Title:

Sequence-based detection and breakpoint assembly of polymorphic inversions

Authors:

Russell B. Corbett-Detig*, Charis Cardeno§, and Charles H. Langley§

Author Affiliations:

* Department of Organismic and Evolutionary Biology. Harvard University. Cambridge, MA 02138.

Title: Sequence-based detection and breakpoint assembly of polymorphic inversions

Keywords: Structural variation, long-range LD, *In(3R)Mo*, *In(2R)NS*, *In(3R)K*, *In(1)Be*, *In(1)A*

Corresponding Author: Russell Brendan Corbett-Detig: russcd@gmail.com
Title:

Sequence-based detection and breakpoint assembly of polymorphic inversions

Abstract:

Inversion polymorphisms have occupied a privileged place in *Drosophila* genetic research since their discovery in the 1920’s. Indeed, inversions seem to be nearly ubiquitous, and the majority of species that have been thoroughly surveyed were found to be polymorphic for one or more chromosomal inversions. Despite enduring interest, however, inversions remain difficult to study because their effects are often cryptic, and few efficient assays have been developed. Even in *Drosophila melanogaster*, in which inversions can be reliably detected and have received considerable attention, the breakpoints of only three inversions have been characterized molecularly. Hence inversion detection and assay design remains an important unsolved problem. Here, we present a method for identification and local *de novo* assembly of inversion breakpoints using next-generation paired-end reads derived from *D. melanogaster* isofemale lines. PCR and cytological confirmations demonstrate that our method can reliably assemble inversion breakpoints, providing tools for future research on *D. melanogaster* inversions as well as a framework for detection and assay design of inversions and other chromosome aberrations in diverse taxa.

Introduction:

Sturtevant (1917) discovered the first chromosomal inversion as a suppressor of recombination in *D. melanogaster*. Shortly after his initial finding, Sturtevant produced evidence that inversions, structurally reversed segments of the linear map order of chromosome arms, account for this observation (Sturtevant 1926). A vast body of empirical work has followed this original discovery, yielding several key results that inform our
understanding of the genetic, and potential evolutionary implications of polymorphic inversions. First, single crossovers within the inverted regions of inversion heterokaryotypes are expected to yield inviable gametes, effectively suppressing exchange between arrangements (Sturtevant and Beadle 1936). Inversion heterokaryotypy redistributes chiasma elsewhere in the genome, both intrachromosomally in colinear segments, and via the interchromosomal effect (Lucchesi and Suzuki 1968).

This primary effect, suppression of recombination in the inverted regions of heterokaryotypes, especially near the breakpoints, is the subject of much of the theoretical population genetic research focused on inversions. Generally, interpretations in the literature favor a model in which inversions achieve high frequencies by suppressing recombination between coadapted alleles located near the inversion breakpoints (Dobzhansky 1951), although there are many other possible mechanisms (Kirkpatrick and Barton 2006, Hoffman and Reisberg 2008). Empirical research on polymorphic inversions has been extensive as well, the central result being that chromosomal inversions are pervasive. Indeed, segregating inversions are found in abundance in the majority of organisms that have been examined in depth, including plants, insects, mammals, and humans (Hoffman and Reisberg 2008). However, the selective forces that govern the evolution of inversion polymorphisms remain largely obscure, with important exceptions being inversions associated with novel sex chromosomes (Charlesworth et al. 2005), sex ratio distortion (Jaenike 2001), and autosomal segregation distortion (Kusano et al. 2003, Lyon 2003).

*D. melanogaster* is highly polymorphic for chromosomal inversions. In fact, since the pioneering work of Sturtevant (1931) and Dubinin et al. (1937), more than 500 segregating inversions have been found in natural populations of *D. melanogaster*, encompassing a broad
frequency spectrum ranging from present at low frequency in single populations, to present in virtually all populations worldwide (Ashburner and Lemeunier 1976, Aulard et al. 2004). This distribution is conventionally subdivided into four classes that correspond to the inversions’ prevalence in natural populations: unique endemic, recurrent endemic, rare cosmopolitan, and common cosmopolitan (Ashburner and Lemeunier 1976, Krimbas and Powell 1992).

