Proposal to Sequence a *Drosophila* Genetic Reference Panel: A Community Resource for the Study of Genotypic and Phenotypic Variation

Trudy Mackay, Stephen Richards, Richard Gibbs

Overview: We propose the sequencing of a *D. melanogaster* genetic reference panel of 192 wild-type lines from a single natural population which have been inbred to homozygosity, and for which extensive information on complex trait phenotypes has been collected. This will create: (1) A community resource for association mapping of quantitative trait loci. Within this project we will demonstrate such mapping and provide candidate quantitative trait polymorphisms for traits relevant to human health. (2) A community resource of common Drosophila sequence polymorphisms (SNPs and indels) with a minor allele frequency (MAF) of 0.02 or greater. These variants will be valuable for high resolution QTL mapping as well as mapping alleles of major effect, molecular population genetic analyses, and allele specific transcription studies, among others. (3) A "test bench" for statistical methods used in QTL association and mapping studies for traits affecting human disease.

The proposed genetic reference panel of sequenced homozygous lines has many advantages and creates a new innovative genetics tool for the *Drosophila* community. First and foremost, each line represents a homozygous genotype that can be made available to the entire community. The same strains can be evaluated for multiple complex traits, including 'intermediate' phenotypes such as whole genome transcript abundance and quantitative variation in the proteome and metabolome. This will facilitate a systems genetics approach for understanding the genetic architecture of complex traits in an economical genetic model organism. Interrogating a common resource population for genetic variation at multiple levels, traits, and environments will provide an unprecedented opportunity to quantify genetic correlations and pleiotropy among traits, as well as to quantify the magnitude and nature of genotype by environment interaction. Trait values can be ascertained with a high degree of accuracy by evaluating multiple individuals per strain. A sample of 192 strains is sufficiently large to include minor allele variants with a frequency of 0.02 or greater, and has the power to detect intermediate frequency variants with moderately small to large effects on complex traits. Re-sequencing a sample of 192 strains is also experimentally and economically feasible, given the small size and high quality of the *Drosophila* reference genome, and the use of massively parallel sequencing technology. The sequence information will be used for association mapping studies for phenotypes that are in the current database to give an immediate payoff in terms of *Drosophila* quantitative trait genes that are candidate genes for human complex traits. These strains will provide a long term resource for the *Drosophila* community. Candidate genes for any complex trait can be identified by quantifying the trait phenotype in the reference panel of sequenced strains. Since the lines are a living library of all common polymorphisms affecting natural variation for any trait of interest, they can be used by members of the *Drosophila* community to identify extreme lines for QTL mapping – the lines are already inbred and therefore can be used immediately to construct mapping populations. They can also be used as a base

population for artificial selection experiments, in which lines can be derived with trait phenotypes that greatly exceed the range of the base population. Additionally, this will facilitate the development of a common set of dense polymorphic markers that can be used to develop an economic and accurate platform for genotyping the thousands of recombinant lines or individuals required for accurate mapping of QTLs.

The Flies – A Genetic Reference Panel for Mapping and Cloning Quantitative Trait Genes:

The Mackay lab has recently derived a set of 192 inbred lines from the Raleigh, NC natural population by inbreeding isofemale lines to homozygosity by 20 generations of full sib mating. The homozygosity of these lines has been verified by analysis of microsatellite markers and re-sequencing of several regions on all three major chromosomes; less than 5% of the lines exhibit residual heterozygosity at one locus on 3R. This genetic reference panel has been extensively phenotyped for a battery of complex traits, and constitutes a long-term resource for further phenotyping and experimentation by the *Drosophila* community. The reference panel will be sequenced to 10-12 X coverage using the 454 500bp XLR pyrosequencing technology, and additionally to 6-8X using the Illumina short read platform for error correction around homopolymers and validation of identified polymorphisms. The long read data will be assembled *de novo* to allow full characterization of insertions, deletions and inversions.

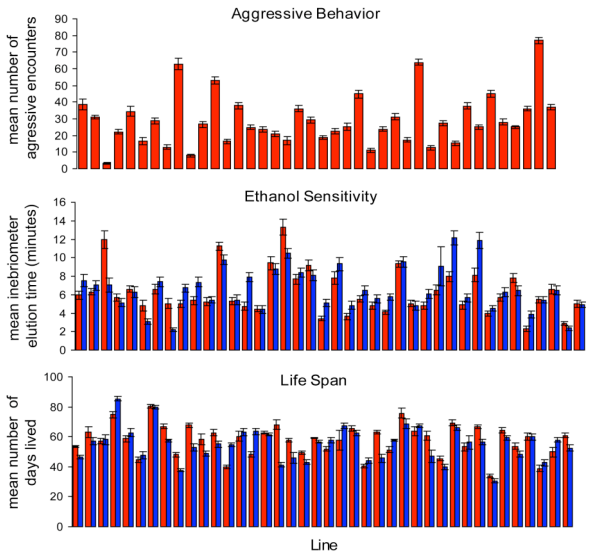


Figure 1. Variation in three quantitative traits in 40 of the proposed 192 line *Drosophila* genetic reference panel. Red: males; Blue: females.

The Full Data Set – Quantifying Variation in Complex Trait Phenotypes: The Mackay laboratory has quantified variation among all 192 of these lines for longevity; resistance to starvation stress and chill coma recovery; aggressive, locomotor, olfactory and mating behavior; alcohol sensitivity; and numbers of sensory bristles. We plan to initiate sequencing on a core set of 40 of the Raleigh lines, followed by the remainder of the strains. Therefore, the community is focusing initially on obtaining phenotypic information on this core set of lines. The lines exhibit a great range of variation for all traits (Figure 1, Appendix 1), with broad sense heritabilities ranging from 0.22 – 0.78 (Appendix 2). In many cases, the range of variation among this panel of lines is comparable to, and in some cases even exceeds, the difference in mean phenotype between lines subjected to divergent artificial selection for the traits (e.g., Edwards et al., 2006).

Currently, the Mackay lab is assaying the 40 core lines for variation in oxidative stress resistance, competitive fitness, sleep, behavioral responses to a battery of drugs (e.g., dopamine, serotonin, caffeine, nicotine, alcohol), and whole genome transcript abundance

(using Affymetrix Dros2.0 GeneChips). There is no doubt that these lines will vary for every complex trait for which a quantitative phenotypic assay can be developed, including traits of direct relevance to human health, such as variation in immune competence, learning and memory, lipid metabolism, responses to addictive drugs, and 'intermediate' phenotypes such as enzyme activity. This subset of the reference panel is also ideal for assessing the magnitude of genotype by environment interaction for complex traits, since the same lines can be reared under multiple environments.

The community phenotyping effort will build upon the extensive foundation provided by the Mackay laboratory. Members of the fly community have already committed to phenotyping the genetic reference panel for a number of traits relevant to the NIH mission. This includes variation in lipid and protein levels (see letter from Dr. Maria DeLuca); learning and memory (letter from Dr. Frederic Mery); immune challenge (letter from Dr. Jeff Leips); foraging behavior (letter from Dr. Marla Sokolowski); adult olfactory behavior in response to a battery of odorants (letter from Dr. Robert Anholt); larval olfactory behavior in response to the same odorants (letter from Dr. Juan José Fanara); development time and adult body size (letter from Drs. Estaban Hasson and Juan José Fanara); ovariole number (letter from Dr. Marta Wayne); circadian rhythm, cuticular hydrocarbons and social behaviors (letter from Joel Levine); wing morphology (letter from David Houle); and sperm precedence (letter from Dr. Kimberly Hughes).

All phenotype data will be publicly available for all traits. As community members add each new phenotype to the database, they will be able to assess genetic correlations with all other traits that have been studied to date, thus building an unprecedented and comprehensive picture of the *Drosophila* phenome that would not be possible if all investigators used different strains.

Whole Genome Association Studies: One immediate utility of the complete genome sequences of the Raleigh inbred lines will be to perform whole genome association studies for the complex trait phenotypes in the database. The database will include variation among the lines in whole genome transcript abundance; therefore, the availability of whole genome sequence for each line will also provide the first opportunity for genome wide assessment of the relationship between DNA sequence variation, variation in transcript abundance, and variation in quantitative trait phenotypes.

Power Considerations: The power of using inbred lines for association studies is much greater than that of outbred individuals for two reasons. First, the genetic variance of a population of fully inbred lines is at least twice that of an outbred population at Hardy-Weinberg equilibrium (Falconer and Mackay, 1996), because all individuals are homozygotes for segregating alleles. Second, the ability to obtain replicate measurements of multiple individuals per inbred line gives an accurate estimate of the mean phenotypic value of each line, greatly reducing the noise due to environmental variance. To illustrate this, consider the power to detect an association for a marker causally affecting the trait at a frequency of $q = 0.5$, under three scenarios: (1) a sample of 192 outbred individuals; (2) one individual from each of 192 inbred lines, and (3) many individuals from each of 192 inbred lines. Standard statistical theory gives the relationship between n, the number of

replicates per group (i.e., individuals or lines with alternate alleles at the polymorphic marker), and the magnitude of the difference in phenotype associated with the marker (δ) as $n \geq 2(z_\alpha + z_{2\beta})^2/(\delta/\sigma_P)^2$ (Sokal and Rohlf, 1981); where σ_P is the within-group standard deviation; α and β are, respectively, the Type I and Type II significance levels set; and z is the ordinate of the normal distribution corresponding to its subscript. Let $\alpha = 0.05$ and $\beta = 0.1$. (1) With 192 outbred individuals and $q = 0.5$, we expect 48 homozygous individuals for alternate marker genotypes, and the power to able to detect differences of $0.661\sigma_P$ between homozygous genotypes. (2) With the core set of 40 inbred individuals and $q = 0.5$, we expect 96 homozygous individuals for alternate marker genotypes, and the power to able to detect differences of $0.468\sigma_P$. (3) If multiple individuals are measured per inbred line, the phenotypic variance is that of line means, or σ_P^2/N, where N is the number of individuals measured per line. If $N = 20$, we will be able to detect effects of $(\sigma_P/\sqrt{20})(0.468) = 0.105\sigma_P$; if $N = 40$, we have the power to detect effects of $0.074\sigma_P$. To put this in perspective, the sample of 40 inbred lines is equivalent to an outbred population of 7,680 individuals for $N = 20$ replicate measurements per line, and an outbred population of 15,360 for $N = 40$ replicate measurements per line. For the core set of 40 lines, and $N = 40$, we have the power to detect effects of $0.162\sigma_P$, equivalent to 3,200 outbred individuals. Appendix 3 shows these effects in real units of measurement for each trait, and as a percent of the population mean. The power declines as gene frequencies depart from 0.5, but the tendency for rare alleles to have larger effects somewhat counteracts this (Carbone et al., 2006). Significant associations between molecular polymorphisms and quantitative trait phenotypes have previously been documented for *Drosophila* studies of this magnitude (Mackay and Langley, 1990; Lai et al., 1994; Long et al., 1998; Lyman et al., 1999; Robin et al., 2002; DeLuca et al., 2003; Carbone et al., 2006). There is growing evidence that the distribution of effects of alleles affecting complex traits is exponential; i.e., many alleles with small effects, but a few with large effects that contribute most of the trait variance (Robertson, 1967; Dilda and Mackay, 2002). We will have the power to detect variants in the latter, more important tail of the distribution, but not to detect variants with very small effects.

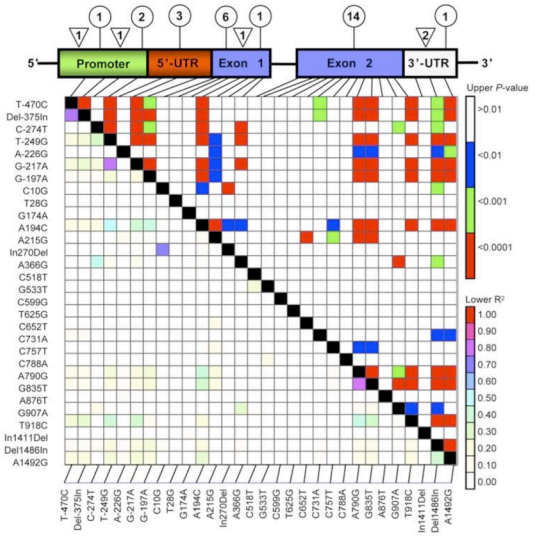


Figure 2. *Catsup* polymorphisms show an absence of haplotype blocks. The *Catsup* gene structure is depicted with the number and distribution of SNPs (circles) and InDels (triangles) in 169 *Catsup* alleles sampled from the Raleigh population. LD in *Catsup* is shown below the gene structure, with *P*-values from Fisher's exact test above the diagonal and estimates of r^2 below the diagonal (from Carbone et al., 2006). Note the very low r^2 values throughout this 2 kb region.

