Abstract

Palsson, Arnar. Molecular Quantitative Genetics of wing shape in *Drosophila melanogaster.* (Under the direction of Greg Gibson.)

Building on quantitative genetic analysis in fruitflies I chose to investigate the molecular genetic underpinnings of natural variation in wing shape. Shape is a complex trait demanding a multidimensional description and was adequately portrayed and quantified with the tools of morphometrics. The results demonstrate uncoupling of size and shape. Shape shows strong degree of integration over the structure, disputing hypothesis of the wing as comprised of independent modules laid down by compartmentalization. However, distinctly local shape effects are also observed in genetic correlations, complementation and association tests, arguing for a continuous distribution along an axis of integration and modularity.

The identification of quantitative trait nucleotides within a QTL was pursed in a two step scheme. First I tested a set of candidate loci, implicated by QTL experiments and/or developmental roles, for contribution to wing shape. The results are consistent with segregating variation of loci in the vein-determining pathways, *hedgehog* (*hh*)*, decapentaplegic* (*dpp*) and *Epidermal growth factor Receptor* (*EGFR*), impacting shape. The second step involved fine-scale mapping, by testing for associations between *EGFR* and wing shape in two geographic populations of *D. melanogaster.* The genotyping was done by sequencing 10.9 kb of the locus from 209 lines demonstrating a mostly neutral locus, possibly experiencing purifying selection. One of two alternate 5'-exons may be evolving more rapidly by positive directional selection. Linkage disequilibrium decays rapidly within *EGFR* increasing the resolution of association mapping. Association tests identified one site (C31365T) with sex dependent effects on wing size, significant after Bonferroni correction. Seven more sites are weakly suggested. The highest of those (C30200T) disrupts a putative GAGA factor binding element and has replicable effects on crossvein placement in three study designs. The work suggests naturally occurring polymorphisms in *EGFR* affecting size and shape of the *Drosophila* wing.

Molecular quantitative genetics of wing shape in *Drosophila melanogaster*

by

Arnar Palsson

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment for the
requirements for the degree of
Doctor of Philosophy.

Genetics

Raleigh

2003

Approved by:

_____	_____
Co-Chair of Advisory Committee	Co-Chair of Advisory Committee

Biography

Born in Reykjavik Iceland on November 21st 1970, raised on a farm under a barren mountain and later in a town with few buses. Acquired early a drawing skill but kept a withdrawn character throughout his teenage years. Traveled to Turkey and bought a rug with great plans in mind. Only later did those materialize in a shady forest on a distant continent. Wrote short stories for radio about a tortured teenage mind in modern society seeking to rid the world of cliques. The stories were seriously quite lighthearted but eventually became the quintessence of what was being criticized and had to be killed off, more then once and more then twice, because good cliques die hard. Educated in a high-school for arts and natural sciences and sought further degree in biology. Received a Bachelor and Master of Science at University of Iceland in 1995 and 1998 respectively. During his studies he fell in love with Solveig S. Halldorsdottir, and the two of them married in the summer of 1998. That same fall he started working on his doctoral dissertation project at North Carolina State University. The rug was but forgotten. Over the course of 5 years they made new friends and begat a son. The buses continue to amaze.

Acknowledgements

I am eternally grateful for the vision and support of my advisor Greg Gibson, the white, a true genius by all standards. Other committee members deserve highest credit, in particular Trudy MacKay for help with experimental design, analysis of quantitative genetic experiments, and comments on the thesis draft. Thanks to Jeff Thorne and Michael Purugganan for discussions and suggestions on how to improve the thesis. Thanks to Marla Sokolowski, Macarena Busto, Bill Ballard, Avis James and Richard Lyman for stocks of *D. simulans* and *D. sechellia*. Particular thanks to Ed Buckler for use of his Tassel software prior to public release and excellent discussions. My gratitude to Justin Fay for help with the H statistic and a grand takk to Lior Pachter for help with AVID and Vista.

This work could not have been completed without the help of several devoted people in the Gibson laboratory. Those contributing directly will be thanked explicitly but others more generally with a bowl of fish. Thanks to Leslie Polzien for help on Chapter 2, and Joy Morris for work contributing to Chapter 4. Thanks to Roland Carillio, Rebecca Berger and Kelli Birdsall for establishing and inbreeding the WE lines, and parts of the Kenyan chromosome substitutions. I am most obliged to Ann Rouse and Rebecca Berger for generating the *D. melanogaster* sequence data and Ian Dworkin for partnering in the clean up and outgroup sequencing. My gratitude to James Dodgson who completed majority of the round-robin experiment. The analysis where refined and substantially improved by direct input and commentary by Ian Dworkin and Naruo Nikoh. The clarity of the text can be attributed to the helpful comments of Greg Gibson, Trudy MacKay and Solveig S. Halldorsdottir, but I am fully responsible for elfish that may have slipped in. I also have great appreciation for Rebecca Berger and other members of the lab who helped produce the volume.

