Whole genome phylogeny for 21 *Drosophila* species using predicted 2b-RAD fragments

Type IIB restriction endonucleases are site-specific endonucleases that cut both strands of double-stranded DNA upstream and downstream of their recognition sequences. These restriction enzymes have recognition sequences that are generally interrupted and range from 5-7 bases long. They produce DNA fragments which are uniformly small, ranging from 21-33 base pairs in length (without cohesive ends). The fragments are generated from throughout the entire length of a genomic DNA providing an excellent fractional representation of the genome. In this study we simulated restriction enzyme digestions on 21 sequenced genomes of various *Drosophila* species using the predicted targets of 16 Type IIB restriction enzymes to effectively produce a large and arbitrary selection of loci from these genomes. The fragments were then used to compare organisms and to calculate the distance between genomes in pair-wise combination by counting the number of shared fragments between the two genomes. Phylogenetic trees were then generated for each enzyme using this distance measure and the consensus was calculated. The consensus tree obtained agrees well with the currently accepted tree for the *Drosophila* species. We conclude that multi-locus sub-genomic representation combined with next generation sequencing, especially for individuals and species without previous genome characterization, can accelerate studies of comparative genomics and the building of accurate phylogenetic trees.
Arun S Seetharam\textsuperscript{1,2}

Gary W. Stuart\textsuperscript{1}

\textsuperscript{1}Department of Biology, Indiana State University, 600 Chestnut Street, Terre Haute, IN 47809, USA.

\textsuperscript{2}Present address: Bioinformatics Core, Purdue University, West Lafayette, IN 47906, USA.

Corresponding Author:

Gary W. Stuart

Department of Biology, Indiana State University, 600 Chestnut Street, Terre Haute, IN 47809, USA

Tel: +1 (812) 237-7898

Email: gstuart@indstate.edu
1 Introduction

Evolutionary relationships of species derived by comparing single orthologous genes or groups of genes can be negatively affected by potential horizontal gene transfers, incomplete lineage-sorting, introgression, and the unrecognized comparison of paralogous genes (Delsuc et al. 2005). However, with the advent of the genomic era, it is now possible for researchers to use the complete genomes of fully sequenced organisms for building trees. Though such trees offer robustness for analysis, it becomes impractical to use traditional methods for constructing large scale alignments and for generating trees from these alignments, mainly because of their large size and their highly heterogeneous nature. As a result, there are now sophisticated methods that don’t rely on alignment and are optimized for large scale data. These methods generally use vector representation of genes (Qi et al. 2004; Stuart et al. 2002) or features such as gene content (Huson & Steel 2004; Snel et al. 1999; Tekaia et al. 1999), gene order (Bourque & Pevzner 2002; Korbel et al. 2002), intron positions (Roy & Gilbert 2005), or protein domain structure (Lin & Gerstein 2000; Yang et al. 2005).

Despite a strong recent interest in the various large-scale non-alignment methods, they are often viewed as somewhat less rigorous and less reliable. In addition, even with the dramatic decrease in the cost of genome sequencing, it is still not attractive to sequence the genomes of those organisms that have little economical value, especially if their genomes are extremely large. On the other hand, the possibility of obtaining a large and representative set of fragments, instead of the whole genome sequence, can be economically feasible even for the lesser known species and can provide a valuable alternative for many types of genomic scale studies, including phylogenomics.

Recently, several approaches have been developed to represent the genome by randomly sampling the entire genome. These approaches give a good reduced representation of the genome and are based on restriction sites on the genome combined with the next generation sequencing methods. Some popular methods include Complexity Reduction of Polymorphic Sequences (CRoPS) (van Orsouw et al. 2007); restriction site-associated DNA sequencing (RAD-seq) (Baird et al. 2008a; Etter et al. 2011); Genotyping by Sequencing method (GBS); double-digest RAD-seq (Peterson et al. 2012), and 2bRAD (Wang et al. 2012). All these methods provides good sub samples from homologous locations within genomes and are widely used to study population genetics (Baxter et al. 2011; Hohenlohe et al. 2010). These methods have the potential to uncover detailed information about a wealth of genomic markers. Complex interactions among markers
can also be extracted at the population level (Baird et al. 2008b; Davey & Blaxter 2010).

