformatics* 32, no. 14 (2016): 2103–2110. doi:10.1093/bioinformatics/btw152, Available at https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw152

32. Suzuki, R. and Shimodaira, H. “Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering” *Bioinformatics* 22, no. 12 (2006): 1540–1542. doi:10.1093/bioinformatics/btl117, Available at https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl117

33. Chakraborty, M., VanKuren, N. W., Zhao, R., Zhang, X., Kalsow, S., and Emerson, J. J. “Hidden Genetic Variation Shapes the Structure of Functional Elements in Drosophila” *Nature Genetics* 50, no. 1 (2018): 20–25. doi:10.1038/s41588-017-0010-y, Available at http://www.nature.com/articles/s41588-017-0010-y

34. Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., Xu, Q., Wang, Z.-X., Wei, X., Han, B., and Huang, X. “Pan-Genome Analysis Highlights the Extent of Genomic Variation in Cultivated and Wild Rice” *Nature Genetics* 50, no. 2 (2018): 278–284.

doi:10.1038/s41588-018-0041-z, Available at http://www.nature.com/articles/s41588-018-0041-z

35. Feulner, P. G. D. and De-Kayne, R. “Genome Evolution, Structural Rearrangements and Speciation” *Journal of Evolutionary Biology* 30, no. 8 (2017): 1488–1490. doi:10.1111/jeb.13101, Available at http://doi.wiley.com/10.1111/jeb.13101

36. Cardoso-Moreira, M., Arguello, J. R., Gottipati, S., Harshman, L. G., Grenier, J. K., and Clark, A. G. “Evidence for the Fixation of Gene Duplications by Positive Selection in Drosophila.” *Genome research* 26, no. 6 (2016): 787–98. doi:10.1101/gr.199323.115, Available at http://www.ncbi.nlm.nih.gov/pubmed/27197209

37. Ting, C.-T., Tsaur, S.-C., Sun, S., Browne, W. E., Chen, Y.-C., Patel, N. H., and Wu, C.-I. “Gene Duplication and Speciation in Drosophila: Evidence from the Odysseus Locus” *Proceedings of the National Academy of Sciences* 101, no. 33 (2004): 12232–12235. doi:10.1073/PNAS.0401975101

38. Yeaman, S. and Whitlock, M. C. “THE GENETIC ARCHITECTURE OF ADAPTATION UNDER MIGRATION-SELECTION BALANCE” *Evolution* 65, no. 7 (2011): 1897–1911. doi:10.1111/j.1558-5646.2011.01269.x, Available at http://doi.wiley.com/10.1111/j.1558-5646.2011.01269.x

FIGURES

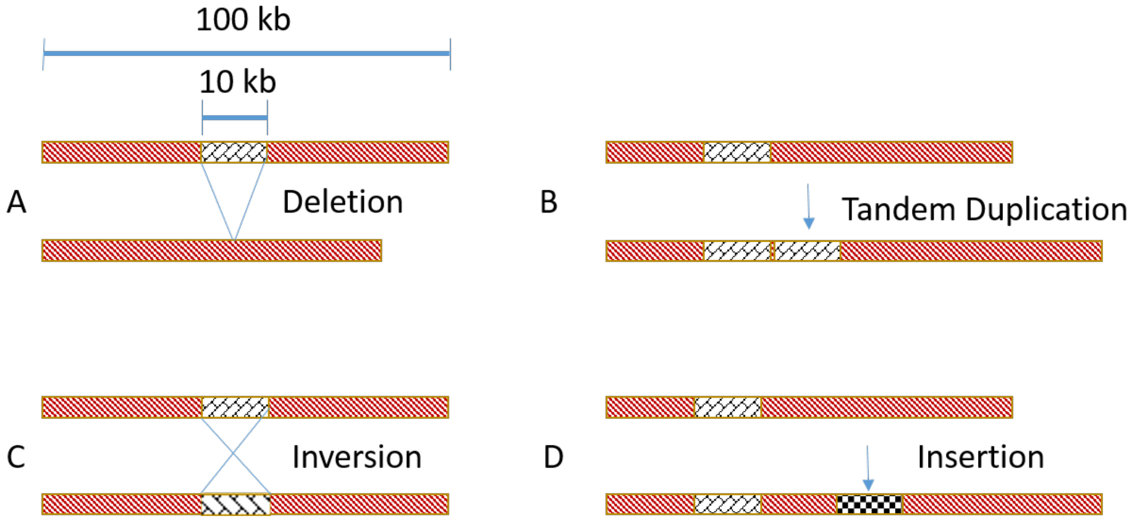

Figure 1-1. Types of Structural Variations. The red bars represent large regions of a chromosome with different patterned segments representing sites of variation. A) 10 kb deletion. B) Tandem duplication of one segment to an adjacent site. C) Inversion. D) Insertion of a new segment of DNA