The whole of the Genetics Department at NCSU, staff, faculty and students, contributed by providing intellectual environment in very hospitable surroundings. Thanks to NCSU Genome Research Laboratory for help with sequencing. I was supported in part by fellowships from the American Scandinavian Foundation, (Thor Thors fellowship) and The North Atlantic Treaty Organization.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Trying to understand the mechanism and dynamics of evolutionary change biologists strife to identify the genetic modifications responsible for phenotypic divergence. The marks evolution leaves in the DNA justifies this kind of dissection. Eventually analysis of these genetic signals may enlighten our understanding of the interplay and respective relevance of evolutionary forces. A more basic question concerns the accuracy of the genotype to phenotype representation and its relationship to genetic variation in populations. This is not purely an evolutionary question for it touches on our understanding of the robustness of molecular and cellular processes.

Dissection of the genetic causes of phenotypic diversity has been pursued by comparisons between distinct taxa, species or subspecies or by analyzing genetic variation at the species level. A typical experiment of the first kind usually starts with a functional characterization of a gene in one species indicating its importance for a particular trait. By comparison to another species, the contribution of attributes like DNA or protein sequence, mRNA expression patterns, or protein distribution or function, to observed differences in the trait of interest is postulated. An obvious problem is the circumstantial nature of the data. Fortunately experimental genetics offer a number of ways to substantiate findings of this kind, for example by rescuing mutants by making transgenic individuals with “homologous” genes from other species. Another interesting method involves extension of classical complementation tests to related species or subspecies (Sucena and Stern 2000, Kopp and Carroll 2001, reviewed in Stern 2000, Gibson and Palsson 2001). Both of those methods have their limitations but have already proven useful for studies of species lending themselves to transgenics or between species hybridization.

The second approach is that of quantitative genetics, building on Darwin’s notion that differences between species start out as heritable variation within a population. An important caveat is that even though a locus is demonstrated to contribute to segregating variation for a trait in a population, it does not follow directly that the locus will contribute to adaptive evolution. This method will therefore expose a pool of variants available, but can not tell us about their evolutionary fate, with the possible exception of major deleterious alleles that have very low probability of being fixed. Identification of the alleles or polymorphisms contributing to standing variation for a trait is not trivial, due to the complexity of genotype to phenotype relationships, vastness of the genome and a lack of connection between statistical and biological significance.

The current toolkits of quantitative genetic can zoom in on a chromosome region, gene or single nucleotide polymorphism (SNP) that is significantly correlated to the trait. As with between species comparisons, the biological importance of a genetic variant can in theory be tested with transgenic experiments. In *Drosophila* this principle has only been implemented for variants in the *Adh* locus (Stam and Laurie 1996). In other systems, transgenic assays in humans cell lines have been used to confirm the molecular action of associated mutations (Rockman and Wray 2002), and also in several plant studies (Frary *et al.* 2000, Cong *et al.* 2002, Tian *et al.* 2003, review by Mauricio 2001 and Remington *et al.* 2001a).

The ultimate tool for analysis of quantitative trait nucleotides is homologous recombination. Yeast is the best developed system for such manipulations but it was only recently that quantitative genetic analyses have been attempted (Steinmetz *et al.* 2002). The authors reported very complex cis and trans-interactions modifying the effects of the nucleotides tested and therefore emphasized the importance of direct experimentation for validating quantitative trait nucleotides. Rong and Golic (2000) developed homologous recombination in *Drosophila*, and proved the principle by rescuing a *yellow* mutation. The utility of the method for reverse genetics is still being explored but a handful of labs have now reported success in knocking out loci of interest (Sawamura *et al.* 2000). For the purposes of quantitative genetics the greater interest is to "knock-in" allelic variants of interest, even just a single nucleotide. Another way of confirming an association is to replicate the study. This can either be done with a new sample of chromosomes or by using an existing panel of alleles, for instance substituting these into a common background or subjecting them to a test cross.

The introduction is divided into three sections. First the project aims are stated. Questions regarding heritable factors contributing to phenotypic variation in natural populations are put forth and the procedure for mapping quantitative traits to individual polymorphisms is discussed. Particular emphasis will be put on recent advances in *Drosophila* and the designs available. The importance of fully embracing the phenotypic space will be stressed. The design of the association mapping experiment demands that we discuss possible complicating factors like population structure. Second, intense focus will be put on the molecular and developmental biology of the structure I chose to work on, the *Drosophila* wing. The last section will focus on the evolution and quantitative genetics of wing shape. I believe that scrutiny of molecular details is justified as the coupling of detailed information with the framework of quantitative genetics will be important for progress in the field. If we want to step out of the black box and make statements about the genetic basis of evolution then we need to understand how the genes bring about phenotypes.

A. Quantitative trait nucleotides

Key questions and project aims

Continuous variation in natural characters has its material basis in the allelic variation of individuals and in the environment they belong to. Identification of the heritable component of trait variation is important for practical purposes of agriculture and medicine in addition to the epistemological value of evolutionary questions.

After a hundred years of pursuing Mendelian inheritance, medical geneticists now aspire to identify allelic variants causing subtle phenotypic variation in complex human diseases. The majority of human diseases are affected by mutations at numerous loci, stressing their polygenic nature. Current efforts to identify genetic agents in disease rely on pedigree analysis or association studies in outbred populations. Any experimental verification of the effects of an allele or mutation is limited to analysis in cell culture or "relevant" animal models. There are several problems with either approach in order to test the true biological effect of every genetic variant identified. Of more importance for the current study is the extension of the human research program to a species amenable to genetic manipulation in order to enrich our understanding of complex traits. In particular the aim should be to identify nucleotide variants correlating with phenotypic values and verify their biological effects experimentally or by repetition.