Absence of Haplotype Blocks Allows Direct Allele Identification: In humans, the average pairwise nucleotide diversity is 0.001/bp, and linkage disequilibrium between polymorphic markers follows a block-like pattern, in which polymorphisms in close physical linkage often forms blocks of markers in strong linkage disequilibrium (haplotype blocks), separated by regions of high recombination (International-HapMap-Consortium, 2005). Thus, the human scenario is excellent for using reduced numbers of markers as proxies for each haplotype block, simultaneously

reducing the genotyping effort in a whole genome association scan while increasing the number of genes and markers in the block that could be causally associated with variation in the trait. In contrast, *D. melanogaster* is highly polymorphic, with an average nucleotide diversity of 0.004/bp for coding regions and 0.01/bp for non-coding regions (Moriyama and Powell, 1996), and linkage disequilibrium between polymorphic sites decays rapidly with physical distance in normal regions of recombination (Long et al., 1998; Carbone et al., 2006). It is not uncommon for *Drosophila* polymorphic sites less than 10 bp apart to be in linkage equilibrium (Figure 2, Carbone et al., 2006). Thus, *Drosophila* is excellent for identifying polymorphisms causally associated with variation in complex traits, but the penalty is that complete sequence information is required.

Multiple Testing, Association Tests and Followup Experiments: The large number of association tests to be done for each trait poses a multiple testing problem. Previously, two variants of permutation tests have been used to address this issue (Churchill and Doerge, 1994; Doerge and Churchill, 1996). The first test asks whether more polymorphic sites in each gene than expected by chance are associated with variation in the trait (Lai et al., 1994; Carbone et al., 2006; Jordan et al., 2006), thus nominating a candidate gene for further study. The second asks whether a particular polymorphic site is more significant than expected by chance (Long et al., 1998; Robin et al., 2002; Carbone et al., 2006), thus selecting individual polymorphisms for further study. False discovery rate methods developed in the context of microarray data analysis (Storey and Tibshirani, 2003) will also be applicable to these analyses. The existence of comprehensive phenotypic and genotypic data is likely to spur the development of further statistical methods (letters from Drs. Rebecca Doerge and Lauren McIntyre). However, a major advantage of using *Drosophila* is that a lenient false positive rate can be tolerated. Individual investigators can test candidate genes of interest for functional significance using complementation tests of mutations in candidate genes to lines with alternative QTL alleles, and expanding the association test by phenotyping other populations for individual polymorphisms, or re-sequencing candidate genes using conventional methods.

Genome Wide Molecular Population Genetics: The proposed genome sequences will enable integration of population genomic analyses with patterns of phenotypic variation. Although on average *Drosophila* is highly polymorphic and linkage disequilibrium decays rapidly with physical distance, there is great variation in polymorphism and linkage disequilibrium throughout the genome, reflecting the interplay of mutation, recombination, natural selection and population history. Thus, whole genome data will be used to assess which regions are evolving according to the neutral expectation and which show the signatures of natural selection, by applying tests for departure from neutrality on a genome-wide scale. These include tests for more putatively functional mutations than expected by chance, tests for an excess or reduction of nucleotide diversity, as expected if polymorphism is maintained by a form or balancing selection or has been reduced by a recent ‘sweep’ of a beneficial allele, respectively; a high frequency of derived alleles, as expected in regions that have undergone a selective sweep; and regions of excess linkage disequilibrium, as expected for recently selected alleles for which recombination has not yet broken down associations with linked variants (Sabeti et al.,

2006). Several of these tests require sequence from closely related species and an outgroup sequence. The recent accumulation of whole genome polymorphism data from *D. simulans* as well as whole genome sequence of *D. yakuba* will be greatly informative in this regard. Application of this battery of tests on a genome wide scale will reveal particular genes and gene regions exhibiting patterns of polymorphism that deviate from the neutral expectation. The description of the pattern of variation along each chromosome using sliding window approaches can reveal regions that have heterogeneous evolutionary histories, which can be particularly valuable in unannotated genomic regions. Genes associated with variation in complex traits often show population genetic signatures of historical natural selection (Robin et al., 2002; DeLuca et al., 2003; Carbone et al., 2006). Merging the inferences about evolutionary history obtained from the population genomics analyses with the inferences about genes affecting quantitative traits from the phenotypic analyses will provide the first large-scale answer to the long standing question of the balance of forces that maintain genetic variation for complex traits in nature. Molecular population genetic analyses of these data will be spearheaded by Drs. Philip Awadalla, Antonio Barbadilla and Ignazio Carbone (letters attached).

High Resolution Sequence Polymorphism Map: The *Drosophila* Genetic Reference Panel is a living library of all common polymorphisms affecting natural variation. The proposed whole genome sequence analysis of these 192 lines will identify a common set of dense polymorphic markers and allow an economic and accurate platform for genotyping Drosophila for any purpose. The 192 lines have a 95% probability to contain alleles with MAF 1.5%. It is important to realize that this will be a polymorphism discovery effort, and that the actual allele frequencies for alleles found in only a single line will have to be independently measured using an independent genotyping platform. The molecular polymorphism data will be curated in the *Drosophila* Polymorphism Data Base (DPDB, http://bioinformatica.uab.es/dpdb/dpdb.asp) by Dr. Antonio Barbadilla (letter attached).

A Test Bench for Novel Experimental and Statistical Methods: *Drosophila* has long been a testing ground for techniques that are later applied to human genetics and other animals. For example, the whole genome assembly in eukaryotes was first tested in *Drosophila* (Myers et al., 2000) before mammals. The central problem in human genetics today is the identification of genetic loci and specific alleles contributing to common disease. Association mapping studies in humans are expensive and some have produced false positives. The *Drosophila* Genetic Reference Panel will serve as a test bed for novel statistical and experimental approaches that seek to increase the accuracy of quantitative trait analysis in human health, as described in the multiple testing section above. It has the advantages of known alleles with described quantitative affects, the ability to replicate experimental results in independent laboratories, and facile experimental methods, as well as tractable genome size, allowing for minimal computation time. As a fair amount of phenotypic information is already available, this test bench will be available for use as soon as the sequencing is completed.

Sequencing Plan: (A) Platforms. New massively parallel sequencing technologies have brought projects of this size to an extremely reasonable cost and size (see question B4

below for cost details). However the different available technologies have different yet complementary error characteristics. Data from a pilot project (see Appendix 5 for details), test sequencing four of the proposed DGRP strains including the original BDGP reference strain (y^1; cn^1 bw^1 sp^1) suggested a mixed platform strategy.

The Illumina platform currently produces read lengths of 36bp, possibly stretching to 50bp with low quality tails. The primary error mode is substitutions, especially high near the end of the reads. The largest problem with short read data is the inherent difficulty in mapping short sequences to the reference genome in the presence of the very polymorphisms we are aiming to identify. This results in the analysis of a smaller proportion of the genome, and often lower coverage in regions that can be partially mapped, requiring additional sequence coverage.

The 454 pyrosequencing platform has longer reads – now 500bp on the XLR platform, but a different yet complementary error mode of homopolymer length issues. These longer reads are much easier to map to the genome in the presence of sequence differences from the reference sequence. Thus the longer reads have the advantage of allowing a larger proportion of the genome to be analyzed than the short reads. Additionally and importantly, the length of the new XLR reads allows facile *de novo* assembly. Comparison of assembled contigs to the reference sequence allows the identification of insertions and deletions longer than the length of a single sequence read, and is particularly useful for the characterization of larger insertions and inversions that otherwise prevent the alignment of single reads to the BDGP reference sequence.

The goal of the Drosophila Genetic Reference Panel is to identify as many polymorphisms in the genomes of the panel as possible, with a high degree of accuracy. Thus we are proposing to use both platforms, the 454 XLR long read pyrosequencing to identify larger (>3bp) insertion/deletion polymorphisms, and Illumina short read sequence to prevent the accumulation of false positives due to homopolymer errors, and to provide a genome scale verification of as many of the sequence changes as possible. This verification can be particularly important as a few of the bases in the genome, despite intensive inbreeding, can still be polymorphic within the strain, and within the DNA isolated from multiple individuals. Confirmation of such polymorphism on two different platforms allows verification of such cases and appropriate handling in downstream statistical analysis.

(B) Coverage considerations. Based on the pilot study, high quality consensus sequence coverage of the majority of the genome saturates at 10-12 fold 454 long read coverage (~4-5 XLR runs at the current time). For the Illumina short read platform a similar saturation profile occurs – with additional mapping problems reducing the overall coverage relative to the input coverage. We plan to use paired end sequencing for this project, which is now available on the Illumina platform; this will mitigate these mapping problems. The Illumina sequences will be primarily used for error correction and polymorphism validation, rather than discovery. Thus, we believe that 6 fold genome coverage is sufficient for the correction of homopolymer errors (and identification of true polymorphisms next to homopolymers).

Polymorphism Identification: In the pilot project we sequenced a few of the proposed DGRP strains and the original BDGP reference strain. From the DGRP strains we identified an average of ~660,000 polymorphisms of different types (Appendix 5). Whilst

methods may change and improve in the future, we are taking a two pronged approach to polymorphism identification. The main approach is read alignment to the BDGP reference sequence. We have successfully used an analysis pipeline based around the mosaik software package (Marth lab) to identify the small polymorphic features (<10 – 20bp). This package has the virtue of allowing alignment of both short and long read data sets simultaneously to the reference. Comparison to Sanger reads in one of the sequenced strains allowed us to estimate a SNP false positive rate of 0.008% on the pyrosequencing platform, after homopolymer error correction by the short read data. The false negative rate was 1.0%, due mainly to lack of coverage. When a sequence difference from the reference is detected by both sequencing platforms, the error probabilities are extremely low.

The second approach addresses longer polymorphic features, above 20bp in length. Here we assemble the long read XLR 454 data and align contigs to the BDGP reference. From a test assembly of 500bp XLR data we obtained N50 contig lengths of 26.7 kb and a longest contig of 267 kb. Additional paired end information produced a scaffold N50 of 3.3Mb (the largest scaffold was 17.3Mb). Alignment of these to the BDGP reference using simple blat alignment and careful parsing allowed facile and accurate identification of larger polymorphisms and their boundaries (Appendix 5). The combination of these approaches, reflecting the combination of sequencing platforms, allows a full characterization of the genomes.

A specific question in analyses such as these is the false positive rate, as a high false positive polymorphism discovery rate might confound statistical analyses. One particularly stringent test of this is to re-sequence the BDGP reference strain, which should be in large part similar to the reference sequence, but contain a small number of differences due to passage of the strain over the last 10 years. This is indeed what we observe, with 730 substitutions identified with high confidence compared to ~ 1,000,000 for the DGRP strains. Thus the proposed methods have demonstrated low error rates.

A Planned and Managed Analysis: The whitepaper authors believe a proactive approach to ensure timely analysis, public dissemination and publication is required. To this end, in addition to submitting all data to FlyBase, Genbank and all other appropriate public databases, we will provide rapid analysis of the QTL data already available, to provide a list of candidate quantitative trait sequence polymorphisms for the many quantitative traits already measured in these strains. We have enrolled a number of collaborators promising to perform additional analysis of traits on these lines (see multiple letters of support), and statistical experts (support letters from Drs. Doerge and McIntyre) to apply novel analyses to this unique dataset and kick start community involvement. A large number of the most promising QTLs identified will be followed up with complementation tests and other functional analyses (carried out by our collaborators). We intend to publish not just a description of sequence variation in *Drosophila* and its impact on population genetics, but also candidate polymorphisms affecting numerous traits already and promised to be measured, with many partially verified by the methods described above.

Finally, to fully leverage the use of this complete dataset, the sequence data, reference strains, all measured phenotypes and the statistical tools will be made publicly available.

The actual reference stocks will be independently maintained in the Bloomington Stock Center (letter from Kathy Matthews attached) and the Mackay laboratory, in duplicate mass cultures at both locations. Keeping the stocks in multiple locations guards against loss. Further, ensuring the stocks are maintained in mass cultures minimizes the impact of new spontaneous mutations. The lines will be checked for contamination annually using 20 polymorphic markers. Thus, *Drosophila* investigators can use these resources to quantify traits of interest in the strains, and use web based tools for analysis with association mapping tools of their choice, rapidly receiving candidate sequence polymorphisms for follow up with complementation tests, mapping or other analyses. With such tools, this dataset brings association mapping for quantitative traits to the entire *Drosophila* community.