The latter class has received by far the most attention. Common cosmopolitan inversions exhibit frequency clines, diminishing from high frequency in equatorial regions, to nearly absent in the subarctic regions. This pattern is replicated independently on several continents, suggesting strong selective forces govern the distributions of these inversions (Knibb et al. 1981). This observation has prompted numerous attempts to identify the traits that are experiencing selection, and several ecologically relevant traits have been associated with common cosmopolitan inversions (Van Delden and Kamping 1991, Kennington et al. 2007, Frydenberg et al. 2003). However, it remains unknown if the genetic elements that confer these traits are the targets of selection, or hitchhiking as a result of reduced recombination and the relatively young age of these inversions (Andolfatto et al. 1999, Matzkin et al. 2004).

Even less is known about the rare cosmopolitan and recurrent endemic inversions, which are comparatively understudied and sometimes not recorded in published frequency assays (Knibb et al. 1981, Krimbas and Powell 1992). The rare cosmopolitans are distributed worldwide, but often entirely absent from populations, while the recurrent endemic inversions may be at high frequency in one geographic region, but have rarely or never been identified elsewhere (Krimbas and Powell 1992). A detailed understanding of their limited
distributions and selective potentials is essential both as a comparison to the more "successful" common cosmopolitan inversions and to a nuanced and complete conception of polymorphic inversions in *D. melanogaster* and to the broader topic of genome evolution.

Despite continuing interest in the inversion polymorphisms of *D. melanogaster*, only three inversions, all common cosmopolitans, can be assayed directly using molecular means (Wesley and Eanes 1994, Andolfatto *et al.* 1999, Matzkin *et al.* 2005). Others must be identified via the laborious original method: crossing to a stock with known chromosomal arrangements and examining the banding patterns of the giant salivary gland polytene chromosomes in the larval progeny. In fact, all inversion breakpoints that have been characterized molecularly in *D. melanogaster* were identified using this convenient cytogenetic feature in combination with fluorescent *in situ* hybridization techniques (Wesley and Eanes 1994, Andolfatto *et al.* 1999, Matzkin *et al.* 2005). By hybridizing larval polytene chromosomes with DNA fragments of known mapping positions, it’s possible to narrow down the breakpoint regions through successively closer hybridizations (Wesley and Eanes 1994). This method is both time consuming and, perhaps most problematic, completely impractical for organisms that lack visible polytene chromosomes. Considering the largely quantitative goals of population genetic research, a more general and efficient means of inversion detection and assay design is essential to furthering our understanding of inversion polymorphisms in natural populations.

Genomic techniques have presented two appealing alternatives for identifying and characterizing structural polymorphisms segregating within populations. One method, originally developed by Tuzun *et al.* (2005), is based on sequencing both ends of short DNA fragments with an approximate known distance and orientation with respect to each other. By
first mapping paired reads to a reference sequence, and subsequently identifying clusters of read pairs that don’t map in the expected orientation or distance relative to one another, it is possible to reliably identify the breakpoints of structural polymorphisms (Tuzun et al. 2005, Kidd et al. 2008, Medvedev et al. 2009). This approach is appealing because it can be used to fine-map structural breakpoint and it has previously been validated as a tool for interrogation of structural polymorphisms in *D. melanogaster* (Cridland and Thornton 2010). While these methods have high sensitivity for breakpoint detection, it is often not possible based solely on the breakpoints to distinguish inversions from other structural rearrangements, such as duplications that reinsert in inverted orientation (Cridland and Thornton 2010). The relatively short length of inserts that are currently used in the majority of resequencing projects, as opposed to long inserts used in the previous landmark studies (Tuzun et al. 2005, Kidd et al. 2008), may exacerbate this issue, as short-inserts provide little resolution beyond detected breakpoints.