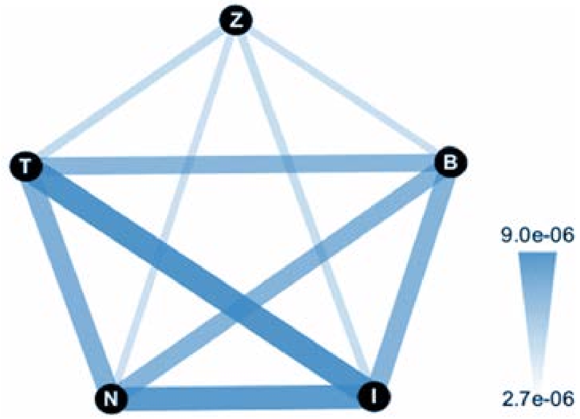

Figure 1-2. Population Distance network. Five populations of *D. melanogaster*: Beijing (B), Ithaca (I), Netherlands (N), Tasmania (T), and Zimbabwe (Z). Genetic similarity is measured by genome-wide FST with connecting bar width and color representing the amount of similarity.

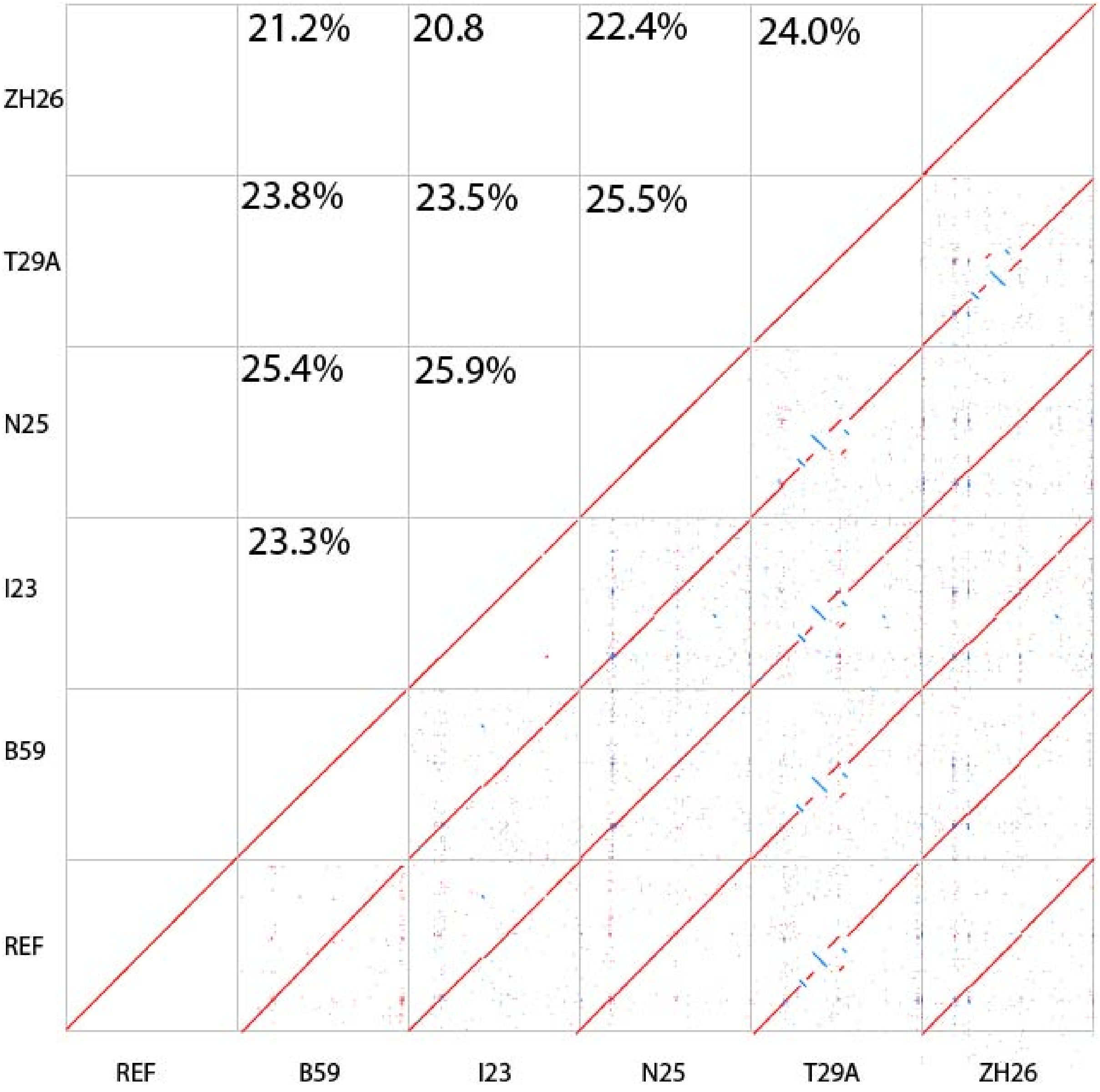

Figure 1-3. Whole genome alignments between the five global strains and the reference genome of *D. melanogaster*. The X-axis represents each genome used as the