Specific Points:

A. Specific Biological/Biomedical Rationales for the Utility of New Sequence Data:
A1. Improving Human Health: This project will provide candidate *D. melanogaster* quantitative trait polymorphisms affecting lifespan, alcohol tolerance, aggression, and many other traits directly related to human disorders and disease. It is likely that a proportion of the identified sequence polymorphisms will have orthologous effects in humans, suggesting new diagnostic tests and suggesting new pathways as targets for drug design. It is also likely that this project will help us better define the role of non-genetic effects in these traits, and better define where lifestyle changes will likely provide better health outcomes.
A2. Informing Human Biology: In the same way that the study of *D. melanogaster* mutants has connected genes and proteins to phenotypes that are often found to be similar in human biology, we expect this study of biological and genotypic variation in *D. melanogaster* to be of use for the study of human variation where there are similar pathways and processes.
A3. Expanding Our Understanding of Basic Processes Relevant to Human Health: Most genetic variation affecting traits relevant for human health is quantitative in nature. Single gene Mendelian variants, whilst easier to understand, affect a much smaller proportion of the population. As well as providing candidate genes for specific human traits where a similar trait can be measured in flies as described in A1, this project will also generate a large number of QTLs. The analysis of this set will enable investigations into critical points in genetic pathways and determine common biological idioms in the evolution of biological redundancy. In short, the *Drosophila* Genetic Reference Panel will provide the data to take understanding of quantitative traits to a medically relevant level of detail.
A4. Providing Additional Surrogate Systems for Human Experimentation: *D. melanogaster* is already a proven surrogate for many aspects of human genetics. This project will improve our ability to measure the genetic determinants of variation of these models in response to drugs and other treatments. In many cases the quantitative trait polymorphisms identified may be more relevant to human variation in response to drug treatments than genes identified by the less subtle effects of mutational screens.
A5. Facilitating the Ability to do Experiments: The project will directly facilitate association mapping of quantitative traits and traditional mapping of quantitative traits.

Furthermore, it will enable the entire *Drosophila* community to perform these experiments with no additional sequencing, only phenotyping will be required. Additionally it will generate a high resolution polymorphism map and allow high resolution low cost genotyping for population studies and other uses in *Drosophila*. Finally it will be used as a low cost test bench for novel statistical and experimental methods prior to their use in human and other organisms with large genomes.

B. Strategic Issues in Acquiring New Sequence Data:

B1. The Demand for New Sequence Data: The *Drosophila* community has a proven history of fully utilizing the excellent sequence resources it already has. In many ways it has been a model example of how genomic sequences can stimulate biological and medical research, and lead to other powerful high-throughput biological tools. Based on initial enthusiasm from the community (see attached letters) and the ease and low cost of performing association studies, once these sequences are available, we believe that community enthusiasm will be significant. Due to the large size of the *D. melanogaster* community, we do not expect any additional expansion of the community due to these sequences. In the larger biology community, there is clamoring for complete genotype and phenotype data sets to better understand the connection between genotype and phenotype. Thus, this dataset will also be used by many outside of the *Drosophila* community, for example bioinformaticians developing better quantitative trait and association mapping methods, population geneticists developing models of the inheritance of non-Mendelian traits, and as a basis data set for systems biology researchers. This has already been the case for the original *D. melanogaster* sequence that is widely used as a test bed for genome assembly and gene prediction software, as well as a platform for high throughput biological research.

B2. The Suitability of the Organism for Experimentation: *D. melanogaster* is a premier model organism for biological experimentation.

B3. The Rational for the Complete Sequence of the Organism: Alternatives to the whole genome association studies, and high resolution whole genome mapping studies, are to use a low resolution map for mapping specific traits and then to study in high resolution regions of interest for that particular phenotype. While this is suitable for a study of a single phenotype, it does not allow the study of many quantitative phenotypes, does not provide a useful community tool allowing the amortization of costs, requires significant financial and labor investment for the high resolution follow up, and does not fully take advantage of the low costs of newly available massively parallel sequencing technologies. We believe the proposed whole genome provides tools for both traditional and association mapping, makes these available to the entire *D. melanogaster* community applicable to any phenotype at a reasonable financial cost.

A related question is our rationale for the number of lines to be sequenced. As discussed above, we believe that the 192 Raleigh lines provide excellent power for association experiments to detect moderately small genetic effects, and yet is a small enough number to be tractable for the average *Drosophila* laboratory measuring phenotypes. Variation for most, if not all, important quantitative phenotypes will be observed in this number of lines, allowing the resource to be broadly applicable. With all the proposed lines, we have the power to detect minor allele frequencies of 0.015 (the probability of not observing a single allele with a population frequency of 0.015 is 0.055

in a sample of 192 alleles). Finally, 40 of the Raleigh lines have been assayed for transcriptional activity using Affymetrix arrays. Previously, samples of smaller size have identified molecular variants significantly associated with phenotypic variation (Lai et al., 1994; Long et al., 1998; Robin et al., 2002). Further, the range of variation embraced by these lines is similar to, and sometimes greater than, the variation seen in lines selected for specific phenotypes.

B4. The Cost of Sequencing the Genome and the State of Readiness of the Organisms DNA for Sequencing: Based on our experience with the pilot sequencing of a number of strains, and previous experience with the reference *D. melanogaster* sequence as one of the original members of the BDGP, *D. pseudoobscura* sequencing and a number of other insects, we foresee no challenges in terms of biological features that will hinder this project. DNA has been isolated from all 192 inbred NC strains and is ready for sequencing.

Due to the amazing wealth of new sequencing technologies evolving at this time, it is impossible to predict the magnitude of the decrease in sequencing costs over the time period of this project. However, at the current time, 10-12X long read 454 genome coverage (~2Gbp) will require four high density XLR platform runs (~ 500Mb/run) and an additional paired end run to provide assembly linking data. For the short-read coverage, one half of an Illumina GA2 paired-end run will be required. Over the period of this work, we expect the throughout of the HD-XLR platform to improve to a maximum of 1Gb/run, reducing the number of runs required by half, and the Illumina platform to have a similar increase in throughput reducing the requirement to half a run per strain. Without improvements, the six runs per strain required at $5,000 US each, add to a total of $30,000 per strain, or approximately five and three quarter million dollars for the entire project. Whilst the sequencing cost of the project could ultimately be reduced by half, the expected improvements will come over the one year time period in which the project will be completed, so the actual cost is likely to be approximately four million dollars.

5. Are There Other (Partial) Sources of Funding Available or Being Sought for This Sequencing Project?

The phenotyping work both ongoing and pledged in letters of support is being funded out of ongoing funding of the individual investigators involved. In total, this amount is comparable to the amount of funds requested for sequencing due to the labor costs of a large number of individual researchers. No other additional sources of funding for the sequencing are being sought at this time.

References:

Carbone, M. A., Jordan, K. W., Lyman, R. F., Harbison, S. T., Leips, J., DeLuca, M., Awadalla, P. and Mackay, T. F. C. (2006) Phenotypic variation and natural selection at *Catsup*, a pleiotropic quantitative trait gene in *Drosophila*. Curr. Biol. *16*, 912-919.

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. Genetics *138*, 963-971.

Clarke, L., and Carbon, J. (1976). A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. Cell *9*, 91-99.

De Luca, M., Roshina, N. V., Geiger-Thornsberry, G. L. Lyman, R. F., Pasyukova, E. G. and Mackay, T. F. C. (2003) *Dopa-decarboxylase* (*Ddc*) affects variation in Drosophila longevity. Nat. Genet. *34*, 429-433.

Dilda, C. L. and Mackay, T. F. C. (2002) The genetic architecture of *Drosophila* sensory bristle number. Genetics *162*, 1655-1674.

Doerge, R. W., and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. Genetics *142*, 285-294.

Edwards, A. C., Rollmann, S. M., Morgan, T. J. and Mackay, T. F. C. (2006) Quantitative genomics of aggressive behavior in *Drosophila melanogaster*. PloS Genetics DOI: 10.1371/journal.pgen.0020154

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edition (Pearson Education Group).

Jordan, K. W. and Mackay, T. F. C. (2006) Quantitative trait loci for locomotor behavior in *Drosophila melanogaster*. Genetics *174*, 271-284.

International-HapMap-Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Lai, C., Lyman, R. F., Long, A. D., Langley, C. H. and Mackay, T. F. C. (1994) Naturally occurring variation in bristle number and DNA polymorphisms at the *scabrous* locus in *Drosophila melanogaster*. Science *266*, 1697-1702.

Long, A. D., Lyman, R. F., Langley, C. H., and Mackay, T. F. C. (1998). Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics *149*, 999-1017.

Lyman, R. F., Lai, C. and Mackay, T. F. C. (1999) Linkage disequilibrium mapping of molecular polymorphisms at the *scabrous* locus associated with naturally occurring variation in bristle number in *Drosophila melanogaster*. Genet. Res. *74*, 303-311.

Mackay, T. F. C. and Langley, C. H. (1990) Molecular and phenotypic variation in the *achaete-scute* region of *Drosophila melanogaster*. Nature *348*, 64-66.

Moriyama, E.N., and Powell, J.R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. *13*, 261-277.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A*., et al.* (2000). A whole-genome assembly of Drosophila. Science *287*, 2196-2204.

Robertson, A. (1967) The nature of quantitative genetic variation. Pp.265-280 in A. Brink (ed.), *Heritage From Mendel* (University of Wisconsin Press).

Robin, C., Lyman, R. F., Long, A. D., Langley, C. H. and Mackay, T. F. C. (2002) *hairy*: A quantitative trait locus for Drosophila bristle number. Genetics *162*, 155-164.

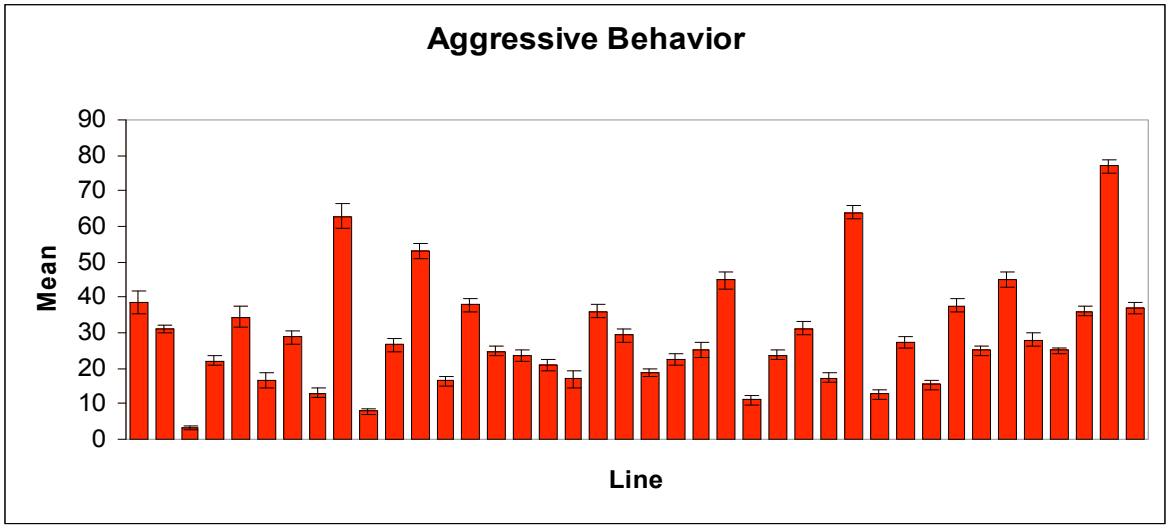
Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006). Positive natural selection in the human lineage. Science *312*, 1614-1620.