Alternative approaches, which rely on data from many densely genotyped individuals, use an expected signature of nucleotide variation to detect polymorphic inversions (Bansal et al. 2007, Sindi and Raphael 2010). While these methods have provided valuable insights and many candidate inversions, the inherent genomic limitations of SNP genotype data has proved to be a major impediment. To accurately predict inversions, these methods require substantial minor allele frequencies of inversion and large sample sizes (Bansal et al. 2007, Sindi and Raphael 2010). Additionally, genotyping approaches offer poor resolution of inversion breakpoints, which may be of interest for population genetic analyses, designing PCR assays, as well as for the study the mechanisms of inversion formation and DNA repair.
Here we present a hybrid method of inversion detection that incorporates features of breakpoint and genotype-based methods outlined above. Briefly, our method infers putative breakpoints using map positions of paired-end reads across many samples. We subsequently identify breakpoints that are shared between samples based on overlap of detected breakpoints. We then filter potential breakpoints by scanning the surrounding regions for a signature of nucleotide variation, heightened $F_{ST}$, which is expected to be associated with inversion breakpoints. We apply this method to more than 100 *Drosophila melanogaster* haploid genomes, in which we identify breakpoints and design molecular assays for five previously uncharacterized polymorphic inversions. We also improve on existing assays for two previously characterized inversions. Cytology as well as previously developed PCR assays confirm that our method is highly accurate. We expect these primer pairs will be immediately useful as tools for researchers interested in assaying inversions directly in individual *D. melanogaster* or existing stocks. Additionally, this general framework could be implemented to detect and design molecular assays for structural polymorphisms in other species.

**Methods:**

Details for all fly stocks used in this study can be found at www.dpgp.org. In short, all short read sequences are derived from 76bp Illumina paired end reads, separated by ~300 bp inserts, and represent numerous African populations and one French population. All genomic regions analyzed are haploid either through chromosomal extractions, inbreeding, or haploid embryo extractions (Langley *et al.* 2011). The average coverage depth is approximately 31x (range 8 to 47).
We aligned short reads for each line to the *D. melanogaster* reference sequence v5.22 (Adams *et al.* 2000) using ELAND v2 as a standard part of the Illumina Casava pipeline. Alignments were performed over the course of more than a year; as such several versions of the Casava pipeline were used (see [www.dpgp.org](http://www.dpgp.org) for details). Eland alignments were ported to MAQ using the ‘export2maq’ utility in the MAQ v0.7.3 software package (Li *et al.* 2008).

We called consensus sequences for each line, requiring that each site called have a minimum read depth of 3 and minimum quality of 30. All other sites were excluded from the resulting assemblies. We identified 7 lines showing excess identity by descent as those with regions of little pairwise divergence on more than one chromosome arm, these are obvious in cursory inspections, so this was done by hand and primarily excluded genomes that are drawn from the same isofemale line. From each pair showing IBD, we discarded the line that was sequenced to lower depth.

From the alignment files, we discarded read pairs that share identical mapping coordinates with another pair, as well as those for which one read maps to an annotated transposable element. Next, we parsed read pairs for which both reads mapped uniquely to the reference sequence but mapped in parallel orientation (>> or <<). We restricted this set of reads to those that mapped to the same chromosome arm. For each line individually, we assigned aberrant read pairs to clusters, requiring that both reads in a pair mapped within 500 base pairs of another read included in that cluster. Thus each cluster contains sets of read pairs for which one of the pair maps to one 500bp region of the genome, and the other read maps to another 500bp region of the same chromosome arm. We further required that all reads in one cluster map in the same orientation and that genomic positions within each cluster be more than 1MB from each other. Clusters supported by fewer than five clones
were discarded. Finally, we parsed from the “export.txt” files read pairs for which one read maps to the same genomic location and in the same orientation as the clusters identified previously, and the other read is unmapped. The unmapped reads are expected to cross the breakpoint. We then folded these reads into the original cluster (Figure 1).