Recently, these fragments have also been used for evolutionary studies (Emerson et al. 2010; Rubin et al. 2012; Yi & Jin 2013).

A novel class of enzymes, known as Type IIB restriction endonucleases (Roberts et al. 2003b), are site-specific endonucleases that cut both strands of double-stranded DNA upstream and downstream of their recognition sequences. These restriction enzymes have recognition sequences that are generally interrupted and range from 5-7 bases long. They produce DNA fragments which are of uniform length, ranging from 21-33 base pairs in length (without cohesive ends) (Roberts et al. 2003a). The fragments are generated from throughout the entire length of a genomic DNA providing an excellent fractional representation of the genome. This method of generating fragments using Type IIB enzymes is termed 2bRAD (Wang et al. 2012) and these fragments have been used for various purposes including population studies, digital karyotyping (Stebbins 1950), for pathogen identification by computational subtraction (Tengs et al. 2004) and genomic profiling to identify and quantitatively analyze genomic DNAs (Dunn et al. 2002). In this study, we show that these fragments can be used for efficient phylogenetic study for determining evolutionary relationships between distinct species. We have tested this method in silico and shown that 13 different types of IIB restriction enzymes can be used to accurately reconstruct the phylogeny of a diverse set of 21 Drosophila species that are currently available.

2 Materials and Methods

2.1 Obtaining datasets

Whole genome, nucleotide sequences for the 21 Drosophila species were downloaded from the FlyBase (McQuilton et al. 2012), NCBI databases and from Princeton University website (Rebeiz et al. 2009) on July 10, 2010.

2.2 Simulated restriction digestion

The PERL program “Phyper” was used to simulate restriction digestion for all 16 Type IIB endonuclease enzymes and for processing the obtained fragments. This program generated a representative list of unique fragments i.e., single-copy fragments (most abundant) and fragments that are present as multiple identical copies (less frequent). The former fragments are most likely to belong to divergent fragment families, within a given genome that display one or a few mutations relative to each other and were identified and removed from the analysis. The representative list of fragments were generated for each genome, for each enzyme separately.
2.3 Fragment comparisons

The representative lists of fragments were then used with another PERL program “Phyppa” for comparative analyses. This program compares each fragment of a genome with every fragment of another genome in order to find identical fragments and similar fragments (fragments with up to 6 mismatches for ensuring more than 70 % similarity among sequences). A total of 210 such comparisons were done in order to generate the full list of shared fragments (identical fragments and similar fragments) for every pair of genomes (both PERL scripts are available upon request). Analyses was performed on a standard laptop with a quad core processor (1.73 GHz Intel Core i7) and with 6 GB RAM. For each enzyme, the scripts required about 6 hours to finish for both fragment generation and comparison between all genomes.

2.4 Distance calculations

The number of shared fragments between a pair of genomes was then used to calculate the evolutionary distance by calculating the ratio of shared fragment to the total fragments and converting them to negative natural log (Equation 1). Conversion to negative natural log was essential to ensure that the distances computed were always positive.

Equation

\[
\text{Distance} = -\ln \left( \frac{\text{Identical fragments} + \text{Similar fragments}}{\text{Total fragments of both species}} \right)
\]

2.5 Building trees

Distance measures for all the pairwise comparisons for a particular enzyme were used to build trees using the neighbor program from the Phylip (Felsenstein 2005) package. A consensus tree was then produced by combining trees for all the enzymes with the consensus program from Phylip. The flowchart for the entire process is given in Figure. 1

3 Results and Discussion

3.1 Datasets

The full nucleotide sequences for 21 Drosophila species downloaded from various sources are listed in Table 1. The genome size ranged from 137.82 mb for D. simulans to 235.52 mb for D. willistoni. D. willistoni had the lowest GC content of all with 37.89% and D. pseudoobscura had the highest GC content (45.43%).
3.2 Type IIB restriction enzymes

The 16 Type IIB restriction endonucleases that could be used for simulating the restriction digestion of *Drosophila* genomes along with their recognition sites, average distance between the restriction sites assuming random distribution of nucleotides and without any compositional bias, and the size of fragment (blunt) that the enzymes leaves behind are given in Table 2 ([Tengs et al. 2004](http://example.com)). Unlike traditional Type II enzymes, Type IIB enzymes cleave on both sides of the recognition sequence (about 7-15 bases upstream and downstream, depending on enzyme) generating a fragment of uniform length. Also, the recognition site is usually split into two parts by some fixed number of random bases. They normally leave 2-3 base overhangs on the generated fragment.