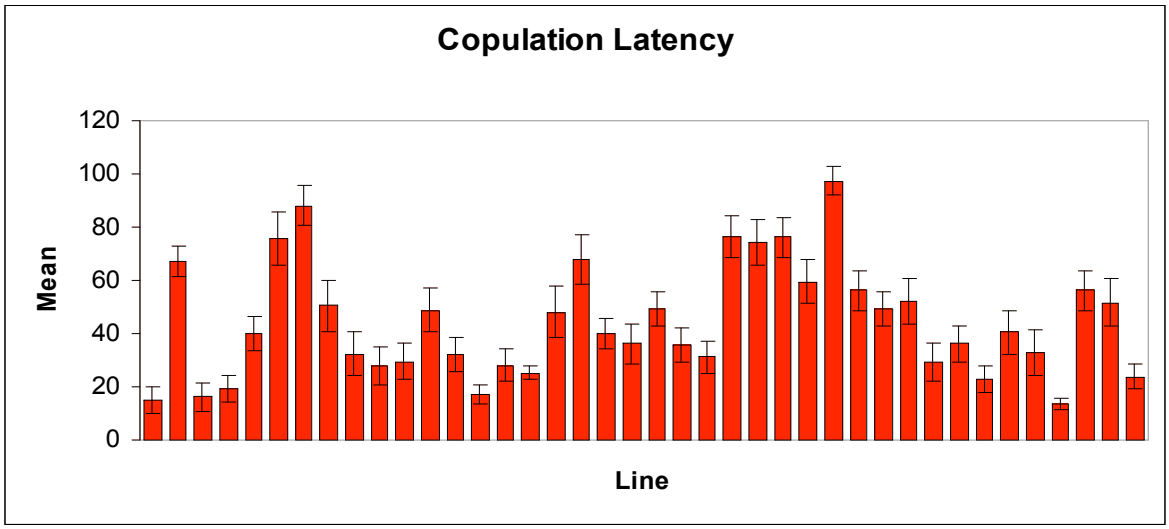
Sokal, R.R., and Rohlf, F.J. (1981). *Biometry*, 2nd edition (W. H. Freeman and Company).

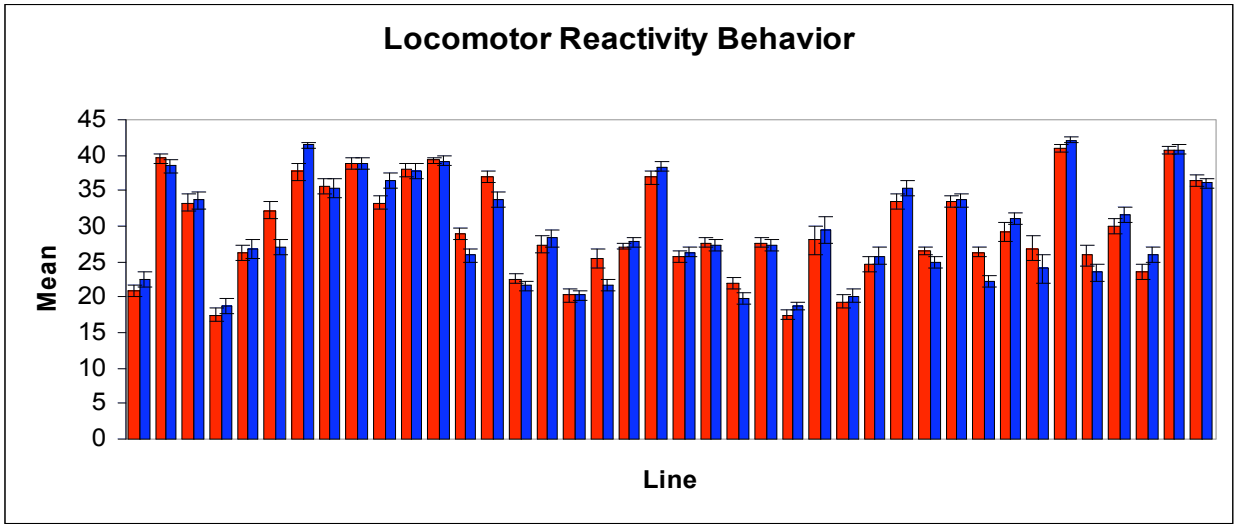
Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.

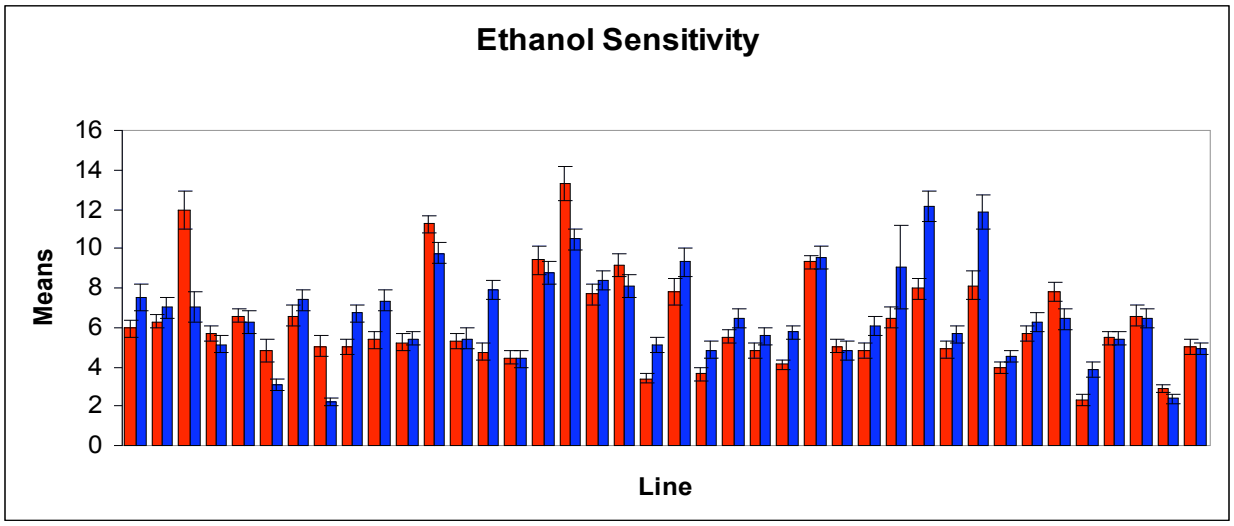
Appendix 1

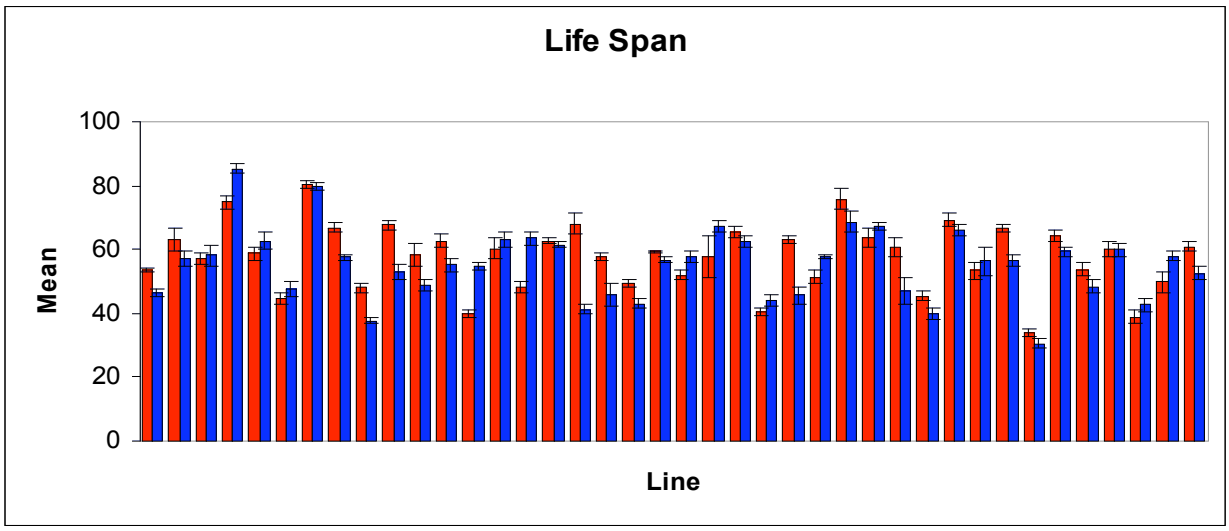
Quantitative Variation Among the Core Set of 40 Lines From the Drosophila Genetic Reference Panel

