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

Trudy Mackay, Stephen Richards, George Weinstock, 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; and an additional 96 lines to be chosen with input from the community. 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 ~ 0.01 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 traits, thus giving an unprecedented opportunity to quantify genetic correlations and pleiotropy among traits, as well as evaluate the same traits in multiple environments 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. Sampling 96 additional strains from other geographic locations will capture additional variation to identify polymorphisms in the whole sample with a minor allele frequency of ~0.01 or greater. 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 core set of lines comprises the genetic reference panel, which has been extensively phenotyped for a battery of complex traits, and which constitutes a long-term resource for further phenotyping and experimentation by the *Drosophila* community. The reference panel will be sequenced to a minimum of 5X coverage using the latest generation massively parallel sequencing technologies.

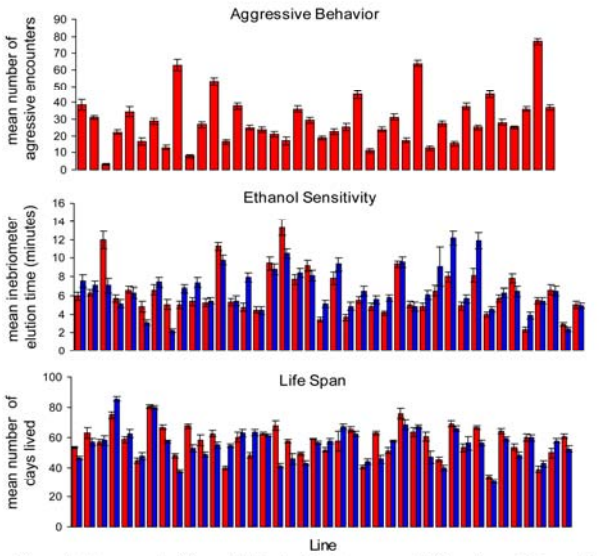


Figure 1. Variation in three quantitative traits in 40 of the proposed 192 line Drosophila genetic reference panel. Red: males, Blue: females.

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 Figure 1), with broad sense heritabilities ranging from 0.22 – 0.78 (Appendix Table 1). 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 $n \gtrsim 2(z_\alpha + z_{2\beta})^2/(\delta/\sigma_P)^2$ (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 Table 2 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.

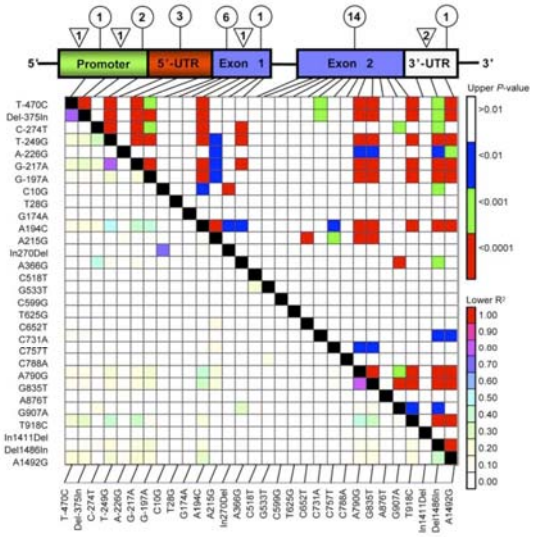


Figure 2. *Catsup* polymorphisms show an absence of haplotype blocks. The *Catsup* gene structure is depicted with the number and distribution of SNPs (circles) and InDels (triangles) in 169 *Catsup* alleles sampled from the Raleigh population. LD in *Catsup* is shown below the gene structure, with *P*-values 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.

Surveying *Drosophila* genetic variation, additional strains: We also propose sequencing 96 additional strains, from outside the Raleigh population. Whilst the Raleigh population has been chosen on grounds of availability, inbreeding status, and the extensive phenotype information available as a base for future efforts, it is but a single data point of world wide *Drosophila* genetic variation. We are proposing the sequencing of an additional 192 lines for several reasons. First, it will provide a more balanced view of *Drosophila* genetic variation, making the polymorphism data more widely applicable for experiments not involving the Raleigh lines. Second, the resolution of the polymorphism map will be increased to enable detection of a minor allele frequency of ~0.01. Finally, we hope to provide some initial public *Drosophila* genotype data to allow data driven experimental design for future population genetics studies. We have started a community dialogue to identify these lines (Appendix Table 4). Initial feedback from members of the *Drosophila* population genetics community focused on two main themes: African populations and populations from a geographic cline. We propose to sequence 48 African lines (including representatives of the Z and M behavioral races) and 48 lines from a cline (most likely along the US east coast, to take advantage of the available Raleigh lines, by adding Maine and Florida populations). The exact composition of the lines will be determined by a continuation and enlargement of the community dialogue, assessing what is available as well as inbreeding status. As some time will be required for these processes, we plan to sequence these lines in the second half of the project and thus will take full advantage of updated versions of the new massively parallel sequencing technologies available at multiple genome sequencing centers, greatly reducing the costs.