3.3 Fragment analyses

The numbers of representative fragments obtained from each genome for each enzyme are listed in Table 3. The most frequent cutting enzymes such as *BslFI* had generally higher numbers of fragments within all genomes compared to other enzymes. Also, *D. pseudoobscura* and *D. persimilis* had relatively higher numbers of fragments compared to other genomes with most of the enzymes. Following fragment extraction, the original genomic sequences downloaded from various source databases were represented as a collection of fragments of uniform length. For each genome a total of 16 fragment sets were generated by using 16 different type IIB enzymes. The number of fragments generated by each genome was not closely related to the size of their genomes but they were related to the GC content. Most of the enzymes used in the analysis recognized a GC rich recognition site which is reflected in the number of fragments generated with GC rich genomes. The genomes that were GC rich such as *D. pseudoobscura* and *D. persimilis* had higher numbers of fragments compared to other genomes. Similarly the genomes that had lower GC content such as *D. willistoni* and *D. grimshawi* generated fewer fragments. Overall, the number of fragments obtained for each species were within the range of expected fragments based on their genome size and estimated distance between restriction cut sites (assuming random sequence without GC content bias). Most enzymes predicted to be frequent cutters generated large number of fragments like *BslFI*. Predicted rare cutters like *PsrI*, *PpiI*, *AloI* and *CspCI* generated fewer fragments than other enzymes.

3.4 Distance Matrices and Phylogenetic Trees

A comparison of fragments between genomes provided a list of fragments that were shared by those genomes. Closely related organisms are expected to share higher numbers of similar
fragments (including identical fragments) compared to other distantly related genomes. Similar fragments are defined as those with 6 or fewer mismatches. Since the average length of fragments generated from various enzymes was around 27 bases, allowing 6 bases mis-match ensured at least 75% similarity among the sequences. The fragments being compared between 2 genomes ranged from 21 bp to 33 bp long (average size of 27 bp). The identical fragments between the 2 genomes are most likely to represent homologous or even orthologous sections of the genomes. Even for a fragment length of 21 bp (smallest fragment size produced by these enzymes), the probability that a particular 21 bp sequence exists one or more times in a genome of 150 Mb is 0.00341%. The pair-wise distance matrices constructed using the similar fragments detected by each enzyme were used to estimate phylogenetic trees (Figure 2). The individual NJ trees obtained for each enzyme were largely consistent with the currently accepted relationships among the various Drosophila groups and subgroups, as was the single consensus tree obtained (Figure 3). Per cent support values were calculated based on number of enzymes supporting the particular branch.

4 Conclusions

The 21 species of Drosophila used here included the subgenus Sophophora and the subgenus Drosophila. The Sophophora group was represented by melanogaster, obscura and willistoni and the Drosophila group was represented by virilis, repleta and mojavensis. Out of the 12 subgroups within the melanogaster group, 9 subgroups viz., ananassae, montium, melanogaster, suzukii, takahashii, ficsphila, elegans, rhopaloa and eugracilis were represented by 15 species. Of these, only 2 subgroups had multiple members within our data set, but both displayed a monophyletic arrangement within the final tree shown in Figure 2. The placement of the 12 well-studied Drosophila species viz., D. simulans, D. sechellia, D. melanaogaster, D. erecta, D. ananassae, D. yakuba, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis and D. grimshawi within our tree corresponds exactly to the currently accepted phylogeny (Clark et al. 2007; Hahn et al. 2007; Haubold & Pfaffelhuber 2012; Stark et al. 2007).

Overall, the topology of our 21 species tree agrees precisely with those presented by Van der Linde and Houle, (van der Linde et al. 2010), Haubold and Pfaffelhuber (Haubold & Pfaffelhuber 2012) and Yang et al. 2012 (Yang et al. 2012) and all the branches were completely resolved. The subgenus Sophophora was clearly distinguished into old world clades melanogaster/obscura and neo world clade willistoni in our tree (van der Linde & Houle 2008). The largest group melanogaster, had multiple subgroups viz., melanogaster, montium, ananassae
and oriental subgroup cluster \((eugracilis, suzukii, takahashii, elegans, rhopaloa, figusphila)\).