The general proposal of my thesis is to focus on the first part of this problem, to establish a new *Drosophila* system to complement sensory bristles (MacKay 1996), for fine-scale genetic dissection of composite traits. The choice of *Drosophila* for this aim is justified by a body of work (Powell 1996, Spradling *et al.* 1999, Adams *et al.* 2000, Held 2002) highlighting the utility of the model. The specific aims are as follows:

A. To define the appropriate morphometric procedures for integrating developmental and quantitative genetic analyses of the *Drosophila* wing

B. To identify a subset of loci for which naturally occurring variation may affect aspects of wing shape.

C. Investigate the molecular evolution and the level of population differentiation of a candidate locus, the *Epidermal growth factor receptor* (*EGFR*).

D. Test for the contribution of naturally occurring allelic variant(s) in *EGFR* on parameters of wing shape.

Fulfillment of these goals will allow me to address several pressing questions about the nature of quantitative genetic polymorphisms. The idea is to survey all the polymorphisms, base

changes and insertion/deletions, in contiguous regions of a candidate locus and then ask directly about the effects of these nucleotide changes being tested. If the test of association between polymorphism and phenotype is significant then we can make statements about the quantifiable effects of those polymorphisms, commonly called Quantitative Trait Nucleotide (QTN). Experiments such as these initially serve simply to identify a handful of possible QTN's and do not therefore justify generalizations. Establishing which properties of QTN's are most interesting is however in order for later synthesis.

1. The exact molecular nature of the QTN is of interest: is it a base-substitution or an indel? In most *Drosophila* genes a 9:1 ratio of substitutions to insertion-deletion polymorphisms is observed. *A priori*, we would expect functional sites to show the same pattern.

2. Where in the locus does the polymorphism reside? For example does the change alter a protein or does it land in an intron. The gene organization is fairly well known for a good portion of the *Drosophila* genome, either from the literature or genomic efforts. Some genes are even characterized down to individual transcription factor binding sites or specific protein domains, making this comparison very exciting. In case of synonymous changes then the respective frequency of codon polymorphisms is also of curiosity.

3. At what frequency is the QTN segregating and what is the ancestral state? This can reflect the evolutionary history of functional sites, and be used to test the theory of mutation selection balance.

4. What are the exact phenotypic effects of the change? For instance, which character state causes a reduction in trait value? We would also like to assess if the QTN has dominant or recessive effects. The connection between the derived allele and direction of effects is also of interest.

5. Is the site in Linkage Disequilibrium with other genotyped sites? If a site experienced positive selection recently then higher than average LD should be retained in its vicinity. General inspection of the pattern of LD around the site can indicate the history of the SNP in the population.

6. What are the values of molecular evolutionary parameters in the region? Contrasts of polymorphism levels and divergence provide a basis for analysis of the evolutionary forces that shape the distribution of variation.

7. If more than one QTN is detected then we can test if multiple functional variants constitute super-haplotype (Stam and Laurie 1996, Lai *et al.* 1994). Can QTL effects be reduced to single QTN's or are they more commonly a property of haplotypes within which multiple small QTN effects accumulate? Similarly one can test for epistatic interaction between QTN's.

In addition to these general points then the design of the association tests allows two more comparisons. The study involves two populations and I can therefore test for population

dependence of QTN effects. It is proposed that those should be interpreted as a polymorphism by genetic background interaction. Second, as the experiment investigates the multidimensional phenotype of wing shape then one has the chance to test for pleiotropic effects of polymorphisms.

Barton and Turrelli (1989) in their inspiring but bleakly titled "Evolutionary quantitative genetics: how little do we know?" review concluded that we should step beyond the selection equations to "uncover the forces that produce differences between taxa. They envisioned the role of laboratory experiments to complement the traditional studies. The field of *Drosophila* quantitative genetics has not quite lived up to this ambitious vision but the past 12 years of research have progressively created a panel of loci and polymorphisms with strong associations to aspects of phenotypes. The progress has been quite impressive (MacKay 2001) and with new techniques and approaches the field is calling for broader sampling of traits and genes. Further refinement of wing shape, a coherently approachable set of traits, as a model system in molecular quantitative genetics will be my contribution. The intent is to search for naturally nucleotides polymorphisms contributing to wing shape variation.

Designs of association tests in *Drosophila*

Tests of association involve the joint analysis of phenotypic and genotypic distributions in a population sample, and may be more powerful than linkage based studies (Risch and Merikangas 1996). Its most common uses are in the search for factors influencing human disease, where the paradigm was established for the identification of Mendelian variants. Quantitative or complex traits provide a weaker genetic signal to map and as such require additional considerations regarding experimental design (Long and Langley 1999). However association tests have also been applied to address evolutionary questions, primarily in *Drosophila* (Aquadro *et al.* 1986, MacKay and Langley 1990, Laurie *et al.* 1991, Lai *et al.* 1994, Long *et al.* 1998, Lyman *et al.* 1999, Long *et al.* 2000, Geiger-Thornsberry and MacKay 1998, Robin *et al.* 2002).