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 288 lines will identify a common set of dense polymorphic markers and allow an economic and accurate platform for genotyping Drosophila for any purpose. The 288 lines have a 95% probability to contain alleles with MAF 1%. 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: 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). As the sequences can be compared to the high quality reference *D. melanogaster* sequence, and because of the inbred homozygous nature of the phenotyped strains, we will take advantage of

new massively parallel sequencing technology at relatively low coverage. Our current default plan is to each line at a minimum 5.5 X coverage using Solexa technology (1 x 1Gb run per line based on 180Mb genome). We will experiment with increased data quality and sequence coverage for the 40 core Raleigh lines, to experimentally assess the best trade off between sequence coverage and experimental value, for what likely be the strains most heavily used by the community. The sequencing strategy will be updated as technologies are rapidly evolving and will allow greater coverage for all lines and further reduced costs. This amount of coverage will ensure that the vast majority of the genome will be covered with more than a single read. Theoretically (Clarke and Carbon, 1976), in an ideal case ~98% of the genome will be covered by at least a single read, and ~92% covered by two or more reads. In practice, single read coverage will be useful to distinguish SNPs and indels in regions of low coverage in any single strain that have been identified and characterized in other strains in regions of higher read coverage.

Polymorphism Identification: The BCM-HGSC has experience identifying sequence polymorphisms on this platform; specifically, SNP and indel detection on patient samples. All massively parallel sequencing technologies to date require PCR amplification of the sample, creating the possibility of polymerase errors appearing in the sequence. We expect such errors will be at orders of magnitude lower frequencies than our detectable minor allele frequency based on prior experience, and such errors will be identifiable due to the proposed coverage. The error model for all the new technologies is different from that of traditional Sanger based sequencing, for example, homo-polymer length for the 454 platform, and sharp declines of quality at the end of reads for the Solexa platform. Because of the massive coverage over the total population, we expect to define error rates with precision with between line sequence comparisons. A number of candidate quantitative trait loci will be confirmed using traditional Sanger sequencing, allowing measurement of false positive and negative rates. This project will provide an excellent dataset for the refinement of such techniques. These data will be presented in the Genboree browser as well as being made available to FlyBase, Genbank and all other appropriate public databases.

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, 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. In addition 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, and 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: Quantitative traits provide the basis of the majority of the genetics of human health, whereas single gene Mendelian traits, 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 however, we do not expect any additional expansion of the community due to these sequences. There is also a need for data sets for the training of statistical methods for QTL identification. It is our belief that this dataset will also be used by biostatisticians outside of the *Drosophila* community. This has already been the case for the original *D. melanogaster* sequence which is widely used as a test bed for genome assembly, and gene prediction software.

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 0.01 (the probability of not observing a single allele with a population frequency of 0.01 is 0.055 in a sample of 288 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: As one of the members of the original *D. melanogaster* sequencing consortium finishing 1/3rd of the *D. melanogaster* reference sequence, and based on this experience as well as with *D. pseudoobscura* and several other insects, we foresee no challenges in terms of biological features that will hinder this project. DNA has been isolated from all 192

NC strains and is ready for sequencing. A minimum of 5X coverage has been chosen to ensure the vast majority of the genome is covered by at least a single read, however the core set of 40 NC strains will be sequenced to a higher quality, as these are likely to be most used by the community. Additionally, some of the polymorphism discovery strains will be sequenced to higher coverage as they are often unable to be inbred, but this will allow additional chromosomes and polymorphism to be sampled.