Many previous studies have failed to completely resolve the nodes within the oriental subgroup cluster \((Da Lage et al. 2007; Toda 1991)\). In our tree, \(ananassae\) group formed the earliest branch in the \(melanogaster\) group followed by \(montium\) subgroup with strong branch support values.

Most of the earlier studies confirmed this topology \((Da Lage et al. 2007; Kopp 2006; Prud'homme et al. 2006)\) except for two studies that placed them together as a sister clade from rest of the subgroups \((Schawaroch 2002)\) or reversed the order of branching \((Yang et al. 2004)\).

Both these studies had poor branch support. The oriental subgroups cluster formed three sub clades. The first sub-clade included \(elegans\) and \(rhopaloa\) with \(figusphila\) as the sister sub-group, the second sub-clade included \(suzukii\) and \(takahashii\) and the third sub-clade included the \(eugracilis\) sub-group. The placements of these sub-clades were controversial among the literature surveyed and was attributed to the explosive radiation of these oriental groups \((van der Linde & Houle 2008)\). The \(eugracilis\) clade consisting of \(D. eugracilis\) is most inconsistently placed clade and it is either placed as sister species of \(melanogaster\) sub group, as in our tree \((Haubold & Pfaffelhuber 2012; Pelandakis & Solignac 1993; van der Linde et al. 2010)\) or as sister species of the sub clade formed by \(suzukii\) and \(takahashii\) \((Yang et al. 2004)\) or as sister species of \(elegans\) and \(rhopaloa\) within the \(elegans\) - \(rhopaloa\) - \(figusphila\) clade \((Yang et al. 2012)\). The placements of the other two clades, \(suzukii\) - \(takahashii\) and \(elegans\) - \(rhopaloa\) - \(figusphila\) within the \(melanogaster\) group in our tree is in agreement with other published studies \((Kopp 2006; Kopp & True 2002)\). The sub-clade formed by \(suzukii\) and \(takahashii\) is well supported by most studies including ours with the strong branch support \((Da Lage et al. 2007; Kopp & True 2002; Schawaroch 2002; Yang et al. 2004)\). Most studies have confirmed that the \(rhopaloa\) subgroup is the sister group of the \(elegans\) subgroup but the \(figusphila\) sub group is considered to be polytomic branching clade in the \(melanogaster\) group \((van der Linde & Houle 2008)\). However, in our tree \(figusphila\) sub group is presented as the sister species of \(rhopaloa\) - \(elegans\) subgroups, albeit with low branch support. Within the \(Drosophila\) subgenus, all three groups \((virilis, repleta\) and \(grimshawi)\) exhibited a topology frequently observed in other studies \((van der Linde & Houle 2008)\).

A variety of sub-genomic sampling methods have been used previously for population studies and are especially effective on non-model organisms, but are rarely used for generating phylogenies for a diverse set of distinct species. We show here that multi-locus data obtained from short sub-genomic fragment sets, essentially 2b-RAD, provides good phylogenetic signal
and produces a well resolved and well-supported species phylogeny. The wide adoption of various RAD-like methods is due to the fact that deep sequencing of the fragments produced can be easily accomplished following two simple steps: adapter ligation, and then PCR. These methods are applicable to any organism irrespective of its genome size. The 2b-RAD approach to fragment generation and characterization in particular is simple, quick and cost effective (Wang et al. 2012). This method also shares some similarity with the recently described, alignment free multi-locus "co-phylog" method (Yi & Jin 2013). Both use a large number of short homologous fragments and, consequently, both can be profitably applied to short sequence reads derived via next generation sequencing, even prior to assembly. However, the co-phylog method is distinct in that it makes use of standard alignment algorithms applied to each locus to generate estimates of relatedness for building phylogenies. Effective application of the co-phylog method generally requires that the genomes being compared be closely related, and this would be expected to be true for our method as well, since effective matching of homologous short fragments in either case requires a significant degree of local sequence similarity. Despite this expected limitation, we note that the Drosophila species compared herein are relatively diverse, spanning approximately 40-50 million years of evolution.
5 References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, and Johnson EA. 2008a. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One* 3.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, and Johnson EA. 2008b. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, and Blaxter ML. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* 6:e19315.