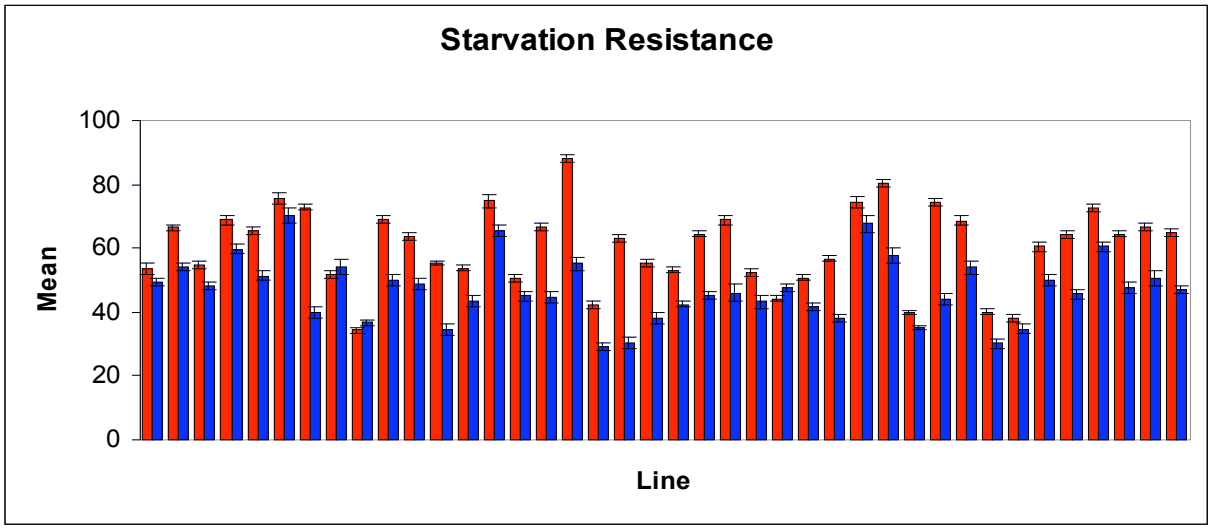


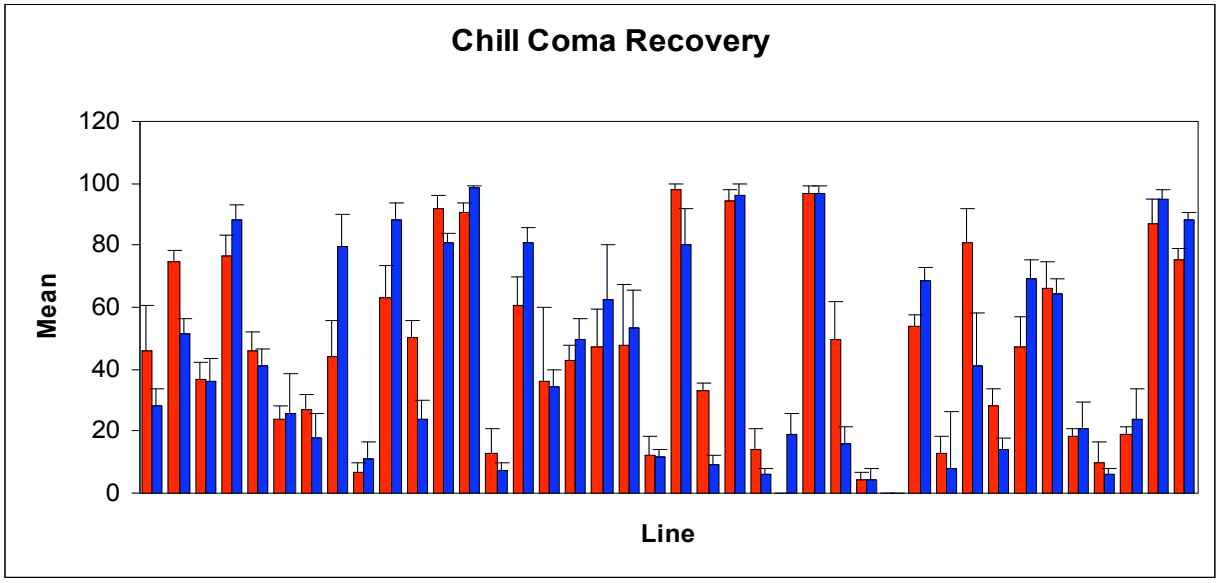


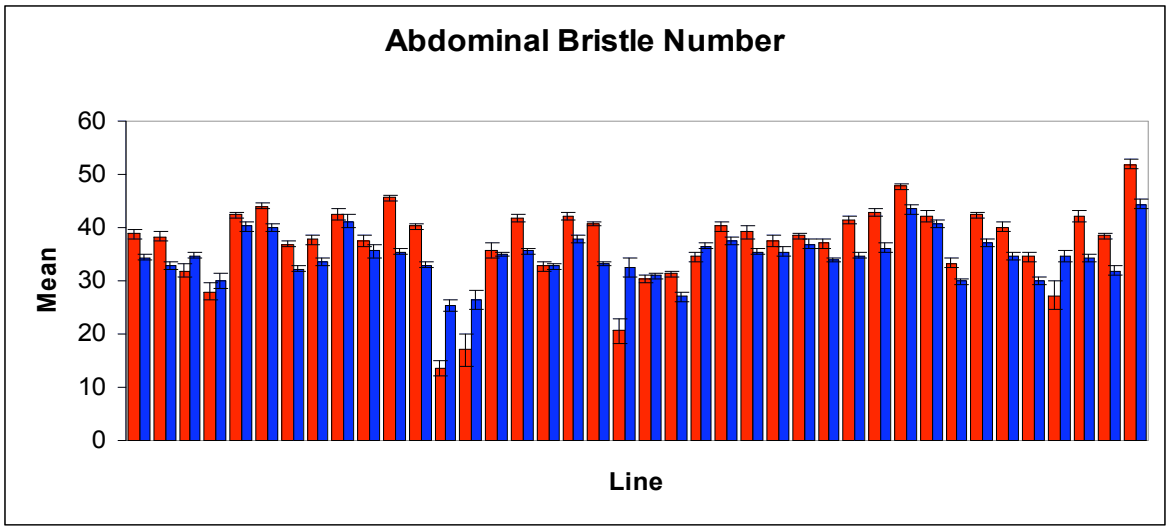

Locomotor Reactivity Behavior
Mean
45
40
35
30
25
20
15
10
5
0
Line

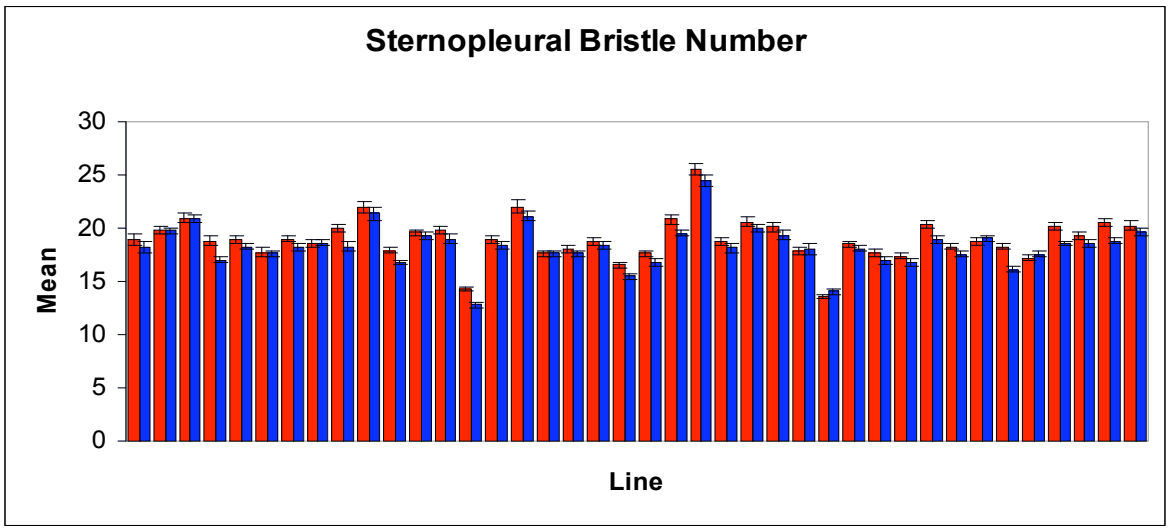

Ethanol Sensitivity
Means
16
14
12
10
8
6
4
2
0
Line


Life Span
100
80
60
40
20
0
Mean
Line


Starvation Resistance
100
80
60
40
20
0
Mean
Line


Chill Coma Recovery
120
100
80
60
40
20
0
Mean
Line


Abdominal Bristle Number
60
50
40
30
20
10
0
Mean
Line


Sternopleural Bristle Number
30
25
20
15
10
5
0
Mean
Line

Appendix 2

Quantitative Genetic Parameters Estimated From the Core Set of 40 Raleigh Inbred Lines

Trait [a]	Mean	σ_G^{2} [b]	σ_E^{2} [c]	σ_P^{2} [d]	H^{2} [e]	CV_G [f]	CV_E [g]
AG	29.22	235.801	64.845	300.646	0.784	52.552	27.559
LR	28.24	27.262	30.463	57.725	0.472	18.489	19.544
LS	54.06	102.126	89.468	191.594	0.533	18.693	17.496
SR	54.11	116.961	93.906	210.867	0.555	19.964	17.910
ER	6.54	3.869	13.674	17.543	0.221	30.148	56.516
CC	48.84	822.175	250.833	1073.008	0.766	58.714	32.430
CL	44.22	398.875	1186.83	1585.705	0.252	45.165	77.907
ST	18.65	3.645	2.313	5.958	0.612	10.237	8.155
AB	35.89	33.147	15.440	48.587	0.682	16.055	10.958
DH(x100)	7.88	19.248	76.040	95.288	0.202	55.648	110.605

[a] AG = aggressive behavior; LR = locomotor reactivity behavior; LS = life span; SR = starvation resistance; ER = ethanol resistance; CC = chill coma recovery; CL = copulation latency; ST = sternopleural bristle number; AB = abdominal bristle number; DH = developmental homeostasis of abdominal bristle number

[b] $\sigma_G^2 = \sigma_L^2 + \sigma_{SL}^2$

[c] σ_E^2 = variance within replicates

[d] $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$

[e] H^2 = broad sense heritability = σ_G^2/σ_P^2

[f] $CV_G = 100\sigma_G/\text{Mean}$

[g] $CV_E = 100\sigma_E/\text{Mean}$

Appendix 3

Power Calculations

A. Core Set of 40 Lines

			δ [c] (σ_P)		δ [c] (% Mean)	
Trait [a]	Units	σ_P [b]	$N = 20$	$N = 40$	$N = 20$	$N = 40$
AG	Number	17.34	3.97	2.81	13.6	9.62
LR	Seconds	7.60	1.74	1.23	6.16	4.36
LS	Days	13.84	3.17	2.24	5.86	4.14
SR	Hours	14.52	3.33	2.35	6.15	4.34
ER	Minutes	4.19	0.96	0.68	14.68	10.40
CC	Percent	32.76	7.50	5.31	15.36	10.87
CL	Minutes	39.82	9.12	6.45	20.62	14.59
ST	Number	2.44	0.56	0.40	3.00	2.14
AB	Number	6.97	1.60	1.13	4.46	3.15

B. Entire Genetic Reference Panel (192 Lines)

			δ [c] (σ_P)		δ [c] (% Mean)	
Trait [a]	Units	σ_P [b]	$N = 20$	$N = 40$	$N = 20$	$N = 40$
AG	Number	17.34	1.82	1.28	6.23	4.38
LR	Seconds	7.60	0.80	0.56	2.83	1.98
LS	Days	13.84	1.45	1.02	2.68	1.89
SR	Hours	14.52	1.52	1.07	2.81	1.98
ER	Minutes	4.19	0.44	0.31	6.73	4.74
CC	Percent	32.76	3.44	2.42	7.04	4.95
CL	Minutes	39.82	4.18	2.95	9.45	6.67
ST	Number	2.44	0.26	0.18	1.39	0.97
AB	Number	6.97	0.73	0.52	2.03	1.45

The power to detect an association for a marker causally affecting the trait at a frequency of $q = 0.5$ with 40 inbred lines, for $N = 20$ individuals and $N = 40$ individuals measured per line. Effects are shown in phenotypic standard deviation units, and as percent of the overall trait means, for complex traits that have been scored on the 40 lines to date.

[a] AG = aggressive behavior; LR = locomotor reactivity behavior; LS = life span; SR = starvation resistance; ER = ethanol resistance; CC = chill coma recovery; CL = copulation latency; ST = sternopleural bristle number; AB = abdominal bristle number

[b] Phenotypic standard deviation

[c] δ = difference in mean between homozygous markers

Appendix 5
Pilot Project: Results from sequencing four inbred *Drosophila melanogaster* strains.

Summary
As a pilot project for the Drosophila Genetic Reference Panel (DGRP), we have sequenced three DGRP lines, and additionally the BDGP original reference strain. One line (DGRP 360) was sequenced at multiple depths on two platforms (454 and Illumina) to assess sequencing strategies for the rest of the project. The main conclusions are:

1. High genome sequencing coverage (12X) is required to get high quality sequence for the majority of bases in the genome. After 12X coverage, additional sequencing brings diminished returns.
2. Short read sequences (36bp) map poorly to the *D. melanogaster* reference sequence, resulting in the analysis of a lesser percentage of the genome. Less mapped coverage per unit input sequence coverage also reduces the consensus quality of regions of the genome that are sequenced. Additionally, the short reads have a relatively high substitution error rate.
3. 11% of SNPs were missed with short reads. Of these, 98% were due to the clustering of polymorphisms preventing alignment. To avoid false alignment, we do not allow alignments with greater than 10% mis-alignment for a 36bp read. Mild clusters of SNPs and other polymorphisms within a 36bp region prevent alignment and analysis. The relatively high polymorphism rate in *Drosophila* exacerbates the problems aligning short reads. Clusters of polymorphisms may have a greater effect on gene function or expression than isolated SNPs. Paired end sequencing and longer read lengths will resolve this problem in the future.
4. Homopolymer errors in long pyrosequencing reads require correction, which is best done on a genome scale with the short read technologies. Such verification is especially important for bases that have remained polymorphic within the inbred strain, to avoid the possibility of sequence error.
5. For larger polymorphisms – insertions, deletions, inversions, long substitutions – *de novo* assembly of long read sequence data is mandatory. For insertions longer than 50-100bp (not represented in the reference sequence) alignment of assembled sequence is the ideal method, and allows accurate delineation of junction sites.
6. A combination of 12X XLR 454 long read coverage and 12X Illumina short read coverage is required. Additionally, a combination of analysis methods – read alignment to both a reference, and a *de novo* assembly, followed by alignment of assembled contigs to the reference – is required for full polymorphism analysis. The short reads need to be aligned to the *de novo* assembly for full utility.

Introduction

We previously submitted a white paper entitled "Proposal to Sequence a *Drosophila* Genetic Reference Panel: A Community Resource for the Study of Genotypic and Phenotypic Variation." where we proposed to sequence 192 *D. melanogaster* strains with extensive quantitative phenotypic data as a community resource for association studies and quantitative trait mapping. This white paper was received enthusiastically, but

questions of methodology were brought up, and it was recommended that a pilot project determine the ideal sequence coverage and technology platform for the project. We sequenced a single strain to high coverage using multiple methodologies and used this comparison to determine the ideal sequencing strategy, and then followed this with sequencing additional strains. The BDGP reference strain (y^1; cn^1 bw^1 sp^1) was included at the request of the NHGRI coordinating committee. This is a particularly interesting choice: The strain was originally isogenized for P1 library construction by the BDGP but had been reared for some time prior to DNA isolation for the BAC and WGS libraries used to generate the majority of *D. melanogaster* reference sequence. Despite the ten years since its creation, the strain is an excellent test for false positives in sequencing technologies, and the data is useful for many researchers using the strain for the Drosophila ENCODE project. However, it is not as relevant for the core task of any re-sequencing project – identifying differences from the reference. For this task the proposed DGRP inbred lines are more informative.

Here we present the results of this demonstration project, and based upon these results we have re-submitted a revised white paper with an updated sequencing plan.