reference for alignment by the other strains. Alignment is shown in order chromosome arms 2L, 2R, 3L, 3R, 4, and X. Percentages represent amount of SV coincidence between each of the strains.

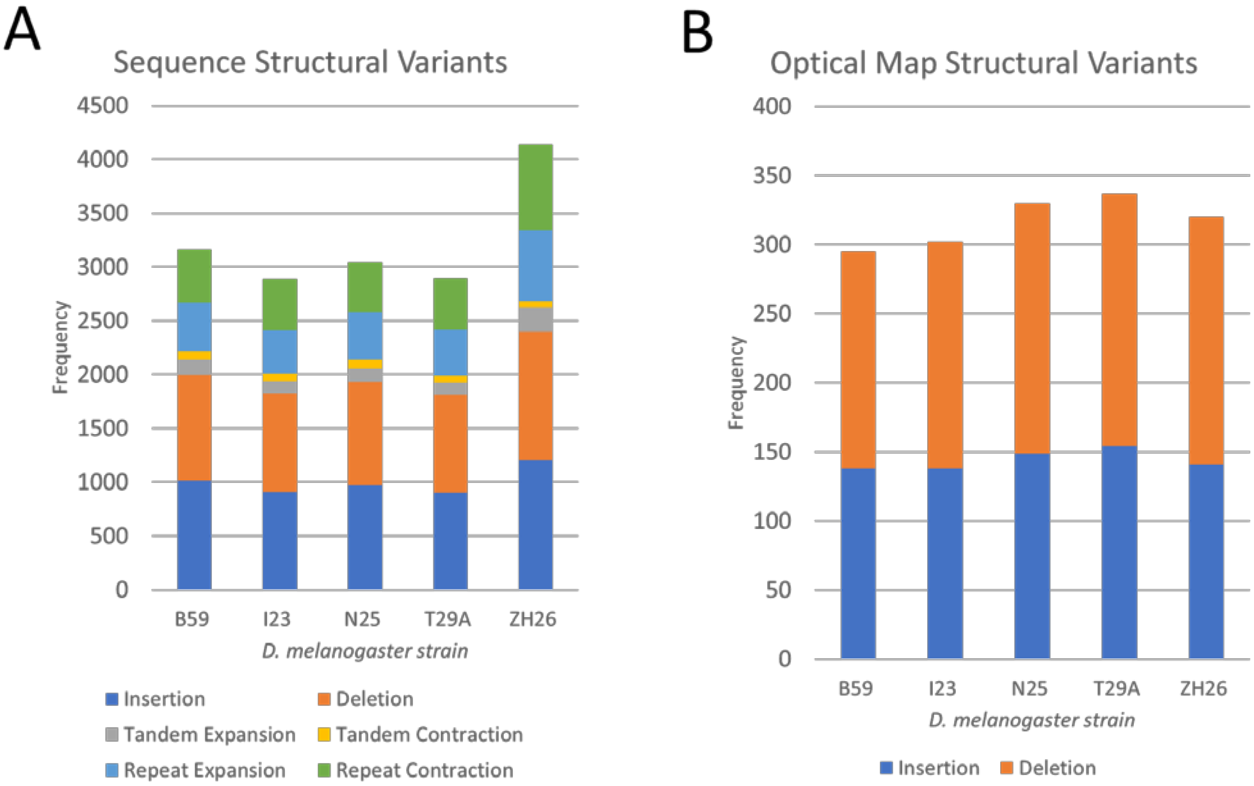

Figure 1-4. Structural Variant statistics of the five global strains of *D. melanogaster*. A) Classification and frequency of sequence based structural variants called by “Assemblyitics”. B) Classification and frequency of optical map based structural variations called by BioNano SVdetect”.