Bourque G, and Pevzner PA. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res* 12:26-36.

Clark A, Eisen M, Smith D, Bergman C, Oliver B, Markow T, Kaufman T, Kellis M, Gelbart W, Iyer V, Pollard D, Sacak T, Larracuent A, Singh N, Abad J, Abt D, Adryan B, Aguade M, Akashi H, Anderson W, Aquadro C, Ardell D, Arguello R, Artieri C, Babash D, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak S, Bradley R, Brand A, Brent M, Brooks A, Brown R, Butlin R, Caggese C, Calvi B, Bernardo de Carvalho A, Caspi A, Castrezzana S, Celniker S, Chang J, Chapple C, Chitteri J, Chinwalla A, Civetta A, Clifton S, Comeron J, Costello J, Coyne J, Daub J, David R, Delcher A, Delehaunty K, Do C, Ebling H, Edwards K, Eickbusch T, Evans J, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia A, Gardiner A, Garfield D, Gavis B, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg A, Griffiths-Jones S, Gross S, Guigo R, Gustafson E, Haerty W, Hahn M, Halligan D, Halpern A, Halter G, Han M, Heger A, Hillier L, Hinrichs A, Holmes I, Hoskins R, Hubisz M, Hultmark D, Huntley M, Jaffe D, Jagadeeshan S, Jeck W, Johnson J, Jones C, Jordan W, Karpen G, Kataoka E, Keightley P, Kheradpour P, Kirkness E, Koerich L, Kristiansen K, Krudma D, Kulathinal R, Kumar S, Kwok R, Lander E, Langley C, Lapoint R, Lazzaro B, Lee S, Levesque L, Li R, Lin C, Lin M, Lindblad-Toh K, Llopis A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado C, Makalowsk W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride C, McKernan B, McKernan K, Mendez-Lago M, Minx P, Mollemhauer M, Montooth K, Mount S, Mu X, Myers E, Negre B, Newfeld S, Nielsen R, Noor M, O'Grady P, Pachter L, Papacite M, Parisi M, Parisi M, Parts L, Pedersen J, Pesole G, Phillipps A, Ponting C, Pop M, Porcelli D, Powell J, Prohaska S, Pruitt K, Puig M, Quesnelle H, Ram K, Rand D, Rasmussen M, Reed L, Reenan R, Reily A, Remington K, Rieger T, Ritchie M, Robin C, Rogers Y, Rohde C, Rozas J, Rubenfield M, Ruiz A, Russo S, Salzberg S, Sanchez-Gracia A, Saranga D, Sato H, Schaeffer S, Schatz M, Schlenke T, Schwartz R, Segarra C, Singh R, Sirot L, Sirota M, Sisneros N, Smith C, Smith T, Spiehi J, Stage D, Stark A, Stephan W, Strausberg R, Strempl S, Sturigil S, Sutton G, Suton G, Tao W, Teichmann S, Tobiari Y, Tomimura Y, Tsolas J, Valente V, Venter E, Venter J, Vicario S, Vieira F, Villeneuve A, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson R, Wing R, Wolner M, Wong A, Wong G, Wu C, Wu G, Yamamoto D, Yang H, Yang S, Yorke J, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin A, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer A, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukgalher B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M,
Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley C, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin C, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hagos B, Hafez N, Henson C, Honan T, Huard M, Hughes L, Hurhula B, Husby M, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lace A, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai A, Mihalev A, Mihova T, Mittelmann R, Molenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, O'Neill B, Osman S, Markiewicz E, Oyono O, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Yu Q, Zembek L, Zhong D, Zimmer A, Alvarez P, Brockman W, Butler J, Chin C, Grabherr M, Kleber M, Mauceli E, and MacCallum I. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203 - 218.

Da Lage JL, Kergoat GJ, Maczkowiak F, Silvain JF, Cariou ML, and Lachaise D. 2007. A phylogeny of Drosophilidae using the Amyrel gene: questioning the Drosophila melanogaster species group boundaries. Journal of Zoological Systematics and Evolutionary Research 45:47-63.

Davey JW, and Blaxter ML. 2010. RADSeq: next-generation population genetics. Brief Funct Genomics 9:416-423.