Results

Table 1: Sequence performed and polymorphic bases identified by Mosaik alignment to the 5.1 *D. melanogaster* reference sequence. (* - see discussion on insertions and deletions below)

Line	Sequence	Polymorphic bases*			SNP rate	Total
		Substitutions	Insertions*	Deletions*		
DGRP 360	12X 454	549,388	91,497	23,075	I in 217	663,960
DGRP 360	12X Illumina	436,894	12,236	8,575	1 in 272	457,705
DGRP 375	12X XLR 454	509,682	96,282	22,442	1 in 234	628,406
DGRP 825	12X 454	579,604	93,022	24,324	1 in 205	696,950
BDGP reference strain	12X 454	730	260	279	1 in 163,054	1,269

Input sequence coverage and aligned sequence coverage. The sequence generated from both platforms was partitioned into bins representing 2, 4, 6, 8, 10 and 12 fold genome coverage, aligned to the genome using Mosaik (at the time of analysis this was the only software platform that could align Illumina, 454 and Sanger reads taking account of indels – thus allowing a single software platform to be used with all three data types). Multiple alignment parameters were tried to obtain optimal alignments, however we believe there is room for improvement in this area, and we continue to test new alignment software as it becomes available.

First we plotted the observed base coverage in alignments in comparison with the input sequence. As can be seen in Fig. 3, the modal read coverage tracks well with the input coverage for the longer reads, but significantly less sequence is aligned for the short reads (in line with theoretical considerations). To assess whether this is a software issue

we repeated the 12X Illumina data alignment with Eland, alignment software provided by Illumina. Although slightly more reads were aligned, there was also a higher false positive alignment rate, and the software does not currently allow indel alignment. In conclusion, 36bp reads are difficult to align to complex eukaryotic genomes. The Illumina paired end kit is currently being deployed (April 2008), and we hope that alignment of paired end reads will reduce this problem.

Longer read lengths align to, and assay a larger portion of the Drosophila genome. More important than coverage is the proportion of the genome that can be sequenced with a given read length. Whilst portions of *D. melanogaster* centromeres, telomeres and other heterochromatic regions are highly repetitive and resistant to current assembly techniques even with 800bp Sanger sequence reads, approximately 30Mb of the *D. melanogaster* reference genome is available in heterochromatin sequence files which added to ~ 110Mb of euchromatin gives a reference sequence length of ~ 140Mb for an estimated total genome size of ~180Mb. We asked what proportion of the genome could be accessed by the different sequencing platforms. As can be seen in Fig. 3, the longer read lengths allow approximately 20Mb of additional *D. melanogaster* sequence to be assayed – or ~11% of the 180Mb genome. The higher quality of the sequenced bases is a function of the increased read coverage – i.e. because of the easier mapping of the longer reads – more reads cover each base, and thus the consensus call quality is higher.

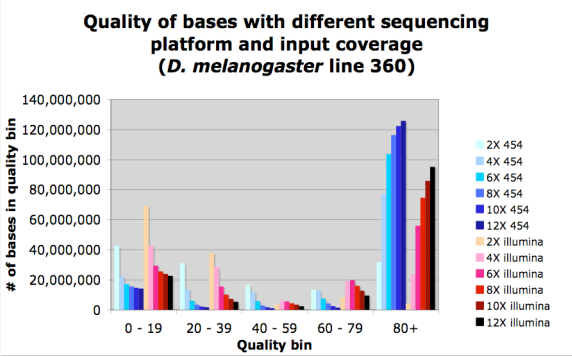


Fig. 3. Total *D. mel* reference bases covered at different aligned consensus qualities. Qualities manufacturers values summed in the alignment with a maximum value of 90.

Increased coverage does not lead to increased error rate. Is it possible that the increase number of sequence reads aligning to the reference sequence increases the number of possibilities that errors may accumulate? Whilst this is clearly true for any individual platform we wondered if the relative increase in aligned sequence coverage between the two platforms would translate into a relative increase in reference bases where more than a single base was called. The number of different bases (a maximum of 6 – ACGTN and *(insertions)) called at every position was tabulated and is presented in Fig 4. As can be seen the number of reference bases with a single base call in the aligned reads is higher for the long read platform, and zero or two or more calls less for the long read platform. We believe this is because of the high error rate at the end of the short reads (the reads are short because the signal noise ratio is poor at the end of the read preventing further cycles of base addition increasing the read length), whereas for the 454 pyrosequencing the main cause of error is homopolymer length determination.

Comparison to a 1Mb Sanger "gold standard" sequence set. Sanger sequence is the gold standard of sequencing as long high quality reads with well defined quality scores allow accurate estimation of final assembled consensus sequence quality. 1Mb of "gold standard" sequence was produced with the following characteristics. (1) Read alignments were filtered for unique read alignments only – this sequence set should be eminently sequencible using short read technologies. (2) A minimum Phred quality score of 40 (note: we used the *Phred* base caller due to its higher accuracy compared to the *AB KB basecaller* software). Because this data is comprised of single reads aligned to the reference, a small number of errors were expected and found. These were removed from the analysis where both 454 and Illumina sequencing disagreed with the Sanger sequence. The error rate of the filtered single read coverage Sanger data set was approximately 1 in 18,500 in line with the minimum Phred score of 40 (1 in 10,000).

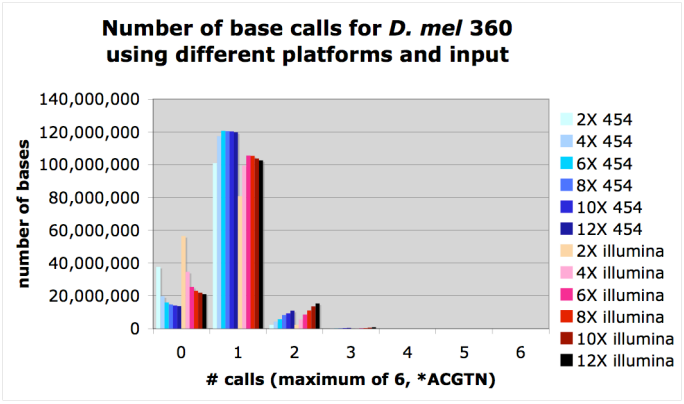


Fig. 4. Number of base calls for each *D. melanogaster* reference base. Red: Illumina, Blue: 454.

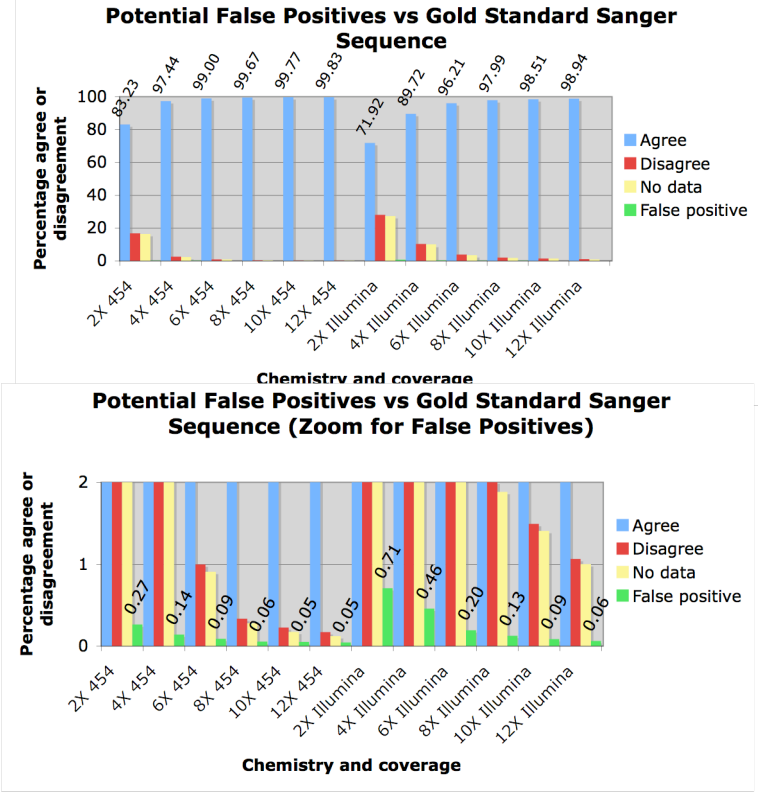


Fig. 5. Comparison of DGRP line 360 sequence to Sanger sequence in agreement with the *D. melanogaster* 5.1 reference sequence. Disagreements are potential false positives. **Top**: Entire Y scale showing up to 99.83% (12X 454) and 98.94% (12X Illumina) agreement with the reference sequence. **Bottom**: Zoom in on disagreements below 2%. Numbers are for the potential false positive rate, although most of these would fail low quality filters for identification of a polymorphism.

Fig. 5. Shows the agreement of different coverage levels of the two platforms with the 1Mb of Sanger sequence that agreed with the *D. mel* 5.1 reference. Although both platforms show good agreement, the additional aligned coverage of the long read dataset leads to almost 1% of additional agreement, and at higher consensus qualities. In the bottom part of Fig. 5,

failures to agree with the gold standard which would manifest as false positives are presented. The major source of disagreement in this case is no data, which would not be seen as false positives. Of the remaining differences, approximately half are due to homopolymers in the case of 454, and substitutions in low coverage regions in the case of Illumina. Another cause of error was poor alignments to the reference in the presence of an insertion. In the absence of the correct data the alignment program often makes mistakes around insertion breakpoints. In almost all cases, the difference to the reference sequence would not pass quality filters to be identified as a polymorphism. In addition, putative polymorphisms filtered out of 454 alignments due to proximity to a homopolymer greater than 5bp in length can be corrected by the complementary short read data.

Polymorphisms: Substitutions. The Sanger gold standard sequence set had 5,364 substitutions – 0.52% or ~ 1 in 200bp of the gold standard sequence, in line with previously published estimates. The longer reads enable a maximum of 98.6% of these to be identified, compared with 86.9% with the short reads. The discrepancy between the low false positive rate of the short reads, and the high false negative rate here is due to alignment problems in the presence of differences to the reference sequence. Specifically, the low number of differences from the reference that can be tolerated by the alignment program is a maximum of 3 substitutions, or a 3 bp indel. Manual inspection of 100 cases of failure to detect a SNP (Table 2) in the Sanger gold standard set due to lack of sequence coverage revealed that in 98 cases other polymorphisms nearby within a window of 36bp surrounding the gold standard SNP pushed the alignment outside the Mosaik alignment parameters. An example is shown in Fig. 7 at the end of this document. Because the SNP rate in *Drosophila* is ~ 1 in 200, with additional insertions and deletions, it is possible for multiple polymorphisms to cluster within a 36bp region. Additionally, an error rate of 2% over the 36bp of the Illumina read (equivalent to ~1% in the statistics reported by Illumina software comparing the first 25bp to a reference sequence, due to the clustering of substitution errors at the end of the read) suggests that 16% of reads will have a single error. In regions of already low coverage only 3

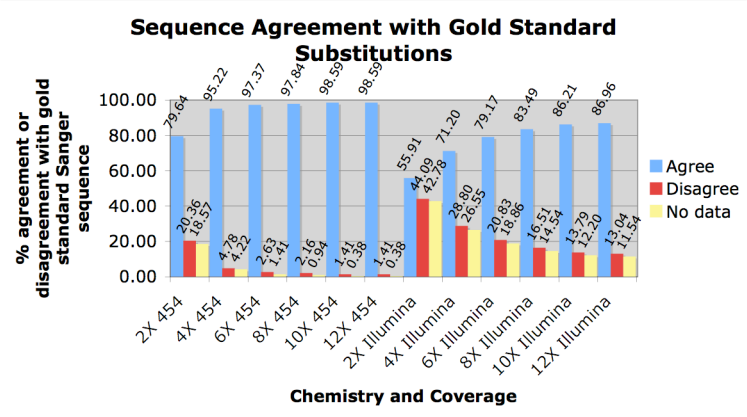


Fig. 6. The Gold Standard sequence contains 5,364 substitutions. Blue bars: percentage of substitutions found by different sequence coverage and platforms included both high and low quality agreements. Red bars: Disagreements with Gold Standard reference substitutions. Yellow: Disagreement due to no sequence coverage of the substitution. * Note that the vast majority of false negatives are due to no sequence coverage. Also the numbers in this graph have improved since an earlier version in Nov 2007 due to the removal of badly aligned Sanger reads in the gold standard set.

substitutions, and some random bad luck (to be expected with ~ 60 million attempted sequence alignments to the genome sequence) may be enough to preclude sequence alignment. Finally, one should note that these alignment issues affect clustered SNPs, and so the numbers of SNPs affected look worse than the numbers of read alignments affected. We fully expect that the latest update in the Illumina chemistry with the improved quality and paired end reads, as well as the possibility of 50bp reads, will greatly reduce these problems when compared to the data generated in the fall of 2007. However, the higher polymorphism rate of *Drosophila* compared with human sequences suggests that longer reads are even more important for the proposed project, and the clusters of polymorphisms may have a greater effect on gene function or expression than isolated SNPs.