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. At least three technologies are likely to compete: Solexa – currently estimated at ~ $3,000/Gb sequence 35bp reads (reagents only, April 2007) has a roadmap to 6Gb per run at the end of 2007. The 454 technologies platform (currently 250bp reads, 120Mb/run) is expected (according to a company representative) to be at 1Gb/run in 500bp reads by the end of 2007, at a competitive price. Finally, Applied Biosystems (AB) demonstrated their ligation based sequencing technology (due late 2007) at a recent sequencing conference, again at the 1Gb in short read format directly competing price wise.

The current default sequencing scheme (which will likely change to utilize updated technologies with lower costs) is a single run using the Solexa technology, producing 5.5 X coverage per strain at low cost. Future platforms from any of the companies will allow additional strains to be sequenced per run, and also higher coverage to ensure greater quality, and the use of technologies with longer read lengths to better characterize small insertions and deletions.

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. When totaled, 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 Table 1

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 Table 2

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.

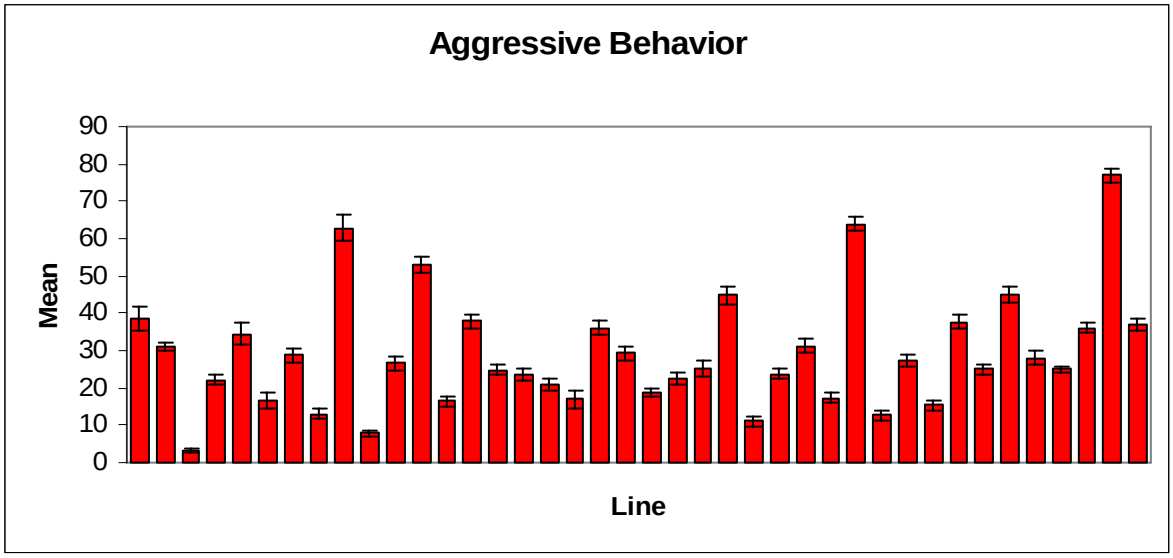
[a] AG = aggressive behavior; LR = locomotor reactivity behavior; LS = life span; SR = starvation resistance; ER = ethanol resistance; CC = chill coma recovery; CL = copulation latency; ST = sternopleural bristle number; AB = abdominal bristle number