Testing for associations in outbred populations requires large sample sizes because they are susceptible to numerous sources of error and bias, including large environmental factors, disparity in age, and population stratification. Some of these can be amended with proper epidemiology and selection of homogenous subjects for the study. Tests of association in outbred populations have yet to be reported in *D. melanogaster*. Previous association studies have utilized the amenability of flies to laboratory rearing and more importantly genetic manipulation. The ease with which flies propagate in the laboratory argues that differences observed under experimental conditions may indeed have ecological relevance. The reduction in environmental variation also increases the power to detect genetic differences and is

applicable to all kinds of quantitative genetic experiments. Control at this level also allows direct estimation of the environmental dependence of effects (Dilda 2002) and therefore tests of multiple aspects of evolutionary theory (Falconer and MacKay 1996, Lynch and Walsh 1998). At the genetic level there are three main designs readily available, with ever increasing control over the genetic background. The perfect experiment of a single polymorphism substitution by means of homologous recombination is a practical possibility (Rong and Golic 2000) but has yet to be implemented for QTN's.

Of the other strategies then analysis of inbred lines comes closest to sampling from outbred population in the sense that each line differs across the whole genome. The main benefit, shared with all the other designs discussed below, is that the phenotypic state of each line can be measured very accurately by scoring multiple individuals in multiple replicates. These effectively clonal populations are established from impregnated females from the wild that found isofemale lines that are later subjected to sib-mating for 20-50 generations. The number of generations and lingering deleterious mutations will determine the effectiveness of the inbreeding, the extent of which is estimated either from theory or with molecular markers. The unique drawback is that the process of inbreeding may alter the allelic pool under investigation (Barnes *et al.* 1998). For instance lethal and deleterious alleles will tend to be selected against unless they are locked in a tight functional or physical linkage balance, which will maintain heterozygosity in the regions. It remains to be seen what the exact consequences of inbreeding are and how greatly inevitable adaptation to laboratory conditions (Curtsinger 1986) affects these kinds of experiments. The current paradigm rests on the assumption that these biases are shadowed by the increased power of the association tests with this and following schemes.

The *Drosophila* system is unique in the level of genetic control, which enables researchers to substitute a whole chromosome into a common background. Key tools enabling these manipulations are balancer chromosomes that carry multiple inversions which effectively prevent recombination. Thus chromosomes from the wild that contain deleterious mutations will be faithfully maintained over a balancer. The chromosome panel can then be tested in the homozygous state or in crosses to each other or major mutations, as will be discussed below. Panels constructed in this way differ only by one of the three main chromosomes thus reducing the noise in the experiment and making the tests more powerful (Mackay 2001). The second advantage is that the chromosomes will be shielded from selection thus giving a more complete representation of the genetic variation in nature.

A further reduction in variance due to genetic background can be achieved by introgressing the chromosome region of interest into a standard isogenic stock. This involves repeated backcrosses of individuals with the region of interest to the same stock. This method is applicable to other species than *Drosophila* and has been used to investigate QTL's in mice and

butterflies and speciation in other *Drosophila* species (Liu *et al.* 1996, True *et al.* 1996). An additional elaboration makes the *Drosophila* protocol more powerful. By starting with chromosome substitution lines then the introgressions can be conducted in replicate that will provide an additional level of control (Robin *et al.* 2002).

One adjustment can be applied generally to all of these methods., This is to outcross the alleles, for instance a round-robin manner. This will generate heterozygous conditions generally but reconstitute homozygotes at individual SNP's, and therefore allows assessment of dominance along with additive effects. The advantage is also that the effects of linked deleterious alleles will be reduced and the SNP's will be tested under more realistic circumstances with prevailing heterozygosity. Another common addition to all of these designs is to test for isoallelism (Thompson 1975) or, in its advanced form, differences in complementation (Long *et al.* 1996, Mackay and Fry 1996). This involves controlled crosses to known mutations in candidate loci that can enhance the differences between the alleles tested. Similar to the association schemes then these "sensitization" crosses can involve whole chromosomes or mutations introgressed into a common background (Long *et al.* 1998, Lyman *et al.* 1999, Robin *et al.* 2002). In addition to the increased power in association tests then this alteration can also be used to investigate the relationship between cryptic and standing variation attributable to polymorphisms in candidate loci.

A systematic experimental comparison of these designs for testing associations has not been conducted. The results from existing studies are in concordance with the expectations on the reduction in environmental contribution to variance, as reports by MacKay and coworkers (Long *et al.* 2000, Robin *et al.* 2002) demonstrate. Those studies utilize two of those schemes, the chromosome substitution and introgression. Long *et al.* (2000) tested for the contribution of polymorphisms in the 110 kb *achaete-scute* region to variation in bristle number, in a follow up to their 1990 paper. The experiment included 56 naturally occurring chromosomes. They confirmed their earlier results (MacKay and Langley 1990) that the presence of large transposon inserts, tested as a composite dummy variable, reduces bristle number. In addition they provided evidence for two indels affecting bristles, one of them in a sex specific manner. More recently Robin *et al.* (2002) tested for bristle QTL's at the *hairy* locus, both by chromosome substitution and introgression of alleles. Those were also tested over a *hairy* mutation and in all but one case a complex insertion deletion polymorphism in the upstream region affected bristle number highly significantly. It was the whole chromosome substitutions that failed to give a significant association while the direction of the effect was consistent with the results from the introgressed alleles. These results might be attributable to one of two things. First the introgression design tests only a small region of the chromosome and is therefore powerful enough to detect differences that are not large enough to be detected in the noisier experiment. Second, the effects of the QTN might be dependent on another factor (or

factors) on the third chromosome that were also segregating in the substitution panel. The introgression would neutralize the effect of any interactions with background. This study therefore demonstrates the increased power of the introgression scheme in contrast to the chromosome substitution scheme to detect polymorphisms of small effect.