We compared each line’s set of read clusters to all other lines’ sets of read clusters. Overlapping clusters in the same genomic position and orientation were identified as potentially confirming the same inversion breakpoint. Due to their unique origins, which results in zero polymorphisms in the inverted population at formation, inversions breakpoints will immediately attain high $F_{ST}$ relative to the standard arrangement population. As genetic exchange is almost completely suppressed near inversion breakpoints (Novitski and Braver 1954, Wesley and Eanes 1994, Andolfatto et al. 1999), genetic differentiation will be maintained between arrangements and is an expected signature of all inversion breakpoints.

Provided an inversion is present in more than two individuals, this expectation suggests an ideal way to sift through identified breakpoints for inversion false positives. For lines sharing identical breakpoints, we compared the consensus sequences in 20KB windows centered on each potential breakpoint, and calculated $F_{ST}$ between lines that share the putative breakpoint and lines that do not. We retained candidate inversions for which both breakpoints’ $F_{ST}$ was greater than .25. We calculated $F_{ST}$ as described in Hudson et al. (1992), using only sites that were called in all lines. Importantly, we did not weight $F_{ST}$ by sample size, which enables the detection of low frequency inversions. Because of this, even immediately after formation, we expect to observe strong genetic differentiation and inversions’ $F_{ST}$ will initially be approximately .5.
We de novo assembled the set of reads corresponding to each remaining potential breakpoint using Phrap v1.090518 (Green 1996). All contigs were aligned to the *D. melanogaster* reference sequence v5.22 (Adams et al. 2000), using Blast v2.2.25 (Altschul et al. 1990), and contigs with significant alignments to both sides of the expected breakpoint were retained. Blast alignments with corresponding e-values below $10^{-10}$ and alignment lengths greater than 30bp were considered significant.

To assist in assigning identities for novel inversions, we compared the cytogenetic positions of the inversion’s breakpoints with reported breakpoint coordinates. To do this, we downloaded the cytologically predicted positions of inversion breakpoints and the map conversion table for cytological coordinates from flybase (www.flybase.org).

Inversion breakpoints may take two forms, cut-and-paste and inverted-duplication, (see Ranz et al. 2007 for a description of breakpoint structure). After aligning breakpoint-spanning contigs to the reference genome, we inferred the structure based on the following criteria. If both breakpoint-spanning contigs appear to map in convergent orientations to within 50bp of each other at both ends of the inversion, they are assumed to be cut-and-paste breakpoints. Otherwise, we assume that the sequence between mapping positions is present as a duplication at the other breakpoint. We confirmed these structural predictions via comparisons against the three inversions that have previously been examined (Wesley and Eanes 1994, Andolfatto et al. 1999, Matzkin et al. 2004), and by comparison against the copy number variation analysis performed by Langley et al. (submitted), whose stocks are known to bear many of these inversions.

To confirm breakpoints, we developed a PCR-based inversion assay. We designed primers using Primer3 (http://frodo.wi.mit.edu/primer3/) that would produce an amplicon
unique to the standard or inverted chromosomal arrangement based on these putative
breakpoints. We extracted genomic DNA from flies using the “Quick Fly Genomic DNA
Prep” provided by the Berkeley Drosophila Genome Project (http://www.fruitfly.org/about/methods/inverse.pcr.html). Briefly, we ground 30 flies in Buffer A (100mM Tris-HCl, 100mM EDTA, 100mM NaCl, 0.5% SDS) and incubated at 65°C for 30 minutes. We added a 1:2.5 solution (1 part 5M KAc to 2.5 parts 6M LiCl) to the samples and incubated on ice for at least 10 minutes. The DNA was precipitated with isopropanol, washed and resuspended in ddH2O.