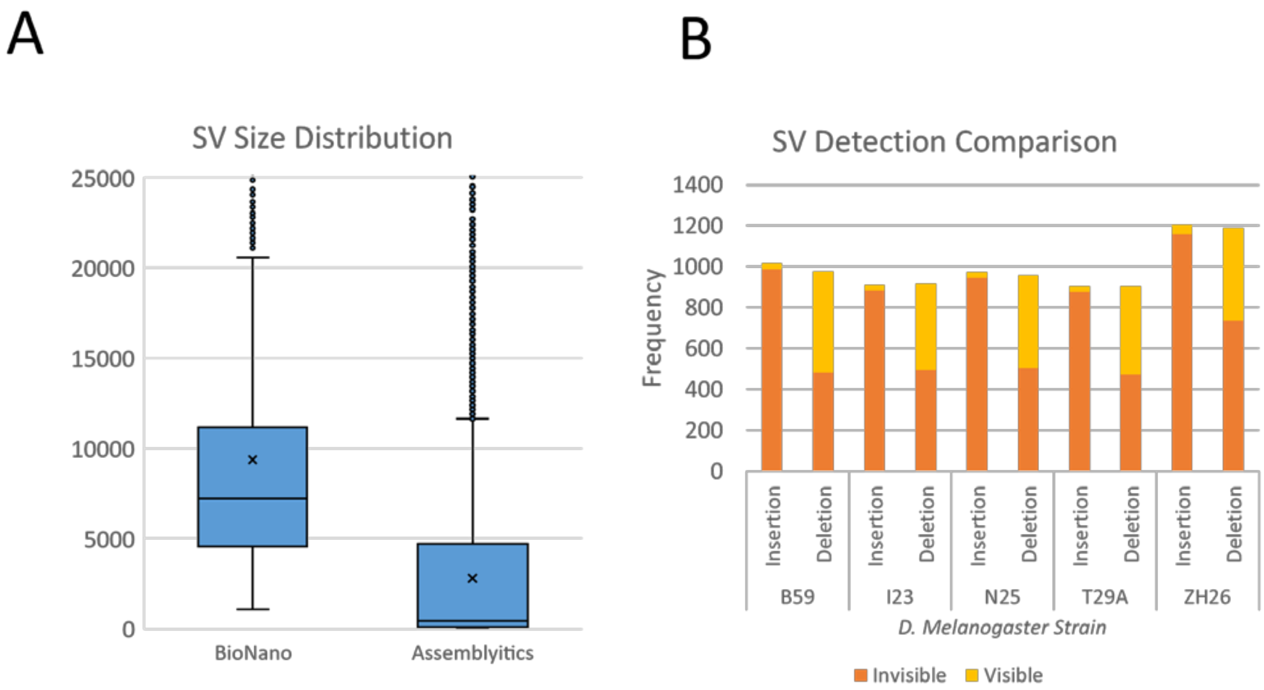

Figure 1-5. A) Size distribution of structural variants called by both long-read sequencing and Optical Map methods. The Y-axis is defined to show majority of variance and does not display some larger SVs detected. B) Sequence SVs detected by long-read sequencing, but not by short read resequencing are classified as "invisible".