Table 2. Categories and sub categories of clustered polymorphisms causing alignment issues for short read SNP false negatives.

No-aligned sequence, false negative category	Number
Total manually inspected	100
polymorphism in 36bp window (3 or more bp affected)	98
2 snps in window (+ bad luck?)	2
clustered subs (including single bp indels)	46
larger indels (>1bp)	25
indels + subs	27

Insertions and Deletions. On both short and long (250bp) read platforms, the ability to identify insertions and deletions is hampered by read alignment strategies. Our read alignment strategy (Mosaik) was used with fairly stringent parameters (95% identity) to prevent false positive alignments. For example, if >5% of a 250bp sequence read was not aligned, the alignment was considered potentially bad. As a result, indels greater than 10bp were not detected. Table 3 shows the maximum sizes of insertions and deletions identified with the Mosaik read alignment pipeline. We also have experience aligning 250bp pyrosequencing reads with Atlas SNP – a pipeline based upon BLAT mapping followed by Smith-Waterman alignment using CROSSMATCH. AtlasSNP identified deletions up to 40kb and insertions up to 10% of the read length when analyzing the genome of James Watson.

***De novo* assembly is required for accurate characterization of insertions and deletions.** In both cases it is clear that for the accurate identification and characterization of inserted sequences, *de novo* assembly is an absolute requirement. To this end, we have taken advantage of longer pyrosequencing reads, improvements in the Newbler assembler (454 Inc.) which can now handle insect sized genomes, and optionally the newly available protocols for 454 paired end sequencing with ~20kb insert sizes. DGRP line 375 was sequenced using the new XLR platform, and assembled in collaboration with 454. Note that the Newbler assembler is already used for the automatic assembly of bacteria genomes for the human microbiome project, and as such is already automated for the 200 assembles required for the proposed project. Table 4 Shows the assembly

statistics for the 12X 500bp read sequencing of DGRP line 375. Whilst analysis of the assembly is progressing and the other strains are in line for assembly, it is clear that this approach yields a full analysis of insertions and deletions. Figure 8 at the end of this document shows some examples of large insertions and deletions that can be easily identified. There are many of these in the *Drosophila* genome. We are currently using a modification of the Atlas SNP procedure of BLAT followed by CROSSMATCH and custom parsing to fully analyze this data. The examples shown however are straight-forward identifications of pure insertions and deletions. The *de novo* assembly also allows characterization and sequencing of far more complex combinations of insertions, deletions and large-scale substitutions, which are impossible to analyze with read alignment methods.

Table 3: Detection of insertions and deletions by the Mosaik reference pipeline with different read lengths.

	454 12X DGRP line 360		Illumina 12X DGRP line 360		1Mb Sanger - Line 360		12X 454 BDGP reference strain	
Indel Size	insertions	deletions	insertions	deletions	insertions	deletions	insertions	deletions
1	54793	12676	10497	8044	1615	714	551	540
2	26276	5330	2108	371	642	174	49	30
3	18651	1292	483	59	387	88	26	10
4	14543	1		25	289	49	14	
5	11238				215	24	7	
6	9372				163	35	6	
7	7158				129	11	6	
8	5822				102	21	7	
9	4590				87	15	9	
10						11		
11						14		
12						14		
13						8		
14						8		
15						5		
16						5		
17						7		
18						5		
19						2		
20						3		
21						4		
22						2		
23						2		
25						3		
27						2		
28						2		
32						1		
33						2		
34						1		
42						1		
50						1		
57						1		
60						1		
85						1		
Read Length	250 bp reads		36 bp reads		750 bp reads		250 bp reads	

Conclusions and Revised Sequencing Plan for the Drosophila Genetic Reference Panel.

This pilot project, performed at the request of the NHGRI sequencing committee, has demonstrated that the originally proposed sequencing plan of 5-6X short read genome coverage is not adequate for the proposed task. Surprisingly, the 36bp reads fail to identify a large number of SNPs where they are clustered, and also fail to analyze larger insertions and deletions, due solely to alignment problems to unknown sequences. Where the sequence data is very similar to the reference, as in the case of the BDGP reference strain, the data look quite good, but the true test is the identification of polymorphism. It is hoped that the advent of longer Illumina reads and paired end sequencing which have become available since this pilot project was completed will greatly resolve these problems. We hope to test these new methodologies as soon as possible.

Statistics for the de-novo assembly of DGRP line 375	
Contig #	12,314
total contig size	116.4Mb
contig N50	26.7kb
largest contig	267
Scaffold #	2,308
total scaffold size	127.7 Mb
Scaffold N50	3.3 Mb
largest scaffold	17.3Mb

Table 4. Newbler assembly of DGRP line 375 using 500bp reads and paired end data with insert sizes of 3kb and 20kb.

Long pyrosequence reads also have homopolymer issues which have to be corrected by the different error profile of the short read technology; however, they cannot be aligned to the reference sequence for a comprehensive analysis. Instead *de novo* assembly is absolutely required, and short read correction of possible homopolymer errors must be done by alignment to the *de novo* assembled sequence to allow the comprehensive alignment as seen against the BDGP reference strain. Finally, all assembled sequence projects must be compared to the reference to identify polymorphisms. 12X (2.1Gb) sequence coverage is required to allow analysis of the vast majority of the genome at high consensus quality. A single paired end run is also required, although we do not at this time believe that the larger insert sized paired end libraries will be required, as the shorter scaffolds will be long enough to identify the majority of larger insertions and deletions in the genome.

Because the 454 HD-XLR platform currently delivers ~ 500Mb of sequence in 500bp read lengths we expect this project will use per strain, 4 runs of the HD-XLR platform, 1 run of the Illumina platform, and 1 5kb insert paired end library. The costs of all these runs are approximately $5,000 each, and so a total cost of $30,000 per DGRP line is calculated at the current time. Whilst we do not expect the cost per run to decrease, it is likely that the yield per run will increase over the time of the project. We thank the NHGRI and the sequencing committee for the funding of this pilot project.

Fig. 7. An example of clustered polymorphisms preventing alignment of 36bp reads to a known SNP
Note: only the 2 center SNPs were not identified.

Illumina 36bp read 12X alignment – 40 bp of alignment at position 14,995,143 on chromosome 2L.
Yellow highlight = no aligned sequence due to polymorphisms in DGRP line 360.

Alignment	----*----		----A----		----C----		----G-----		----T----		----N----		consensus	consensus	ref	ref
Position	#calls	qual	#calls	qual	#calls	qual	#calls	qual	#calls	qual	#calls	qual	cal	qual	base	position
15040586	0	0	2	47	0	0	0	0	0	0	0	0	A	47	G	14995143.0
15040587	0	0	0	0	0	0	2	47	0	0	0	0	G	47	G	14995144.0
15040588	0	0	0	0	0	0	0	0	2	47	0	0	T	47	T	14995145.0
15040589	0	0	2	47	0	0	0	0	0	0	0	0	A	47	A	14995146.0
15040590	0	0	1	24	0	0	0	0	0	0	0	0	A	24	A	14995147.0
15040591	0	0	0	0	1	19	0	0	0	0	0	0	C	19	C	14995148.0
15040592	0	0	1	24	0	0	0	0	0	0	0	0	A	24	A	14995149.0
15040593	0	0	0	0	0	0	0	0	1	24	0	0	T	24	T	14995150.0
15040594	0	0	1	14	0	0	0	0	0	0	0	0	A	14	A	14995151.0
15040595	0	0	1	19	0	0	0	0	0	0	0	0	A	19	A	14995152.0
15040596	0	0	0	0	0	0	0	0	1	24	0	0	T	24	T	14995153.0
15040597	0	0	1	9	0	0	0	0	0	0	0	0	A	9	A	14995154.0
15040598	0	0	1	14	0	0	0	0	0	0	0	0	A	14	A	14995155.0
15040599	0	0	0	0	0	0	0	0	1	17	0	0	T	17	T	14995156.0
15040600	0	0	0	0	1	6	0	0	0	0	0	0	C	6	C	14995157.0
15040601	0	0	0	0	1	6	0	0	0	0	0	0	C	6	C	14995158.0
15040602	0	0	0	0	0	0	0	0	1	13	0	0	T	13	T	14995159.0
15040603	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995160.0
15040604	0	0	0	0	0	0	0	0	0	0	0	0	N	0	A	14995161.0
15040605	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995162.0
15040606	0	0	0	0	0	0	0	0	0	0	0	0	N	0	A	14995163.0
15040607	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995164.0
15040608	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995165.0
15040609	0	0	0	0	0	0	0	0	0	0	0	0	N	0	G	14995166.0
15040610	0	0	0	0	0	0	0	0	0	0	0	0	N	0	A	14995167.0
15040611	0	0	0	0	0	0	0	0	0	0	0	0	N	0	C	14995168.0
15040612	0	0	0	0	0	0	0	0	0	0	0	0	N	0	G	14995169.0
15040613	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995170.0
15040614	0	0	0	0	0	0	0	0	0	0	0	0	N	0	A	14995171.0
15040615	0	0	0	0	0	0	0	0	0	0	0	0	N	0	T	14995172.0
15040616	0	0	0	0	0	0	0	0	0	0	0	0	N	0	G	14995173.0
15040617	0	0	0	0	0	0	0	0	0	0	0	0	N	0	A	14995174.0
15040618	0	0	0	0	0	0	0	0	1	23	0	0	T	23	A	14995175.0
15040619	0	0	0	0	1	23	0	0	0	0	0	0	C	23	C	14995176.0
15040620	0	0	1	23	0	0	0	0	0	0	0	0	A	23	A	14995177.0
15040621	0	0	0	0	0	0	0	0	1	23	0	0	T	23	T	14995178.0
15040622	0	0	0	0	0	0	0	0	1	23	0	0	T	23	T	14995179.0
15040623	0	0	0	0	0	0	1	23	0	0	0	0	G	23	G	14995180.0
15040624	0	0	0	0	1	23	0	0	0	0	0	0	C	23	C	14995181.0
15040625	0	0	1	23	0	0	0	0	0	0	0	0	A	23	A	14995182.0
15040626	0	0	1	23	0	0	0	0	0	0	0	0	A	23	A	14995183.0

Filtered alignment of 1Mb Sanger reads to same region.
Green highlight = substitutions making alignment of 36bp read unlikely in this region.

Alignment	----*----		----A----		----C----		----G-----		----T----		----N----		consensus	consensus	ref	ref
Position	#calls	qual	#calls	qual	#calls	qual	#calls	qual	#calls	qual	#calls	qual	cal	qual	base	position
14995896	0	0	1	61	0	0	0	0	0	0	0	0	A	61	G	14995143.0
14995897	0	0	0	0	0	0	1	61	0	0	0	0	G	61	G	14995144.0
14995898	0	0	0	0	0	0	0	0	1	57	0	0	T	57	T	14995145.0
14995899	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995146.0
14995900	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995147.0
14995901	0	0	0	0	1	61	0	0	0	0	0	0	C	61	C	14995148.0
14995902	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995149.0
14995903	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995150.0
14995904	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995151.0
14995905	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995152.0
14995906	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995153.0
14995907	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995154.0
14995908	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995155.0
14995909	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995156.0
14995910	0	0	0	0	1	61	0	0	0	0	0	0	C	61	C	14995157.0
14995911	0	0	0	0	1	61	0	0	0	0	0	0	C	61	C	14995158.0
14995912	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995159.0
14995913	0	0	1	61	0	0	0	0	0	0	0	0	A	61	T	14995160.0
14995914	0	0	1	57	0	0	0	0	0	0	0	0	A	57	A	14995161.0
14995915	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995162.0
14995916	0	0	1	55	0	0	0	0	0	0	0	0	A	55	A	14995163.0
14995917	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995164.0
14995918	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995165.0
14995919	0	0	1	52	0	0	0	0	0	0	0	0	A	52	G	14995166.0
14995920	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995167.0
14995921	0	0	0	0	1	61	0	0	0	0	0	0	C	61	C	14995168.0
14995922	0	0	0	0	0	0	1	44	0	0	0	0	G	44	G	14995169.0
14995923	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995170.0
14995924	0	0	1	51	0	0	0	0	0	0	0	0	A	51	A	14995171.0
14995925	0	0	0	0	0	0	0	0	1	51	0	0	T	51	T	14995172.0
14995926	0	0	0	0	0	0	1	61	0	0	0	0	G	61	G	14995173.0
14995927	0	0	1	33	0	0	0	0	0	0	0	0	A	33	A	14995174.0
14995928	0	0	0	0	0	0	0	0	1	61	0	0	T	61	A	14995175.0
14995929	0	0	0	0	1	52	0	0	0	0	0	0	C	52	C	14995176.0
14995930	0	0	1	55	0	0	0	0	0	0	0	0	A	55	A	14995177.0
14995931	0	0	0	0	0	0	0	0	1	55	0	0	T	55	T	14995178.0
14995932	0	0	0	0	0	0	0	0	1	61	0	0	T	61	T	14995179.0
14995933	0	0	0	0	0	0	1	51	0	0	0	0	G	51	G	14995180.0
14995934	0	0	0	0	1	61	0	0	0	0	0	0	C	61	C	14995181.0
14995935	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995182.0
14995936	0	0	1	61	0	0	0	0	0	0	0	0	A	61	A	14995183.0

Fig. 8. Example insertions and deletions comparing the de-novo assembled DGRP line 375 to the *D. melanogaster* 5.1 reference sequence.