All of the PCR inversion assays (except for the standard chromosomal arrangement of In(3R)P) used standard PCR reaction conditions: 2.0mM MgCl2, 0.2mM each of dNTPs, 0.5mM each of forward and reverse primers, 1U of Taq and 50ng of DNA. Specific PCR conditions for each reaction are described in Supplemental Table 1. Appropriately sized amplicons were identified with agarose gels. We used long PCR to assay the In(3R)P standard chromosomal arrangement. This was necessary because the inverted duplications present at each breakpoint are too long for standard PCR. We followed the manufacturer’s PCR general reaction mixture and conditions (TaKaRa LA Taq™) with a few exceptions – a final MgCl2 concentration of 1.75 mM (Supplemental Table 1) and the annealing/elongation time was 5 minutes. We included the reference strain, y; cn bw; sp, as well as several other standard orientation lines as negative controls for inversion-specific primers and positive controls for standard-specific primers. For inversion positive controls, we included obtained several lines known by cytology to harbor the putative inversion from the Bloomington stock center and Langley et al. (Submitted, Supplemental Table S1).
For each inversion that had not been detected previously, we sequenced via Sanger PCR at least one breakpoint to further validate Phrap assemblies. Sequences were assembled from forward and reverse chromatograms using phredPhrap, which is distributed as a part of the Consed package. We inspected all assembled PCR fragments by hand in Consed v1.090518 (Gordon et al. 1998). These sequences are available in supplemental file 1.

Results:

Across all genomes, we recovered more than 15000 breakpoints that map in parallel orientation to a single chromosome arm. After pooling across all samples, we found over 200 breakpoints that were present in more than one line. Finally after applying the \( F_{ST} \) filter, we found twelve aberrant read clusters whose corresponding consensus sequences showed increased \( F_{ST} \) around both breakpoints. Heightened \( F_{ST} \) is an expected signature of nucleotide variation between inverted and standard arrangements owing to the unique origin of inversion and suppressed exchange between arrangements immediately surrounding each breakpoint (Novitski and Braver 1954, Wesley and Eanes 1994, Andolfatto et al. 1999).

Importantly, heightened \( F_{ST} \) at both breakpoints is not expected for breakpoints associated with other rearrangements that may occur at higher frequencies. This is because only a single breakpoint actually harvests the novel insertion event; the other ‘breakpoint’ reflects reads that map uniquely to the single-copy present in the reference sequence, but are not actually linked to this genomic location. We also surveyed all clusters present in more than one line for breakpoints consistent with cytologically known inversions that have been identified in populations from sub-Saharan Africa (Aulard et al. 2002). We did not find any additional
breakpoints that were consistent with these inversions, thus $F_{ST}$ appears to be a successful
filter for inversion true positives.

We successfully assembled contigs that spanned all breakpoints identified. Since five
pairs of breakpoints were in perfect linkage disequilibrium and the breakpoint coordinates are
very similar, we surmised that these breakpoints corresponded to the same inversion. Thus,
we were able to recover both breakpoints for five inversions, and only one breakpoint for two
others (Supplemental Table S1). That we only recovered a single breakpoint for $In(2L)t$ and
$In(3R)P$ is likely due to the presence of fixed repetitive elements immediately adjacent to the
proximal breakpoints of each inversion (Andolfatto et al. 1999, Matzkin et al. 2005).

In a recent study, Langley et al. (Submitted) found a pattern of excess long distance
linkage disequilibrium and significantly decreased nucleotide diversity associated with a
large paracentric inversion, $In(3R)Mo$, in a Raleigh, NC population of $D. melanogaster$.

Because the sequence data used in this study was derived from single-end reads, it is not
suitable for direct comparison via our method, which relies on independently mapped paired-end
reads. We did identify one line, FR310, which shares this pattern of long distance linkage
disequilibrium and that had been sequenced using paired-end reads. Because of this
inversion’s unexpected prevalence in this Raleigh, NC population, we surveyed this line
specifically for potential breakpoints, without requiring that the identified read clusters be
corroborated by clones derived from another line. We found two breakpoints whose positions
are consistent with our expectations for this inversion based on the observed pattern of
nucleotide diversity, and confirmed these breakpoints via PCR in the eight lines identified by
Langley et al. (Submitted).
We recovered two classes of breakpoints, simple cut-and-paste breakpoints, and staggered break plus inverted duplication breakpoints (Figure 2). We were able to confirm structural predictions for three inversions whose breakpoints were previously characterized (Wesley and Eanes 1994, Andolfatto et al. 1999, Matzkin et al. 2004). In(3R)K, In(3R)Mo, In(1)A and In(2R)NS are all present in the sample analyzed by Langley et al. (submitted) and we were able to support breakpoint structure predictions for each based on comparisons to the copy number variation analysis included in that work. Five of the eight inversions contain inverted-duplications at their breakpoints (Supplemental Table S2), which is similar to the proportions that were found by a study that focused on inversions that fix between species in the Melanogaster subgroup (17 out of 29, Ranz et al. 2007).