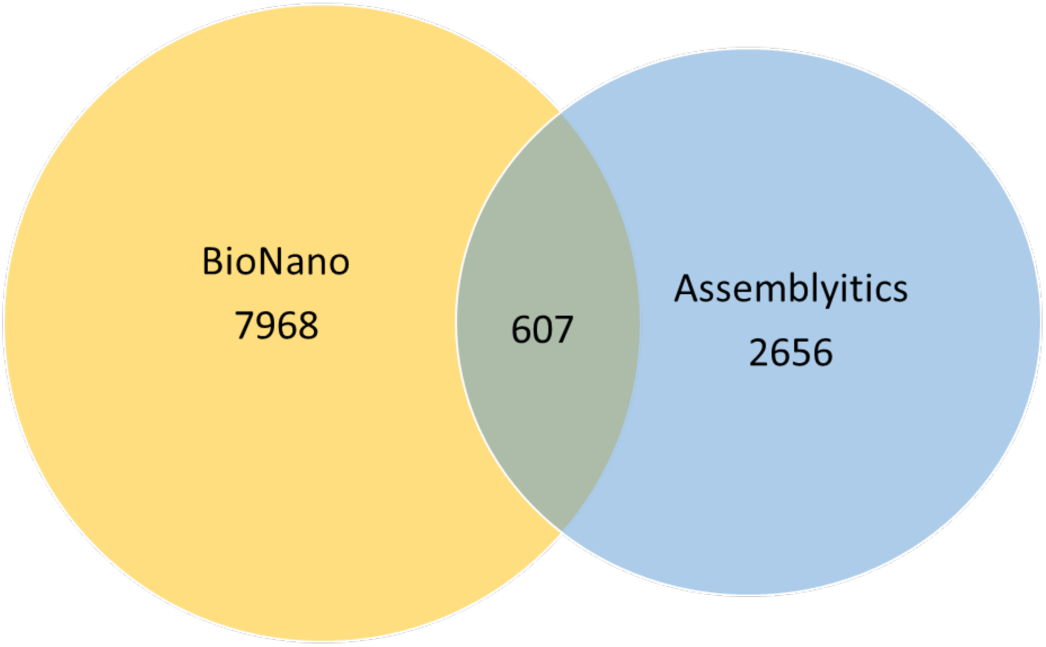

Figure 1-6. Venn Diagram displaying the average number of exons coinciding with SVs detected by the two SV detection methods. For conservative estimates, only exons coinciding with both methods were evaluated.