Delsuc F, Brinkmann H, and Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-375.

Dunn JJ, McCorkle SR, Praissman LA, Hind G, Van Der Lelie D, Bahou WF, Gnatenko DV, and Krause MK. 2002. Genomic signature tags (GSTs): a system for profiling genomic DNA. Genome Res 12:1756-1765.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, and Holzapfel CM. 2010. Resolving postglacial phyleogeography using high-throughput sequencing. Proc Natl Acad Sci USA 107:16196-16200.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, and Cresko WA. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol 772:157-178.

Felsenstein J. 2005. PHYLIPT (Phylogeny Inference Package) In: author Db, editor. 3.6. ed.

Hahn MW, Han MV, and Han SG. 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet 3:e197.

Haubold B, and Pfaffelhuber P. 2012. Alignment-free population genomics: an efficient estimator of sequence diversity. G3 (Bethesda) 2:883-889.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, and Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6:e1000862.

Huson DH, and Steel M. 2004. Phylogenetic trees based on gene content. Bioinformatics 20:2044-2049.
Kopp A. 2006. Basal relationships in the Drosophila melanogaster species group. *Mol Phylogenet Evol* 39:787-798.

Kopp A, and True JR. 2002. Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction. *Syst Biol* 51:786-805.

Korbel JO, Snel B, Huynen MA, and Bork P. 2002. SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics* 18:158-162.

Lin J, and Gerstein M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* 10:808-818.

McQuilton P, St Pierre SE, and Thurmond J. 2012. FlyBase 101—the basics of navigating FlyBase. *Nucleic acids research* 40:D706-714.

Pelandakis M, and Solignac M. 1993. Molecular phylogeny of Drosophila based on ribosomal RNA sequences. *J Mol Evol* 37:525-543.

Peterson BK, Weber JN, Kay EH, Fisher HS, and Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135.

Qi J, Luo H, and Hao B. 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic acids research* 32:W45-47.

Rebeiz M, Ramos-Womack M, Jeong S, Andolfatto P, Werner T, True J, Stern DL, and Carroll SB. 2009. Evolution of the tan locus contributed to pigment loss in Drosophila santomea: a response to Matute et al. *Nature* 440:1050-1053.

Rubin BE, Ree RH, and Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7:e33394.

Schawaroch V. 2002. Phylogeny of a paradigm lineage: the Drosophila melanogaster species group (Diptera: Drosophilidae). *Biological Journal of the Linnean Society* 76:21-37.
Snel B, Bork P, and Huynen MA. 1999. Genome phylogeny based on gene content. *Nat Genet* 21:108-110.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, and Kellis M. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450:219-232.

Stebbins GL. 1950. Variation and evolution in plants. Columbia University Press N.Y.

Stuart GW, Moffett K, and Leader JJ. 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol* 19:554-562.

Tekaia F, Lazcano A, and Dujon B. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res* 9:550-557.

Tengs T, LaFramboise T, Den RB, Hayes DN, Zhang J, DebRoy S, Gentleman RC, O'Neill K, Birren B, and Meyerson M. 2004. Genomic representations using concatenates of Type IIB restriction endonuclease digestion fragments. *Nucleic Acids Research* 32:e121.

Toda MJ. 1991. Drosophilidae (Diptera) in Myanmar (Burma) VII. The Drosophila melanogaster species-group, excepting the D. montium species-subgroup. *Oriental Insects* 25:69-94.

van der Linde K, and Houle D. 2008. A supertree analysis and literature review of the genus Drosophila and closely related genera (Diptera, Drosophilidae). *Insect Systematics & Evolution* 39:241-267.

van der Linde K, Houle D, Spicer GS, and Steppan SJ. 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet Res (Camb)* 92:25-38.

van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, and van Eijk MJ. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* 2:e1172.

Wang S, Meyer E, McKay JK, and Matz MV. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat Methods* 9:808-810.

Yang S, Doolittle RF, and Bourne PE. 2005. Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* 102:373-378.

Yang Y, Hou ZC, Qian YH, Kang H, and Zeng QT. 2012. Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the Drosophila melanogaster species group (Drosophilidae, Diptera). *Mol Phylogenet Evol* 62:214-223.