We designed assays for each class that amplifies a product of unique length for either the standard or the inverted haplotype. Cut-and-paste breakpoints can be assayed easily as described in Andolfatto et al. (1999). Because in staggered-break inversion structures there is no single breakpoint that is unique to the standard arrangement, primers that span a single breakpoint cannot be used to distinguish between inversion heterozygotes and inversion homozygotes. Our solution is to design primers that span the duplicated regions at either end of the inversion (Figure 2). This produces a PCR product that is unique to the standard arrangement, and may be more robust than an allele-specific PCR approach (e.g. Anderson et al. 2005).

To make use of this advantage, we designed new primers for the standard arrangement of In(3R)P (Supplemental Table 1), for which only ambiguous or allele-specific primers were previously available. Likely due to the presence of a repetitive element immediately adjacent to the proximal breakpoints of In(3R)P (Matzkin et al. 2005), we only
recovered a single breakpoint for this inversion. Fortunately, Matzkin et al. (2005) have previously sequenced both breakpoints via long-range PCR for \(\text{In}(3R)P\). We downloaded these sequences from Genbank (Ascension numbers: AY886890–AY886892), and used these to design a novel set of primers that yield a unique amplicon for the standard arrangement. All PCR fragments that we sequenced are identical to the phrap assemblies, except in the low quality bases towards the ends of the traces (Supplemental File S1).

For the three inversions that have previously been characterized at the molecular level, breakpoint coordinates are known and we identified these breakpoints directly in our sequence data. We also confirmed our results for these inversions using published primers (Supplemental Table S3). For \(\text{In}(3L)P\), the existing primers did not work reliably. We elected to design new primers (Supplemental Table S1), and have found these to be more reliable. For four other inversions, all putative inversion identities were confirmed by positive controls (Supplemental Table S1).

For all previously uncharacterized inversions, the cytologically-derived mapping positions based on previous surveys were within 100kb of the breakpoint we identified. In most previously uncharacterized inversions, it was also possible to test our primers on stocks known via cytology to bear the inversion. This was not possible for one inversion on the X chromosome for which no independent positive controls are available. However, the breakpoint coordinates, geographic distribution, and frequency of this inversion are all consistent with \(\text{In}(1)Be\) and do not suggest any other known inversions. Hence, although we cannot be certain of the identity, we refer to this inversion as \(\text{In}(1)Be\) (Supplemental Table S1).

**Discussion:**
While our method was quite successful and has immediate applications to many short-read sequencing projects, it should be noted that there are two important drawbacks, both of which will be ameliorated by imminent advances in sequencing technologies. First, our method requires accurate mapping information and a well-characterized reference genome. Already several species’ genomes have been fully assembled using next-generation sequencing technologies (e.g. Li et al. 2009); hence the anticipated availability of many additional reference sequences may make the proposed method widely serviceable. Second, the extent to which transposable elements contribute to the formation of chromosomal inversions remains an open question (Caceres et al. 1999, Mathiopoulos et al. 1998).

Although this does not appear to be a common mechanism of inversion formation in the D. melanogaster subgroup (Ranz et al. 2007), it is possible that transposable elements contribute more to structural polymorphisms in other species. Because of the modest insert lengths used in sequencing, our method has little power to detect inversions that form via ectopic recombination between repetitive elements. However, this limitation will also diminish in importance with the increasing availability and quality of larger insert sizes in library preparation, which will be able to span individual repetitive elements. Hence, if anything, the applicability and usefulness of this approach will increase as sequencing technologies continue to progress.