For my thesis I opted for analysis of inbred lines over the other schemes described. Effects detected at a candidate locus with this approach are added up over genetic backgrounds and are therefore more likely to be relevant in nature, at least in species like *D. melanogaster* and *Homo sapiens sapiens* where individuals homozygous for large portions of the genome are rare. The method utilizes the laboratory conditions to reduce environmental sources of variation but also makes the association tests less powerful than more controlled genetic designs. Tests of association in inbred lines have been conducted in *Zea maize* with good results (Remington *et al.* 2001b), but have not been used in flies.

The protocol for genotyping has been a pilot study of nucleotide variation in the locus or region of interest, to establish the distribution of molecular variation in the locus LD and recombination parameter (Hudson 1987). The second phase involves scoring of a selected subset of variants in the larger study population. The association tests therefore rest on linkage between marker and QTN, in addition to magnitude and variance of the effect, frequency of marker and QTN and size of the sample (MacKay 1995, Long and Langley 1999). None of previous association studies in *Drosophila* sampled the complete genotype so that the association could be with a site or a linked QTN. In light of the rapid decay of LD in *Drosophila* then the results are consistent with the contribution of allelic variation in the surveyed loci to sensory bristle number. The next question after determining the involvement of specific loci should be to identify the exact polymorphisms causing the effect. That may only be achieved by full genotyping by sequencing. Long and Langley (1999 p730) stated that "Until further technological advances are made, it is likely that markers will not be discovered and typed at a density high enough to justify the assumption that one of the typed polymorphisms is likely to be causative". Following the genomic revolution and drop in cost of key reagents then it is safe to say that the methods are ripe for association tests with more extensive genotypes.

Theoretical considerations regarding association tests

The significance of associations will depend on the magnitude of the effect, the number of alleles surveyed, the accuracy of the phenotypic estimate and the frequency of markers and causative sites. Long and Langley (1999) assessed the power of regular association tests and compared them to haplotype based and Transmission Disequilibrium tests (TDT). Their simulations were tailored to mimic a continuous trait in a human pedigree, by a coalescence model (Hudson 1987). The results demonstrate that marker based permutation association

tests outperform TD and haplotype tests. The TDT tests are still important for analysis of human disease and can also be used to study cryptic variation in *Drosophila* (Ian Dworkin, Arnar Palsson, Kelli Birdsall and Greg Gibson submitted). Second, association tests have considerable power if 500 or more individuals are scored. Likewise adequate repeatability of associations will only be achieved with studies of similar proportions. Note that trait values in the simulation have larger error bars than estimates of phenotypic states in fly studies where 30-80 individuals are usually scored per genotype. Finally, Long and Langley (1999) show that more power is achieved by increasing sample size than the number of markers. They utilized coalescence to generate genotype matrices, but now several real genotype dataset are available (Clark *et al.* 1998, Remington *et al.* 2001b, Chapter 3) that incorporate the history of mutation, selection, drift and recombination in addition to the stochastic element. Such datasets should be used in a follow up simulation study into the relative power of our association testing procedures.

Population stratification

Population structure can complicate the identification of the heritable factors affecting continuous traits in natural populations and in worst case lead to erroneous results. If the distributions of a trait differ significantly between populations then any molecular markers varying in frequency between the populations can potentially yield a significant association. The pitfall of false positives due to population subdivision can be avoided by using family based association tests, like the Transmission Disequilibrium Test (TDT). The main drawbacks of family based tests are the need for high relatedness of subjects resulting in less statistical power in comparison to association tests. Tests of associations in samples from natural populations have now become feasible with molecular markers becoming more affordable and parallel extension of population genetics theory. As human geneticists are proposing whole genome association tests for major diseases then it becomes a priority to estimate structure in populations and try to account for its effects. Success of large scale association tests will rest on the ease by which population structure will be estimated and integrated into existing procedures.

Population stratification can be detected in the distribution of phenotypes and/or genetic markers. Distributions of quantitative or meristic traits are an indirect read-out of all segregating variants contributing to the trait in the study populations. An analysis of the trait values in study populations is commonly conducted with ANOVA or non-parametric methods. The design of the current study reduces the potential biases due to environmental or other systematic causes that in particular trouble studies in medical genetics. Here populations are represented by inbred lines derived from single fertilized females, and the individuals studied were all grown in a

common environment. The design effectively neutralizes the environmental component, and the analysis asks if the populations vary in genetic factors contributing to the trait (either directly or by interaction with the environment). Lack of evidence for phenotypic differences between populations reduces the false positive rate of association tests and greatly simplifies interpretation of the results. It must also be acknowledged that absence of phenotypic disparity does not rule out possible population distinct alleles or factors varying in frequency affecting the trait. Also the relationship between genetic structuring of populations and consequential phenotypic divergence is unknown.