Example 1: 1,414 bp chromosome X deletion in line 375 relative to D.mel 5.1 ref

```
Query: 4681     accagaatttgtaaattaattaaaaagcgcttgttgtttccttttcttttttttgggtgga 4740
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10444706 accagaatttgtaaattaattaaaaagcgcttgttgtttccttttcttttttttgggtgga 10444765

Query: 4741     aaatgtccgaaaattgtttcgaggctcccggaggccattgatggggtcaagttattggca 4800
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10444766 aaatgtccgaaaattgtttcgaggctcccggaggccattgatggggtcaagttattggca 10444825

Query: 4801     gggtcgagatcccaaaccgccaaccaaaaggaactggccatttttaaaatttttacgact 4860
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10444826 gggtcgagatcccaaaccgccaaccaaaaggaactggccatttttaaaatttttacgact 10444885

Query: 4861     ttcgaagcgaagcaagtgcggaaaagtgaaaaaatgtgccaatgggaatg 4910
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10444886 ttcgaagcgaagcaagtgcggaaaagtgaaaaaatgtgccaatgggaatggttgggacca 10444945

Sequence deleted in Strain 375, but present in D.mel 5.1 reference sequence
Sbjct: 10444946 gagaactgcaagggtggcacttttttaccactcgactcacaccctacaattttgtgtgcg 10445005
Sbjct: 10445006 ggtgctactcgccacgcacatcgcgggtacttacaaacacacagtataaatctgaacatg 10445065
Sbjct: 10445066 cagacaagacaccccgttgtgtgcgcacccgaatcaatacggtgttttgcgtcgcgggtg 10445125
Sbjct: 10445126 ccgctcacacagtgcctaaaaagggatgagtgagaaaaacacttgtgggtataccgttaa 10445185
Sbjct: 10445186 acacatgggtgtttccaaaaatactcgggtgtttccaaaaatactcgagtggtctcgtag 10445245
Sbjct: 10445246 gtagtcgagtcaaatggcgccatacataatgattgttgagttcttgtgtctttggtccag 10445305
Sbjct: 10445306 tgtctcggctgttaattgccccttttttgtttttacgatgcaattactagcttgttagg 10445365
Sbjct: 10445366 attcagtattatttggaagccaaaggaaaaggtcacaataatggcagaagcggctgattt 10445425
Sbjct: 10445426 cgttaaaaataaaattaacaatggaacatactcagttgccaataaacataaaggaaaaag 10445485
Sbjct: 10445486 tgttatttggagcattttatgtgacattttaaaggaagatgaaactgttctggacggatg 10445545
Sbjct: 10445546 gctgttctgcaggcaatgccagaaagtgctcaaatttttacacaaaaacacctccaattt 10445605
Sbjct: 10445606 atcccgccataaatgttgtctaacattaagacgaccaacggaattaaaaattgtttcgga 10445665
Sbjct: 10445666 aaacgacaagaaagtagctattgaaaaatgcacccaatgggttgtccaagattgtcggcc 10445725
Sbjct: 10445726 gttttctgcagtaaccggagccggatttaaaaatttggtgaagtttttcctacaaatcgg 10445785
Sbjct: 10445786 cgctatctatggggaacaggtagacgtcgatgacttactacctgatccaacaacattaag 10445845
Sbjct: 10445846 tccttatatttaatagatacttttaagcccactatgtttttattatttagattgagaca 10445905
Sbjct: 10445906 ttaaaaaacgtaaaaatcaacaaatgccgtctttaattgcaattactttatgtgtttgaa 10445965
Sbjct: 10445966 atgggaggcacccattgagtccatcaaagagcaaagacatgagcacaaaaattttcttgg 10446025
Sbjct: 10446026 gtattccctttaccccttcatttcttatacccgtcacgcttccacccatacaaattttag 10446085
Sbjct: 10446086 gcgtacaaaaaatgaccagagaactgcagcccgcatacaaaaaatgacctgcggccgatc 10446145
Sbjct: 10446146 gttgactgtgcgtccactcacccatacggctcttgcgcagcaggcctcgggtggtttttt 10446205
Sbjct: 10446206 tactcgtaacaaaaacacaacgtcggtaaaacactcgagtattttgtgttgccgcaagta 10446265
Sbjct: 10446266 gggtgtcaaaaaaaacggggtgcctagagtaccgagtgtttatcgggtggacgtagagtg 10446325
Sbjct: 10446326 cgagtggcgggctgcagttctctg 10446349

Score = 686 bits (1770), Expect = e-198 Identities = 351/351 (100%) Strand = Plus / Plus

Query: 4911     gttgggaccactgttgtcctgtaattaaataatgcccttgctgctgccacgcccccaatg 4970
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10446350 gttgggaccactgttgtcctgtaattaaataatgcccttgctgctgccacgcccccaatg 10446409

Query: 4971     ccatctgaccccgacgaccgcccactttacggcatgaaaaacacaaacagaaagtatctt 5030
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10446410 ccatctgaccccgacgaccgcccactttacggcatgaaaaacacaaacagaaagtatctt 10446469

Query: 5031     aacggcatttgaagttgaaggagccgaggtcttgcggtggacagaagttatatccgtgtt 5090
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 10446470 aacggcatttgaagttgaaggagccgaggtcttgcggtggacagaagttatatccgtgtt 10446529
```

Example 2. 126 bp insertion into DGRP line 375 chromosome 2R Vs 5.1 reference

```
Query: 4832   tacagatagatgtttaaactgtcccaccccgtaaaagtactgtcaaaagttcaaagttca 4891
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860872 tacagatagatgtttaaactgtcccaccccgtaaaagtactgtcaaaagttcaaagttca 860931

Query: 4892   cccattcaacagtctcagaaacaatgagatgcctcaaaaaggagaacccagggagagatt 4951
              |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct: 860932 cccattcaacagtctcagaaacaatgagatgcctcaaaaaggagaacccaggaagagatt 860991

Query: 4952   tcaaggaacttttagatagggcagcttatgagtacagctatagttcctattcagctacga 5011
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860992 tcaaggaacttttagatagggcagcttatgagtacagctatagttcctattcagctacga 861051

Query: 5012   aaaagtataagtattagaagtattcttcggcagaaaagtgaagacaaacccggactaata 5071
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 861052 aaaagtataagtattagaagtattcttcggcagaaaagtgaagacaaacccggactaata 861111

Query: 5072   cgaagcagccagccaagataacattgaaagcatttaaaacaaacattttattccccaata 5131
              |||||||||||||||||||||||||||||||||||||||||||||||||| ||||
Sbjct: 861112 cgaagcagccagccaagataacattgaaagcatttaaaacaaacattttcttcc------ 861165

Query: 5132   tctacccatatcccagaaaaattatgaaatttcgcgttcgcactcacactagctgagtaa

Query: 5192   cgggtatctgatagtcgggaaactcgactacagcattctctcctgtttttttttataatt

Query: 5252   taaaaaaaaat 5262 — BOLD = 126 bp insertion in DGRP line 375 vs reference

Score = 646 bits (1667), Expect = e-186 Identities = 332/334 (99%) Strand = Plus / Plus

Query: 5263   ccttttgccacgaccattctaaggcactaaaaccgcacaaaaatgccaactcaaagagg 5322
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 861166 ccttttgccacgaccattctaaggcactaaaaccgcacaaaaatgccaactcaaagagg 861225

Query: 5323   attgatgtgagcaacatagaattccacgtcatagccgacgaaggcatgaaattcattgcg 5382
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 861226 attgatgtgagcaacatagaattccacgtcatagccgacgaaggcatgaaattcattgcg 861285

Query: 5383   attttaggtagaactcgttttctaatgatagacatgaaaaccaggaagggtcgtactaaa 5442
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 861286 attttaggtagaactcgttttctaatgatagacatgaaaaccaggaagggtcgtactaaa 861345

Query: 5443   gtcataccaagatcagacagggaggtagtgacggggctcaaaaagcaggtccaggatgga 5502
              |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 861346 gtcataccaagatcaggcagggaggtagtgacggggctcaaaaagcaggtccaggatgga 861405
```

Example 3. A more complicated insertion in DGRP 375: 2 nearby insertions, 497bp and 106bp on chromosome 3R.

```
Query: 3480     ggttgagcttacgtgactcagttgaaaaatgttgaggtcctggtaaattcaaaattaaaa 3539
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct: 24465291 ggttgagcttacgtgactcagttgaaaaatgttgaggtcctggtaaattcaaaattgaaa 24465350

Query: 3540     gcaaagatctctcttaagtgacaaaaagtcactgcaatacattaaagcgccgctttgttt 3599
                ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24465351 gcaaatatctctcttaagtgacaaaaagtcactgcaatacattaaagcgccgctttgttt 24465410

Query: 3600     tgcagatatatatttagtgtcctgccgcgtggatttaaagggcgaaagtgctggcattgc 3659
                ||||||||||||||| |||||||||||||||||||||||| ||
Sbjct: 24465411 tgcagatatatatttattgtcctgccgcgtggatttaaagggccaa 24465456

Query: 3660     aacagtaagtccatccagcagctgccacactctccagtgccatggggcaggtgatgctgg 3719
Query: 3720     tgcagaaagctggagccggaggtcgagttctggcgaagatgttattcagggtgatgaaaa 3729
Query: 3780     cactgtgcggtgcattggaggaggtgcggaccaccttgtagagccacaggggcattatgt 3839
Query: 3840     tcttggtgccggcgcccagaataatgcgacgctggcgcatccgattgtcaggaagatgga 3899
Query: 3900     gatcaccattagtcccggtcttgatcctcagcgatccgccggcacttatgcgagttcgaa 3959
Query: 3960     tccaccgctcgatggtctcccaattgcgatcgttgatgctcttaaacatgggtgccacat 4019
Query: 4020     tgacgtacttgaaggtggcgcacatctggtctccatacaggtagtctcccgatggcgtca 4079
Query: 4080     gatgtccccgattgatgatcacatcgcgttcgttgggaatatacctctggttattgccga 4139
Query: 4140     agatggtgcggaatgcacggtaaatgctgcg 4170

Score = 92 bits (238), Expect = 5e-19 Identities = 48/49 (98%) Strand = Plus / Plus

Query: 4171     aggcagccagtggtctttggcagcgcattgggaatggcctgctgaagtt 4219
                ||||||||||||||||||||| |||||||||||||||||||||||||
Sbjct: 24465457 aggcagccagtggtctttggcatcgcattgggaatggcctgctgaagtt 24465505


Query: 4220     gttgtgcgtccgctgacaggtcgtctgtacgatgtcgttcggctggcagatcatgtggag 4279
Query: 4280     cacctcgttggtgcggactctctgcagacgcttcaggtcgtactg 4324


Score = 2007 bits (5179), Expect = 0.0 Identities = 1042/1057 (99%) Strand = Plus / Plus

Query: 4325     gccggcggcattgcggttcgtaaagatcctattggtgtcctcgatcattgctctggagag 4384
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24465506 gccggcggcattgcggttcgtaaagatcctattggtgtcctcgatcattgctctggagag 24465565

Query: 4385     cggacagtccgcctgggacaggagcactgcaatgagcattatttggtgcattacggaagg 4444
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24465566 cggacagtccgcctgggacaggagcactgcaatgagcattatttggtgcattacggaagg 24465625

Query: 4445     attccttattgatatgaatttcatttagaatcaccaggttataagtggtttatatacata 4504
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24465626 attccttattgatatgaatttcatttagaatcaccaggttataagtggtttatatacata 24465685
```