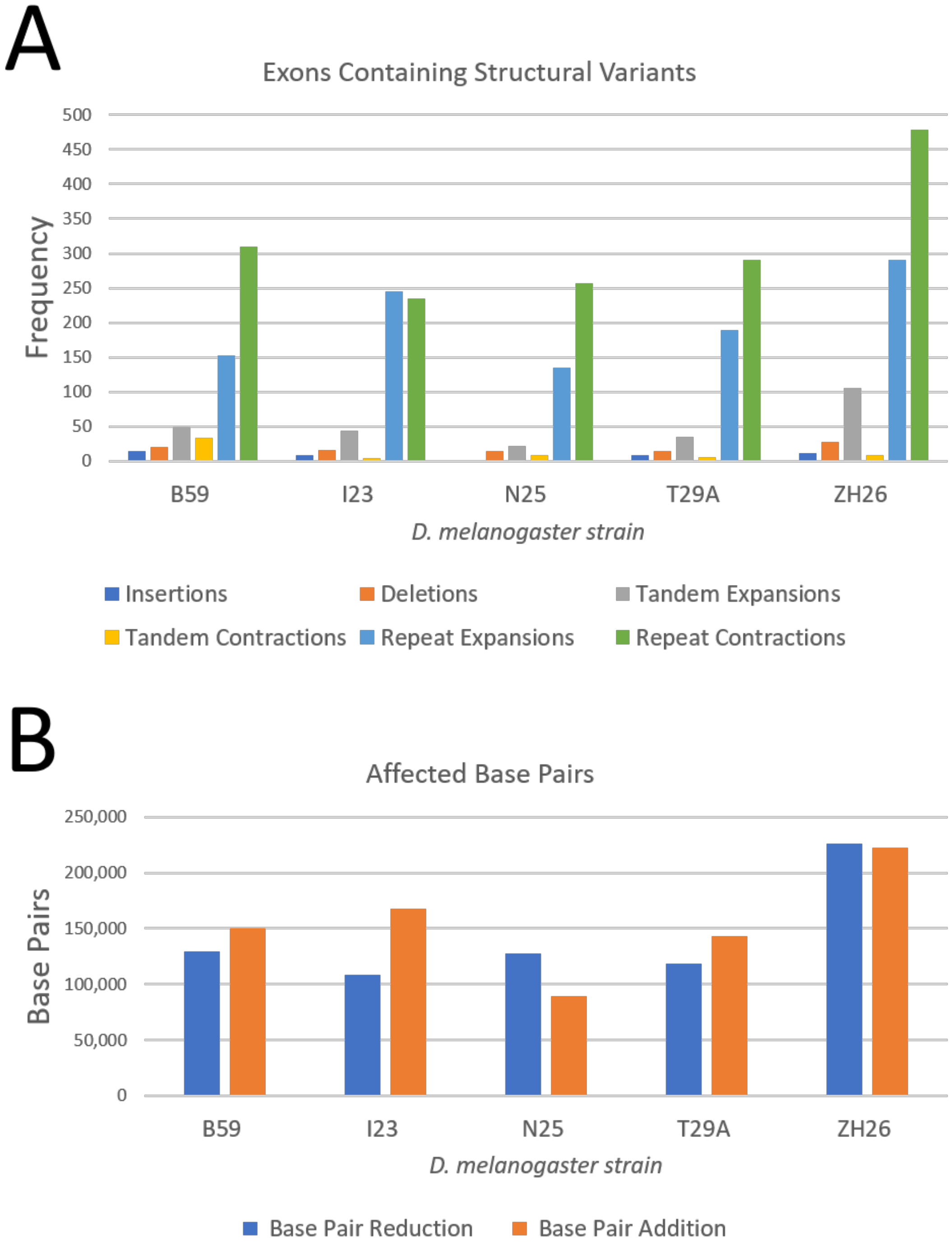

Figure 1-7. Exons containing Structural Variants in the five global strains of *D. melanogaster*. A) Classification and frequency of structural variants within exonic regions of the genome. B) Total amount of base pairs within exonic regions of the genome affected by structural variants. Insertions, repeat expansions, and tandem expansions were classified as base pair additions, while deletions, repeat contractions, and tandem contractions were classified as base pair reductions.