Alternatively a direct comparison of genetic polymorphisms can also be used to evaluate the degree of structure and relatedness among populations. Analysis of the frequency spectra of segregating sites in a chromosome region or gene of interest has been the baseline mode of analysis. The obvious drawback to restricting the testing of population structure to the region under attention is the lack of information about the remainder of the genome. These regions could harbor substantial genetic differences as a result of the history of the populations, and if these differences impact the trait of study will also bias the results. This can be addressed by scoring additional markers across the genome to assess population structure. Relatedness of individuals can also be captured with F_{ST} which assess the significance of deviations in allele frequencies between populations (Weir and Hill 2002). Pritchard *et al.* (1999) devised a procedure to identify sub-populations by clustering subjects into groups by minimizing LD between unlinked markers and avoiding violations of Hardy Weinberg equilibrium. Recently reports describe the coupling of association tests and whole genome scans for population structure, using on the order of 40 to hundreds of micro-satellite markers (Ardlie *et al.* 2002. Remington *et al.* 2001b).

Embracing composite traits

Understanding of the genetics of complex traits may require unbiased exploration of the constituents of the phenotype in question and the pool of naturally segregating variants in the genome or a candidate region. Recent focus on the potential of genomic approaches to uncover natural polymorphisms affecting traits has triggered a shift in the way researchers tackle complex phenotypes (see for example Shimomura *et al.* 2001). The norm has been to find proximate variables that could be used in diagnostics of disease or for mapping. For instance the onset of Schizophrenia shows correlation to the function of dopaminergic neurons in the brain (Shastry 1999, Abi-Dargham *et al.* 2000, Harrison 2000) that could confer a quantifiable liability underlying the threshold character of mental disease. But are those neurogenic changes the biological cause, part of the disease manifestation, or a secondary symptom? Comprehensive understanding of complex diseases like Schizophrenia and cancer may require

a more exploratory approach to the phenotypic space, a strategy that builds on classical epidemiology, pharmacology and the brute force mindset of genomics (Scherf *et al.* 2000). The philosophy is that seemingly minor pieces of information may be unknowingly useful, and the challenge has become extracting attributes relevant for the biological problem at hand. A parallel argument can be made for why evolutionary geneticists should embrace the full dimensionality of trait space. Organisms are without exception composite beings and understanding the relationships between individual constituents of form or phenotype will elucidate lineage modification and diversification. Focus on the genetics of a single trait ignores both the evolutionary context of pleiotropic polymorphisms and relation of the trait to other features of the developing organism. Moreover fitness, the most composite of all traits, summarizes the effects of segregating alleles in a population as they manifest in many traits of an organism. Unbiased exploration of the phenotypic space will uncover the common axis of variation that can indicate developmental constraints, mutational biases and shared genetic components (Stoll *et al.* 2001).

Several evolutionary and medical related studies have attempted to embrace the phenotypic space. Natural differences in kinetics of metabolic enzymes have been studied (Clark and Wang 1997). Likewise modern epidemiology has stepped up its scale of investigation. For example the dissection of cardiac factors in rats (Stoll *et al.* 2001). Morphologies of adult individuals have several advantages over other complex phenotypes. First the heritability of morphological traits is generally higher than for life-history or fitness traits (Roff 1997), which is helpful for mapping purposes. Furthermore adult morphologies are generally easier in quantification, which leads to better estimates of heritability. Moreover *D. melanogaster* has been a useful model organism, and with the complete genomic sequence, growing P-element mutant collection aimed at saturating the genome, and homologous recombination (Celniker 2000) added to the experimental repertoire, its utility for evolutionary genetics is only going to increase in the next decade.

B. Wing development

Waddington (1940) urged that genetic studies of developmental processes rest on two pillars. One is the general ease of genetic manipulation in the organism, especially availability of mutants affecting the trait. The second pillar is the necessity of building a comprehensive picture of the developmental mechanism prior to genetic or environmental perturbation. Studies on the development of the wing of *Drosophila melanogaster*, pioneered by Waddington, are examples of where establishment of both pillars have triggered an enormously successful research program. Progress is fueled by novel tools to study and manipulate gene expression as well as more cell biological and biochemical analysis of the process. In the following discussion I will start by describing mechanistically the wing development, from the larval imaginal disc to full structures. A suite of genes is known to affect the process, many of which have mammalian orthologs. The wing has emerged as a great system for study of those genes and more specifically their functional relationships. These are, not unsurprisingly, remarkably conserved between insects and vertebrates. Here particular attention will be devoted to genes contributing to vein placement and differentiation, especially components of the EGF/Ras pathway.