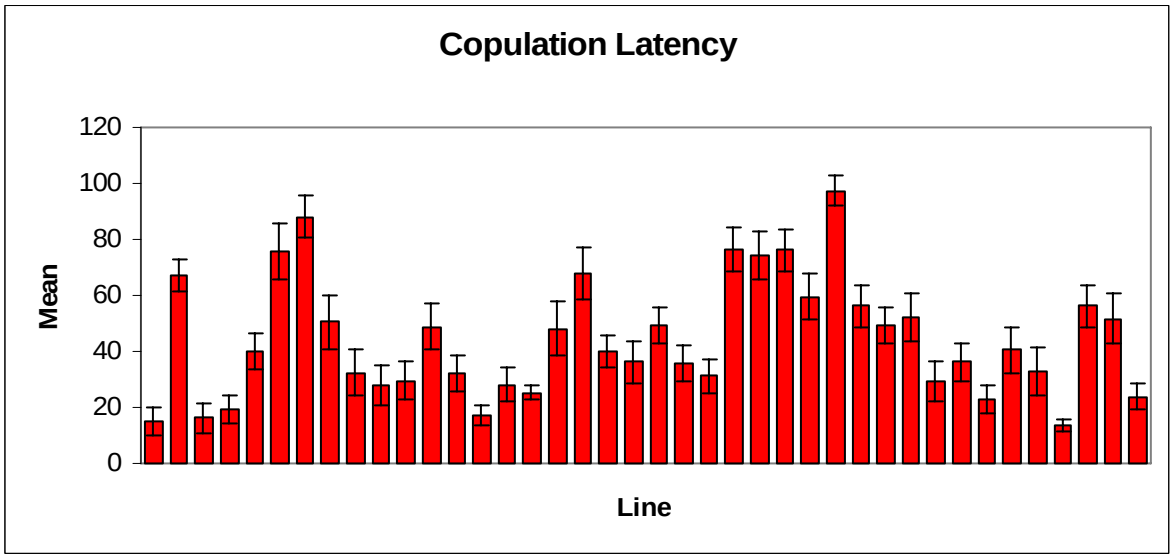
[b] Phenotypic standard deviation

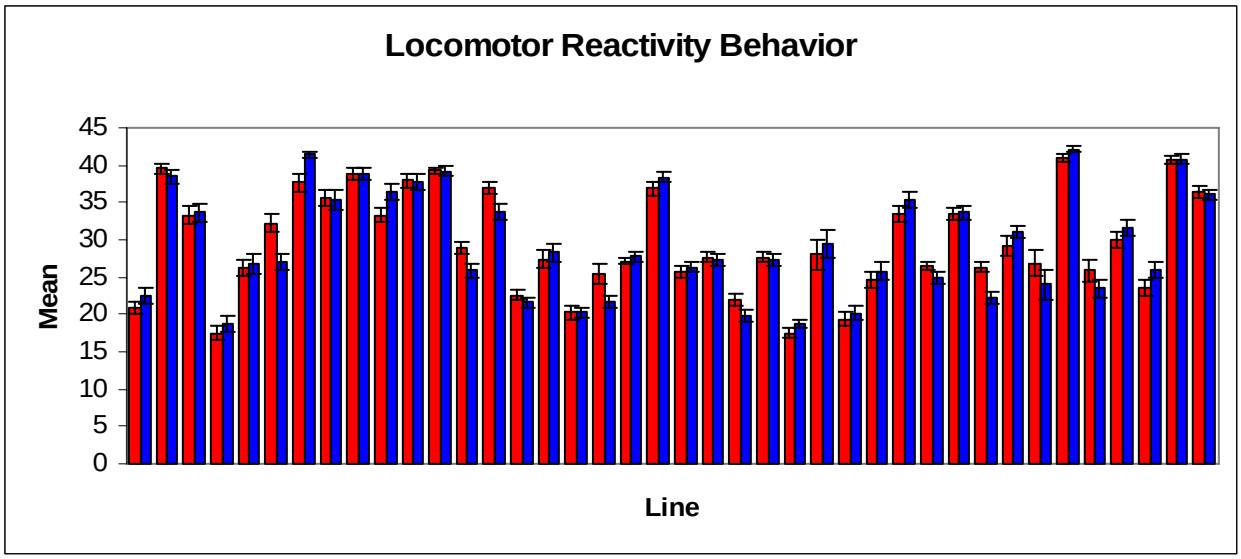
[c] δ = difference in mean between homozygous markers