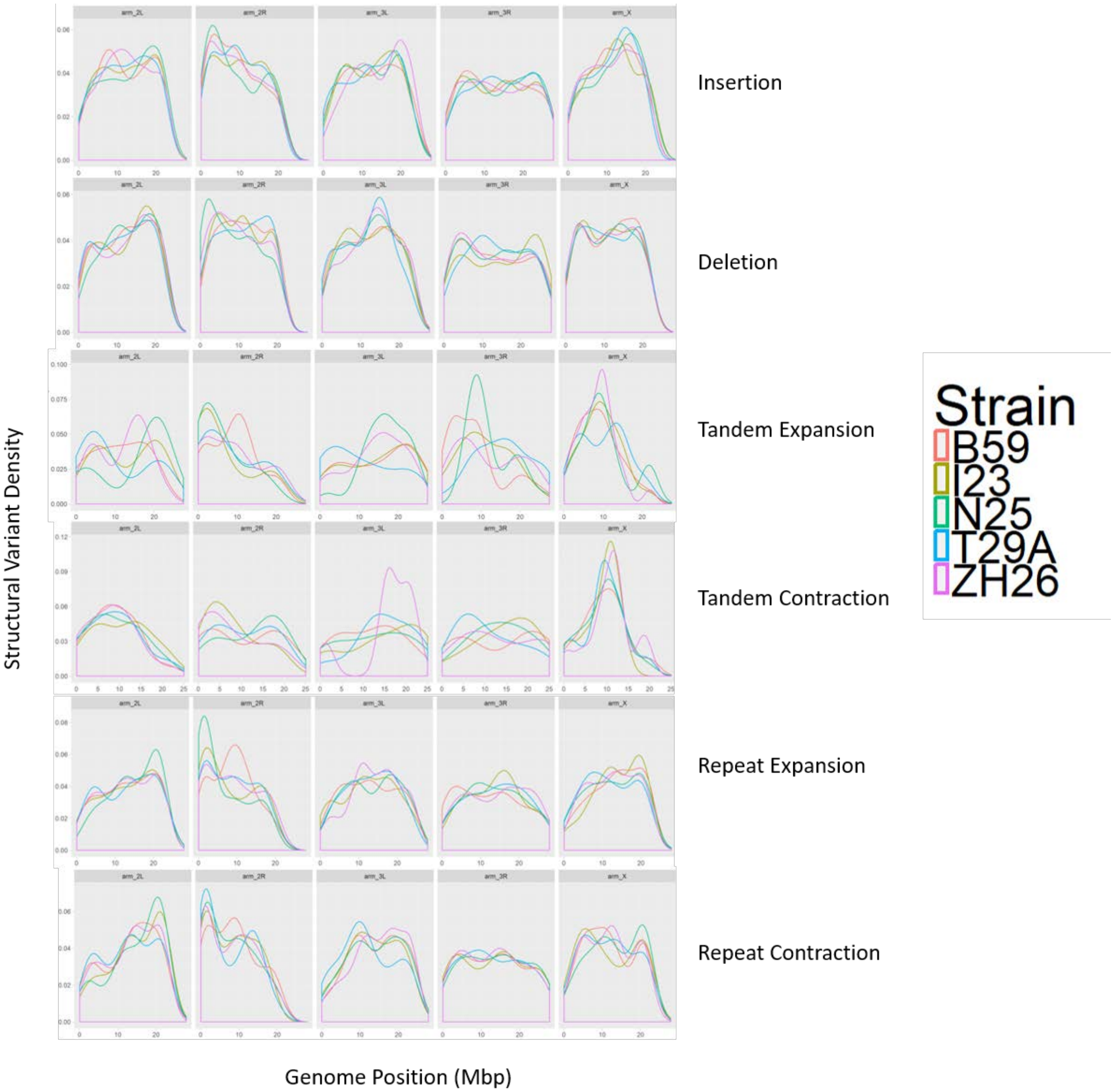

Figure 1-8. Structural Variant Distribution across each arm of the chromosome of *D. melanogaster*. Chromosome 4 was omitted due to its small size. The five global strains are indicated by line color.

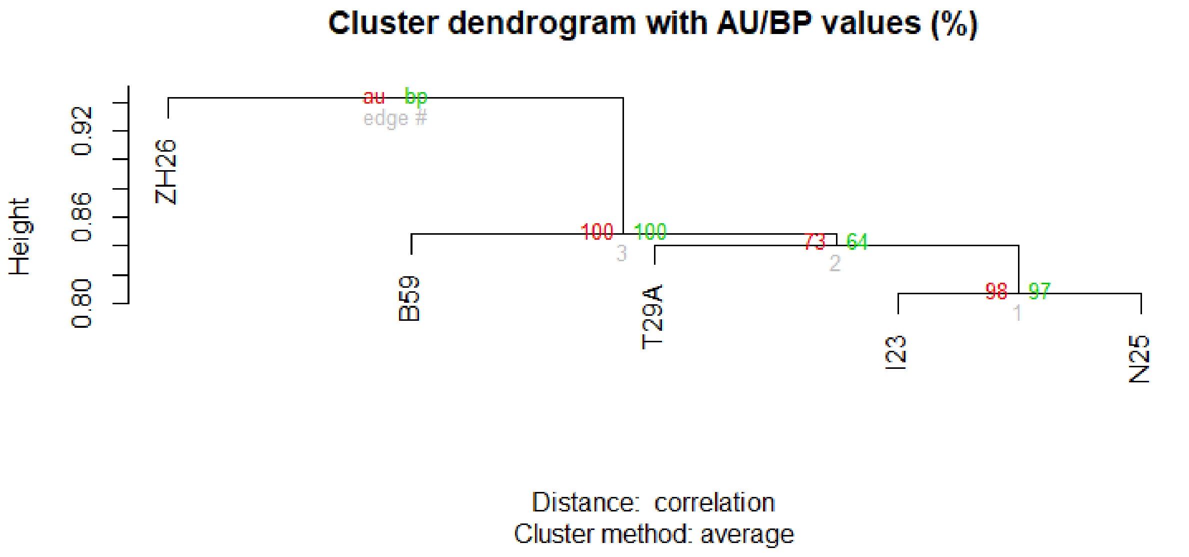

Figure 1-9. Evolutionary relationships based on coincidence of sequence structural variants between strains. The tree was created using pvclust package in R. Two types of p-values are shown for each branch node: Approximately Unbiased (AU) and Bootstrap Probability (BP).

TABLES

Table 1-1. Global Strains of *D. melanogaster* collected from their respective locations to create a representative panel.

***D. melanogaster* Lines**	**Locations**
B59	Beijing (China)
I23	Ithaca (New York)
N25	Netherlands
T29A	Tasmania
ZH26	Zimbabwe

Table 1-2. Assembly statistics for sequence and optical map assemblies of the five global strains of *D. melanogaster*. Optical map alignment is evaluated against the assembled portion of the *D. melanogaster* ISO1 release 5 reference genome.