Drosophila wings are composed of two layers of cells, dorsal and ventral. Their configuration is sustained by rigid tube-like structures called veins and the wing margin. Fruitfly wings originate from early larval structures called the imaginal discs and reach full form after metamorphosis with the aid of a biological hydraulics mechanism. Imaginal discs are clusters of cells that are defined early in larval development and will each give rise to certain parts of the adult epidermis. For instance the *Drosophila* anterior (head and thorax) is made of 20 imaginal discs, including the antennal – eye disc, labial disc and leg discs, while the posterior arises from the genital disc and cells of the imaginal histoblast nests. The distinction between those two epidermal precursor cell types is based on cell number in third instar larvae, where imaginal discs consist of up to 50,000 cells (wing and antennal- eye discs being largest) while histoblast nests contain 6-15 cells (Fristrom *et al.* 1993). At this stage the disc has undergone considerable patterning in addition to growth, as demonstrated by fate mapping experiments (Bryant 1978). Waddington (1940) recognized the major stages of wing development. The wing first develops as a sac-like structure inside the larvae, where a field of cells at one side of the sac gives rise to the wing itself, the rest to the thoracic surface. The wing forms when the structure shoots into the sac, making a blade proper. The two surfaces (dorsal and ventral) are separated and then realign during early pupal development. A major transition is when the wing is transported to the pupal exterior with a process called evagination. Further refinement takes place through the later pupal stages but most spectacularly the wing inflates shortly after eclosion and only then unfolds to give the adult form. As my interest is in the continuous

variation in shape of the wing then it is important note that the details of wing imaginal disc development differ in numerous ways between *Drosophila* and other insects (Waddington 1940). To my knowledge a systematic documentation of these differences has not been attempted. It could however provide valuable insights into the developmental and genetic constraints on wing development. Coupling of these observations and the extensive information *Drosophila* wing development could pose exact hypothesis about a subset of loci harboring segregating variation affecting shape.

During embryo patterning certain cells are selected to give rise to the imaginal discs. This depends on the function of morphogenic genes like *wingless* (*wg*), *hedgehog* (*hh*) and *decapentaplegic* (*dpp*) to name a few. Morphogens trigger cellular responses in a concentration dependent manner and detailed examples will be given below in the context of wing development. Further refinement of the wing imaginal disc classifies cells destined to become the wing itself, compose the hinge region and numerous thoracic structures. A wealth of research on wing development has focused on events occurring after this initial patterning but recent efforts are concentrating on those early stages (Zecca and Struhl 2002). The processes described below occur during late 3rd instar larvae stage and the first 36 hours of pupariation[1].

Initial development and patterning

The wing imaginal disc is a single layer of epithelial cells connected to the larval epithelium as an invaginated sac. These cells have apical-basal polarity with the apical surface facing into the sac and the basal side resting on the basal membrane. The sac of third instar larvae has two types of cells, columnar and squamous, forming distinct sides of the sac, the posterior part being rich in columnar cells. Waddington (1940) recognized this thick bundle of cells as the wing primordia that later change form to make the wing proper. It is worth stressing that the wing primordia is a single layer of cells at this stage, only later will it take the form of folded epithelium. Columnar cells are 30 times taller then they are wide which becomes important when the field of cells expands and takes shape. The squamous cells resemble regular epithelia and are sometimes called the peripodial epithelia. Both cell types will give rise to the adult cuticle (Fristrom *et al.* 1993). Cell fate experiments show that in 3rd instar larvae the fates of cells are determined. Naturally wing cells are separate from other disc cells, but more importantly the two main axes of the adult wing, dorsal-ventral and anterior-posterior have been laid down. Understanding how this is achieved requires a closer look at the disc in 3rd instar larvae. The cells of the prospective wing form a circular field that is divided into quadrants on

[1] There are 5 main Drosophila developmental stages. Embryogenesis is the construction of a functional larvae from a fertilized egg (takes 24 hours). There are 3 larvae stages, named 1 – 3 instar larvae each taking 24 hours, except the 3rd instar which takes 2-3 days. The pupae stage takes 4 days and involves radical rearrangement of the animal to give an adult fly.

the basis of expression of two genes, *engrailed* (*en*) and *apterous* (*ap*). The dorsal part of the wing is defined by the expression of *ap* and the posterior part of the disc is defined by *en* expression, thus providing "selective" information for cells of each quadrant. Both genes are examples of what are regularly dubbed selector genes (Garcia-Bellido 1975), and they also pattern the non-wing parts of the disc. The details of how *en* expression triggers the venation patterning via the morphogens *hedgehog* and *dpp* will be discussed later. Meanwhile a review of how *ap* determines dorsal-ventral fate serves to introduce the dynamics and challenges of wing development.