Yang Y, Zhang YP, Qian YH, and Zeng QT. 2004. Phylogenetic relationships of Drosophila melanogaster species group deduced from spacer regions of histone gene H2A-H2B. *Mol Phylogenet Evol* 30:336-343.

Yi H, and Jin L. 2013. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res* 41:e75.
Figure 1

Workflow of the entire process of generating phylogeny from the Type IIB fragments.
Collecting whole genome nucleotide sequence data from public databases.

*Phyper*: Extracts fragments and generate a representative list of fragments for each genome, for every enzyme.

*Phyppa*: Pairwise comparison of each of the 21 genomes with other genomes to calculate number of shared fragments between them.

Calculate pairwise distance using $- \ln \left( \frac{\text{Shared fragments}}{\text{Total fragments}} \right)$

Phylogenetic tree using *neighbor* program from the PHYLIP package
Figure 2

The consensus phylogenetic tree obtained by combining the trees obtained for each of the 13 enzymes.

The phylogenetic tree for each enzyme was calculated by extracting the corresponding fragments and then counting the number of shared fragment between every pair of species. The upper branch support values represent the percentage agreement over 13 enzymes and the bottom values indicate number of enzymes out of total 13 enzymes supporting the branch.
Figure 3

Single enzyme tree (A101 enzyme) showing the branch length.
Table 1 (on next page)

Various Drosophila species and source databases used for the analysis. The GC % for each genome was calculated using infoseq from the EMBOSS package.
| Genome          | GC %  | Size      | Source     |
|-----------------|-------|-----------|------------|
| D. ananassae    | 42.56 | 230.99 mb | FlyBase    |
| D. biarmipes    | 41.82 | 168.58 mb | NCBI       |
| D. bipectinata  | 41.62 | 166.39 mb | NCBI       |
| D. elegans      | 40.31 | 170.51 mb | NCBI       |
| D. erecta       | 42.65 | 152.71 mb | FlyBase    |
| D. eugracilis   | 40.90 | 156.31 mb | NCBI       |
| D. ficusphila   | 41.93 | 151.04 mb | NCBI       |
| D. grimshawi    | 38.84 | 200.46 mb | FlyBase    |
| D. kikkawai     | 41.38 | 163.57 mb | NCBI       |
| D. melanogaster | 42.05 | 168.73 mb | FlyBase    |
| D. mojavensis   | 40.22 | 193.82 mb | FlyBase    |
| D. persimilis   | 45.29 | 188.37 mb | FlyBase    |
| D. pseudoobscura| 45.43 | 152.73 mb | FlyBase    |
| D. rhopalosa    | 40.07 | 193.90 mb | NCBI       |
| D. santomea     | 38.52 | 165.75 mb | Princeton University |
| D. sechellia    | 42.53 | 166.57 mb | FlyBase    |
| D. simulans     | 43.06 | 137.82 mb | FlyBase    |
| D. takahashii   | 40.01 | 181.00 mb | NCBI       |
| D. virrilis     | 40.80 | 206.02 mb | FlyBase    |
| D. willistoni   | 37.89 | 235.51 mb | FlyBase    |
| D. yakuba       | 42.43 | 165.69 mb | FlyBase    |
List of enzymes used for the fragment generation from the 21 Drosophila species.

Frequency indicates estimated distance between cut sites given a random sequence with all the 4 bases in equal probability and length refers to blunt tag length.
| Enzyme | Recognition sequence  | Frequency | Length |
|--------|-----------------------|-----------|--------|
| AlfI   | GCANNNNNTGC           | 4096      | 32     |
| AloI   | GAACNNNTCC            | 8192      | 27     |
| BaeI   | ACNNNTAYC             | 4096      | 28     |
| BcgI   | CGANNTGC              | 2048      | 32     |
| BplI   | GAGNNNCTC             | 4096      | 27     |
| BsaXI  | ACNNNCTCC             | 2048      | 27     |
| BslFI  | GGGAC                 | 512       | 21     |
| Bsp24I | GACNNNTGG             | 2048      | 27     |
| CspCI  | CAANNTGG              | 8192      | 33     |
| Fall   | AAGNNCTT              | 4096      | 27     |
| HaeIV  | GAYNNRTC              | 1024      | 27     |
| PpiI   | GAACNNNCTC            | 8192      | 27     |
| Psrl   | GAACNNNTAC            | 8192      | 27     |
Table 3 (on next page)