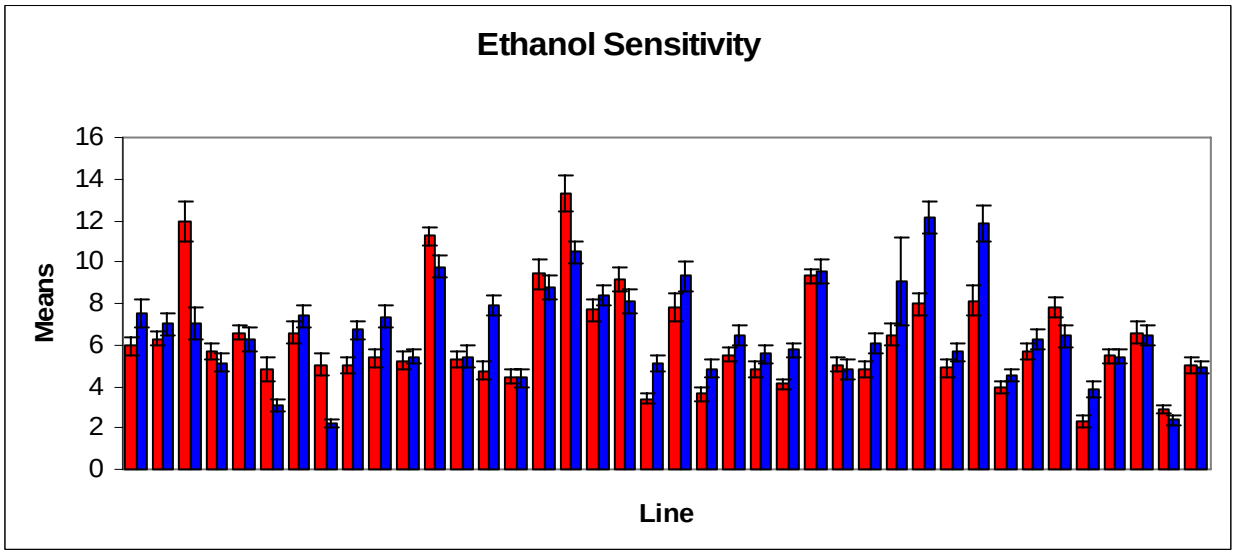
Appendix Figures

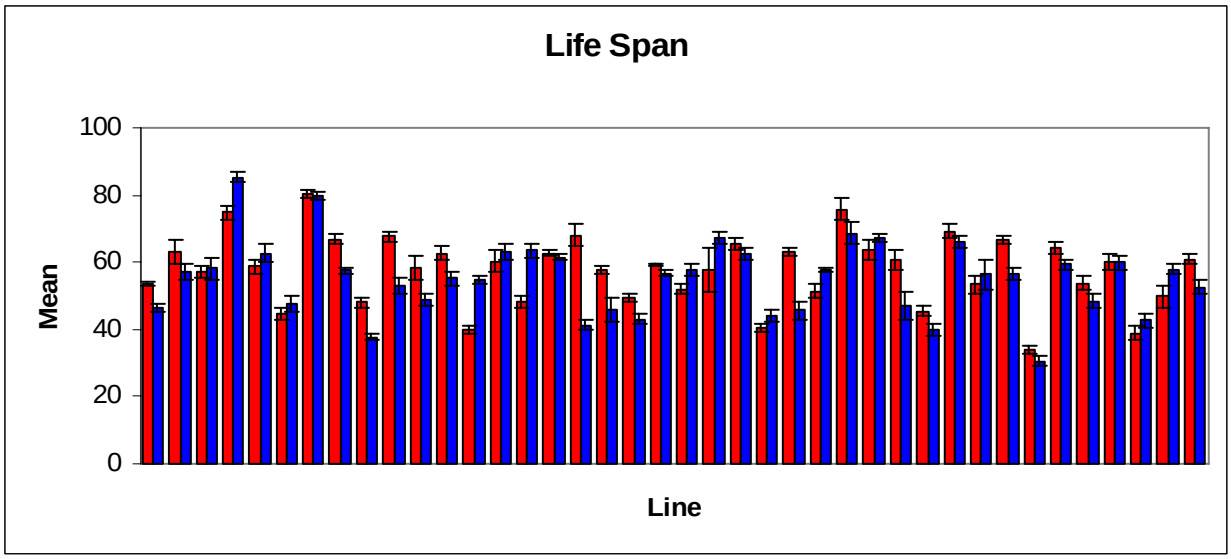
Quantitative Variation Among the Core Set of 40 Lines From the Drosophila Genetic Reference Panel