	B59	**I23**	**N25**	**T29A**	**ZH26**
Sequence Assembly Length (Mb)	144.4	132.95	146.07	139.06	177.22
# Of Contigs	283	314	306	205	960
Contig N50 (Mb)	5.97	1.08	6.47	11.37	1.24
BUSCO %	95.7	85.3	96.9	96.3	92.0
Optical Map Length (Mb)	138.2	130.1	166.3	148.5	144.1
Optical Map N50 (Mb)	1.01	0.897	1.255	1.144	1.297
Alignment %	93.9	91.4	90.8	94	92.8
Hybrid Scaffold Length (Mb)	144.97	135.92	148.16	139.75	179.28
# Of Scaffolds	259	218	276	184	918
Scaffold N50 (Mb)	10.24	5.63	15.39	21.47	2.55

SUPPLEMENTAL METHODS

Supplemental Methods 1) Commands and Parameters

Creating consensus structural variants called by both Assemblyitics and BioNano:

```
Bedtools/intersectBed –a Assemblyitics_SV.bed –b BioNano_SV.bed > Consensus_SV.bed
```

Creating list of Exons coinciding with SVs:

```
Bedtools/intersectBed –a Dmel_Exons.bed –b Consensus_SV.bed > Exonic_SV.bed
```

Creating whole genome alignment visualization:

```
minimap2 -cS Reference.fasta Query.fasta > Alignment.paf

miniasm/minidot  Alignment.paf > Alignment.eps
```

Calculating the number of bases within Exons affects by SVs:

```
awk '{print $5-$4}' Exonic_SV > number_bases_exon_repeat_exp

 paste -sd+ number_bases_exon_repeat_exp | bc
```

*The Only exception is insertions where we used the size column of Assembylitics to find out how much was inserted in the Exon

Canu Assembly parameters:

```
canu -d Directory_name -p Directory_name genomeSize=180m maxMemory=200g maxThreads=10 corMhapSensitivity=normal corOutCoverage=40 merylMemory=100g merylThreads=10 ovsMethod=parallel gridOptions="--qos=jaudall --time=24:00:00" gridOptionsOVS="--mem-per-cpu=10g --time=24:00:00" gridOptionsExecutive="--mem-per-cpu=24g --time=4:00:00" gridOptionsCORMHAP="--mem-per-cpu=10g --time=1:00:00" gridOptionsOBTMHAP="--mem-per-cpu=10g --time=1:00:00" gridOptionsUTGMHAP="--mem-per-cpu=10g --time=1:00:00" gridOptionsCOROVL="--mem-per-cpu=6g --time=24:00:00" gridOptionsOBTOVL="--mem-per-cpu=6g --time=12:00:00" gridOptionsUTGOVL="--mem-per-cpu=6g --time=12:00:00" gridOptionsRED="--mem-per-cpu=8g --time=1:00:00" gridOptionsOEA="--mem-per-cpu=8g --time=2:00:00" gridOptionsOVB="--mem-per-cpu=4g --time=1:00:00" -pacbio-raw reads.fastq
```

BioNano Assembly parameters:

Minimum length= 125 kb

P-value cutoff threshold, initial assembly = 5.68e-9

P-value cutoff threshold, extension and refinement= 5.68e-10

Evaluating shared SVs using R paired with “pvclust” package:

```
#Load in Assemblyitics structural variant files
Bedone <- read.table("strainone_assemblyitics.txt")
Bedtwo <- read.table("straintwo_assemblyitics.txt")

#create 0 vector for each SV in Bedone
G1 <- cbind(rep(0, length(Bedone)))

#For every entry in the first Bedfile, this loops through the second bedfile to evaluate wether the SV has an identical match in the other strain.  There is a margin for error of 3 basepairs.
for(x in 1:length(Bedone)){
  for(w in 1:length(Bedtwo)){
    if(smapone[x,4] > Bedone1[w,4] - 3 & smapone[x,4] < Bedone1[w,4] + 3 & smapone[x,5] > Bedone1[w,5] - 3 & smapone[x,5] < Bedone1[w,5] + 3 & smapone[x,2] == Bedone1[w,2] & smapone[x,6] == Bedone1[w,6])
# If the SV in Bedone has a match in Bedtwo then the 0 in out vector is replaced with a 1
    {G1[x] <- 1
    #This ends the loop once we've found a match
    w <- length(Bedtwo) -1}
  }
}
#Once this is performed pairwise for each strain, we built a matrix with 1 columns for each strain, showing its binary coincidence with the other strains SVs
library(pvclust)
#Creates pvclust tree for the SV coincidence matrix
result <- pvclust(SV_coincidence_matrix, method.dist = "cor", method.hclust = "average", nboot=1000)
plot(result)
```