Total number of fragments generated using 13 different Type IIB restriction enzymes for each of the 21 Drosophila genomes.
| Genomes          | Alfl  | AloI | Bael | Bcgl | BplI | BsaXI | BslFI | Bsp24I | CspCI | FallI | HaeIV | PpiI | Psrl |
|------------------|-------|------|------|------|------|-------|-------|--------|-------|-------|-------|------|------|
| D. ananassae     | 34804 | 11421| 6151 | 51646| 21457| 52433 | 101183| 46042  | 16405 | 38109 | 74174 | 11193| 8344 |
| D. biarmipes     | 41242 | 12667| 6875 | 63518| 22752| 51248 | 109404| 44554  | 18178 | 41284 | 75291 | 12177| 10210|
| D. bipectinata   | 35642 | 10893| 6616 | 51208| 20363| 50001 | 98937 | 45563  | 17131 | 39286 | 73197 | 10545| 8622 |
| D. elegans       | 43207 | 11314| 6068 | 59905| 18764| 45496 | 93763 | 43259  | 18466 | 41866 | 75238 | 11027| 9753 |
| D. erecta        | 42781 | 10517| 5914 | 60434| 18161| 43684 | 85735 | 40020  | 17793 | 31931 | 66412 | 9979 | 8677 |
| D. eugracilis    | 36455 | 10170| 5699 | 60448| 20161| 50001 | 98937 | 45563  | 17131 | 39286 | 73197 | 10545| 8622 |
| D. ficusphila    | 38374 | 11698| 5338 | 60448| 20161| 50001 | 98937 | 45563  | 17131 | 39286 | 73197 | 10545| 8622 |
| D. grimshawi     | 49667 | 5891 | 5212 | 61420| 17341| 50001 | 98937 | 45563  | 17131 | 39286 | 73197 | 10545| 8622 |
| D. kikkawai      | 39192 | 10361| 5516 | 60434| 18161| 43684 | 85735 | 40020  | 17793 | 31931 | 66412 | 9979 | 8677 |
| D. melanogaster  | 39711 | 9908 | 6037 | 5699 | 18764| 43684 | 93763 | 43259  | 18466 | 41866 | 75238 | 11027| 9753 |
| D. mojavensis    | 50786 | 11885| 7567 | 60434| 18161| 43684 | 93763 | 43259  | 18466 | 41866 | 75238 | 11027| 9753 |
| D. persimilis    | 39711 | 9908 | 6037 | 5699 | 18764| 43684 | 93763 | 43259  | 18466 | 41866 | 75238 | 11027| 9753 |
| D. pseudoobscura | 43650 | 10461| 7466 | 60434| 18161| 43684 | 93763 | 43259  | 18466 | 41866 | 75238 | 11027| 9753 |
| D. rhopalao      | 36920 | 10920| 6177 | 6203  | 18139| 44894 | 93524 | 41357  | 17133 | 40153 | 76711 | 10442| 9247 |
| D. santomea      | 40344 | 9877 | 5957 | 67771| 17044| 41850 | 80010 | 38107  | 17037 | 32142 | 67070 | 9414 | 8378 |
| D. sechellia     | 38976 | 10371| 5808 | 59204| 17430| 42659 | 83936 | 39380  | 17276 | 31541 | 68359 | 9792 | 8289 |
| D. simulans      | 38549 | 9815 | 5547 | 56820| 16777| 40735 | 79826 | 37436  | 16666 | 30304 | 64321 | 9148 | 7773 |
| D. takahashii    | 37489 | 11463| 5431 | 58887| 19189| 45240 | 91825 | 39992  | 26269 | 37277 | 74002 | 10801| 8987 |
| D. virrilis      | 58785 | 6943 | 5774 | 64912| 18097| 31951 | 66710 | 38679  | 15733 | 37692 | 65275 | 9290 | 8551 |
| D. willistoni    | 34033 | 7083 | 6177 | 43299 | 15103| 35578 | 70085 | 39996  | 17240 | 42202 | 77102 | 7941 | 9626 |
| D. yakuba        | 42202 | 10300| 6165 | 59442| 17885| 43748 | 83095 | 39920  | 18007 | 33024 | 69632 | 9887 | 8765 |