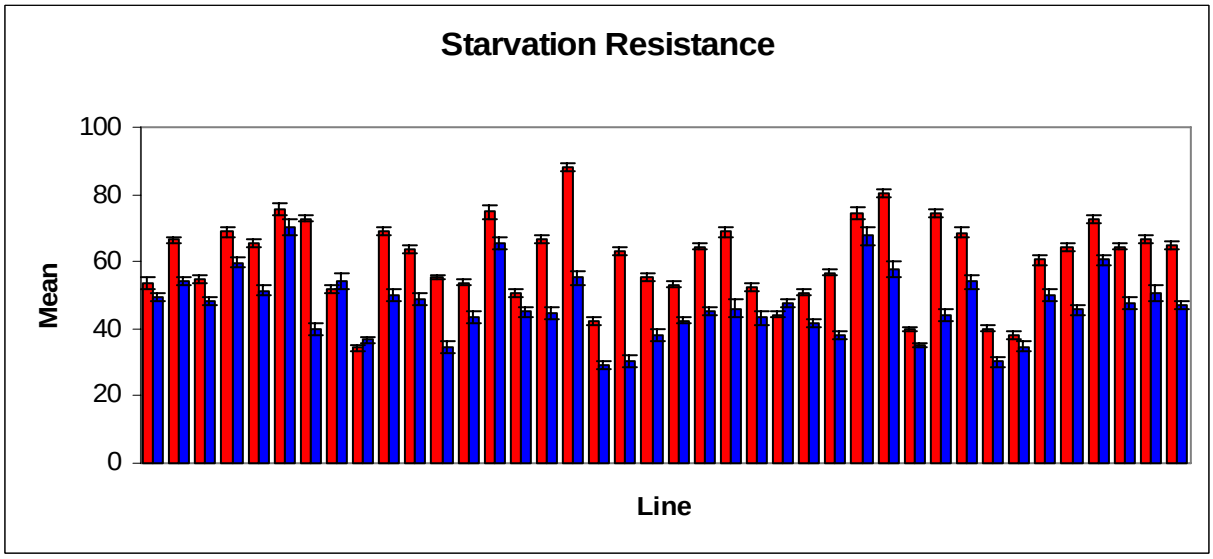


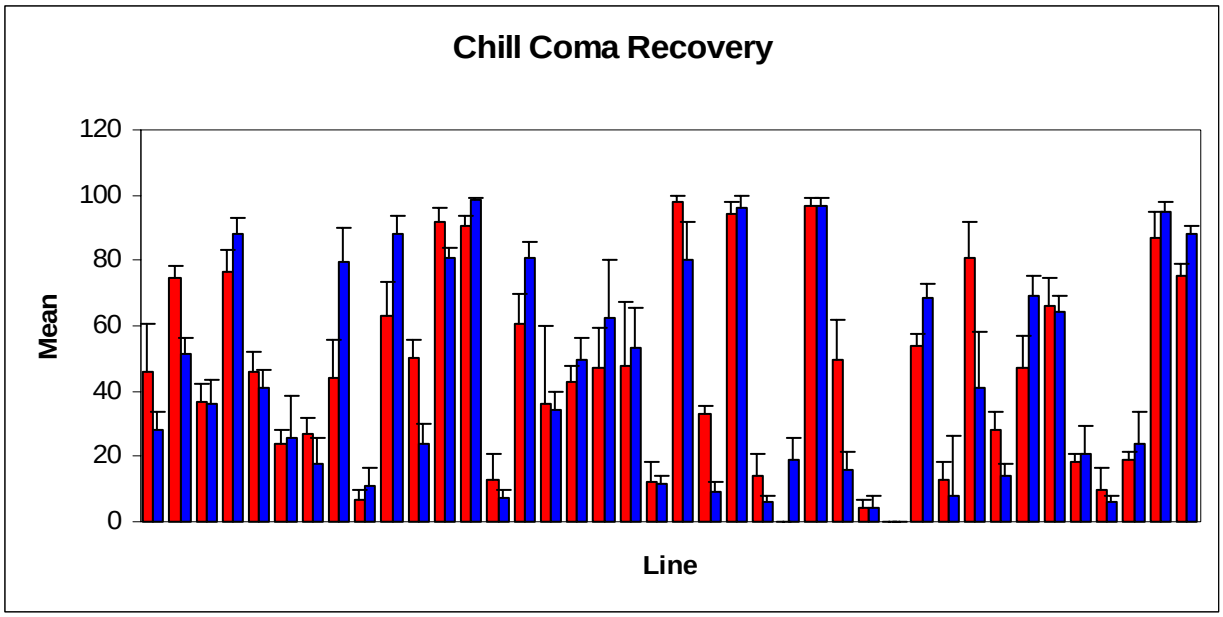


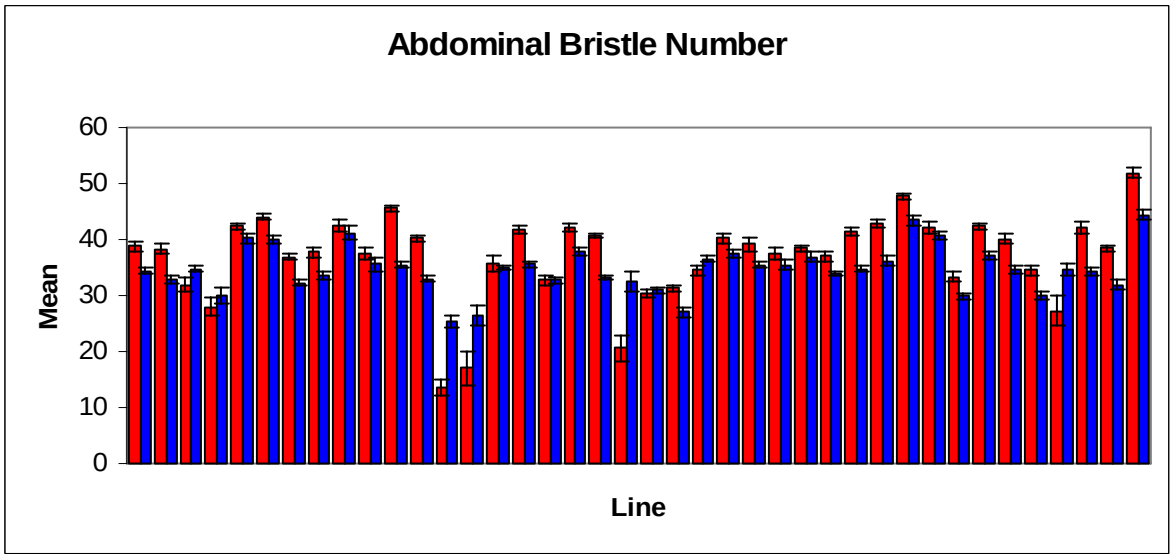

Locomotor Reactivity Behavior
Mean
45
40
35