Despite these potential drawbacks, our method has performed well. A recent survey of African D. melanogaster inversion polymorphisms (Aulard et al. 2002) reported generally the same set of polymorphic inversions at moderate frequencies. So, while we cannot estimate a true false negative rate, this suggests we have recovered the majority of inversions that are likely to be segregating at frequencies greater than two in this sample. Requiring $F_{ST}$
calculations means that our method will miss inversions present in only one individual. This
drawback is unavoidable, since there are thousands of aberrant read clusters in individual
genomes, and it is not always possible to distinguish between inversions and other structural
variants based solely on breakpoint coordinates. Regardless of the error rates, our method is a
vast improvement over conventional methods, and it allows us to rapidly characterize and
develop novel molecular assays for five chromosomal inversions and to improve on two
existing assays. This more than doubles the available assays, providing a substantial
improvement in the tools available for studies of the polymorphic inversions of \(D.\)
\textit{melanogaster}.

Another advantage of this approach is its broad applicability. Cytological methods,
before being time consuming, require visible polytene chromosomes, as well as an
approximate idea of where inversion breakpoints might be expected to occur and in what
strains. Our method circumvents these issues, and it allowed us to examine numerous
individuals simultaneously and to identify polymorphic inversions without requiring any
prior knowledge of the lines or inversion content of the genome. Hence, we expect this will
be a useful framework for researchers interested in characterizing and developing molecular
assays for polymorphic inversions, especially in developing model systems. Although we
analyzed haploid data, this method could feasibly be extended to accommodate diploid
samples provided sufficient sequencing or sampling depth.

It should also be emphasized that clustering putative breakpoints across samples
would allow detection not only of inversions but also of other types of chromosome
aberrations. These need not necessarily be aberrations transmitted through the germ line. For
example, our approach may have applications in the study of rearrangements among somatic
or cancer cells where relevant independent sampling can be conducted.

Chromosomal inversion polymorphisms are a ubiquitous evolutionary phenomenon. They are present in virtually all species, and may have potent evolutionary effects ranging from resisting gene flow in hybrid zones, to maintaining co-adapted gene complexes, to the long-term maintenance of epistatically interacting segregation distortion systems (Hoffmann and Reisberg 2008). However, a complete understanding of the selective effects of polymorphic inversions is elusive. Even—perhaps especially—in the species in which chromosomal inversions were originally discovered, *D. melanogaster*, inversions remain an enigmatic, and intriguing, feature of virtually all populations. A central impediment to quantitative studies of these polymorphisms, especially rare cosmopolitan and recurrent endemic inversions, is a lack of low-cost efficient assays. Here, we provide these tools. Although there are certainly numerous interesting population genetic questions that could feasibly be addressed using this data (many of which are subjects of ongoing research), our goal with this work is to make these resources available to the community as soon as possible.

**Acknowledgements:**
We thank John Pool, Daniel Hartl, and two anonymous reviewers for helpful comments on this manuscript, as well as Shu Fang for providing cytology information. We also acknowledge funding sources. NIH grant HG02942 to CHL. RBC is supported by a Harvard Prize Fellowship.

**Literature Cited:**
Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J


Langley et al. DPGP. Submitted.


Figure 1: Mapping positions of reads used for *de novo* assembly of the breakpoint sequences. Above, reads mapping positions on the reference sequence, and below their inferred positions on the inverted haplotype. Reads pairs that span the breakpoint are shown in shades of blue, while reads for which one end maps, and the pair crosses the junction on the reference sequence are shown in shades of green. All reads shown are used to produce *de novo* assemblies. For simplicity, reads corresponding to this inversion’s other breakpoint are not shown.
Figure 2: We found two types of inversion breakpoints: ‘cut-and-paste’ breakpoints (above), and staggered breakpoints (below), which create inverted duplications at the breakpoints. The duplicated regions are shown as purple and blue “genes”. PCR primers for the standard (in orange) and inverted (in teal) arrangements were designed based on assembled contigs that span the breakpoints, and amplify a unique product for either arrangement.