30
25
20
15
10
5
0
Line

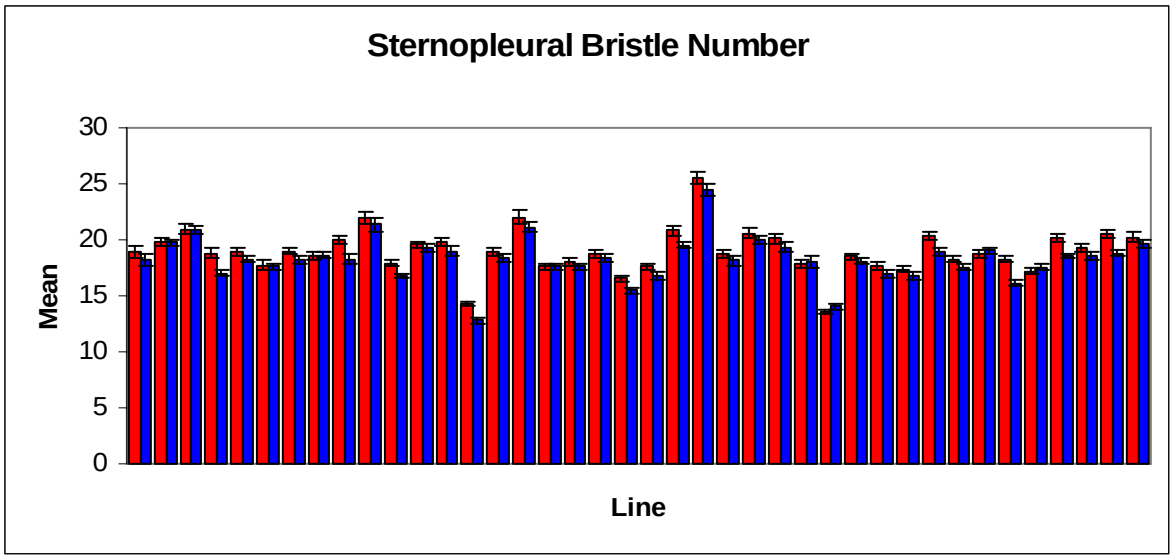

Ethanol Sensitivity
Means
16
14
12
10
8
6
4
2
0
Line


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