Dorsal-ventral compartmentalization

The multiplicity of gene action and developmental events required for wing development is exemplified by the way *ap* expression assigns dorsal fate to the respective cells. The locus is first transcribed, in response to EGFR signaling (Wang *et al.* 2000), in the wing disc in 2nd instar larvae as a part of the wing compartmentalization. A great deal of controversy has been over how *ap* assigns dorsal identity, but two main theories have been put forward. First, re-aggregation experiments show that cells of ventral and dorsal surfaces have different adhesive properties and sort out when mixed. This suggests that organizing genes induce differential expression of cell adhesion molecules leading to assortment of cellular populations. Second, loss of *ap* expression results in a remnants of a wing, suggesting that signaling between compartments is required for growth and patterning. Genetic experiments by O'Keefe and Thomas (2001) suggest that both theories may be relevant, but the regulative role of *ap* is better supported. They rescued *ap* null-mutants by coexpressing *fringe* (a component of the *Notch-Delta* lateral inhibition pathway) and integrin chain α_{PS1} (required for cell adhesion) in the dorsal part of the wing. Astonishingly the resulting wing had perfect vein arrangements and proportions resembling regular wings, but both surfaces had characteristic ventral features, seen in bristle arrangements and vein bulging. This independence of patterning from dorsal identity highlights the importance of signaling between compartments for dorsal-ventral distinction without undermining the role of cell-adhesion molecules. O'Keefe and Thomas also provided evidence for a role of *ap* in initiating the *Notch* lateral inhibition pathway via *fringe*, leading to the localized *Notch* expression along the dorsal-ventral boundary. Ultimately this leads to expression of *wg* in the dorsal-ventral boundary cells driving differentiation, proliferation and wing outgrowth. A direct relationship between cell adhesion genes and *ap* has not been established. The study leaves open the question of how *ap* mediates dorsal identity. Currently the best candidate is a locus called *Dorsal wing*, which in mutated form surrenders the dorsal surface to ventral fate. Likewise ventral misexpression of *Dorsal wing* is enough to give dorsal like structures (Tiong *et*

al. 1995). For the analysis of wing shape these results suggest that shape is regulated by a distinct panel of loci with little overlap with loci involved in initial patterning and cell adhesion.

Evagination, elongation and eversion

Before the particulars of vein placement and differentiation are articulated, it is essential to consider the morphogenetic transformation of the wing imaginal disc from a single layer of columnar cells into a two layered wing blade. Waddington (1940) was the first to describe the phenomena in any detail but a more recent review by Fristrom *et al.* (1993) provides a more comprehensive picture. The process is called evagination and includes extension of the limb and its translocation from the protected larval interior to the surface of the developing adult. While the processes of elongation and eversion take place nearly simultaneously, a clear distinction can be made between them.

Elongation of the wing starts about 6 hours (hr) prior to pupariation in response to a rise in ecdysone hormone, and lasts about 12 hr with the most rapid elongation after pupariation. Looking at the wing disc in 3rd instar larvae, elongation occurs along the dorsal-ventral boundary of the wing primordia. Essentially, *wg* expressing cells at the prospective margin extend out of the plane, adding the 3rd dimension to the picture. Condic *et al.* (1990) have also observed that cells in the center of the imaginal disc of 3rd instar larvae are markedly smaller than those at the periphery. These cells will compose the distal part of the wing and this arrangement facilitates elongation of the appendage. Both unfolding of the epithelia and cell shape changes play a role in the extension of the blade but only account for parts of these radical shape changes (Fristrom *et al.* 1993). Elongation requires detachment of the wing cells from the basal lamina, and during the rest of wing development the membrane attachments of these cells will be in flux as discussed below. The most distal parts of the future wing are free of connections to basal matrix components for a couple of hours while microtubules build out into the structure. As folding proceeds then a looser form of extracellular matrix starts too accumulate between the regular basal lamina and basal side of imaginal disc cells (Brower *et al.* 1987). While considering the elongated wing blade and the basal lamina it is worth noting that cytoplasmic processes of some cells can extend along the basal lamina of wing discs (Fristrom *et al.* 1993). Ramirez-Weber and T. B. Kornberg (1999) gave the phenomena the name cytonemes after a more careful study and postulate its importance for mediation of long-range signals in a field of growing cells. Sadly the fragility of cytonemes makes them refractory to experimental dissection, complicating assessment of their biological importance. The fact that they have also been observed in other species, along with the potential range of the cellular extensions, suggests that cytonemic involvement in mediation of long range signaling must be considered a possibility.

Simultaneous with elongation, changes take place in the peripodial cells leading to eversion of the wing. During the first hours of pupariation these cells accommodate the elongation of the wing proper by stretching and flattening. But by the 4th hour contraction begins, and the area of the peripodial epithelia decreases as the cells take columnar shape. In conjunction the imaginal disc sac begins to open up to the surface of the pre-pupae. The combination of elongating appendage, contracting sac, and open escape route, forces the baby wing blade to the exterior of the animal. This is completed by 4 hour of pupariation after the successful detachment of larval epithelia and the pupal case has provided a safe environment for the new limb. It is commonly assumed that evagination affects neither the patterning nor shaping of the appendage (Milner *et al.* 1983, Fristrom *et al.* 1993). Experimental verification of this assumption is not available nor has the importance of epithelium folding on wing development been addressed. While cell shape changes are crucial for many events of evagination, most obviously for the folding of ventral and dorsal surfaces, the debate is still as to whether folding of the true imaginal disc is a function of limited physical space or elaborate genetic control. This classical perception of the wing disc as a two dimensional field has been also been challenged for the model of the wing and leg disc epithelial sac. Experimental study of *dpp* signaling showed that there is signaling between the two surfaces, the squamous and columnar (wing) cells (Gibson *et al.* 2001). The messenger molecular, in this case *dpp*, was excreted into the sac and mediated responses on the columnar surface. In light of these results, the processes of folding and evagination could potentially be important for the patterning of the wing or other *Drosophila* structures derived in a similar manner.

Synchronization by Ecdysone

The series of events described above are assumed to be synchronized in part by changes in ecdysone levels. Insect metamorphosis is driven by the steroid hormone ecdysone and imaginal disc morphogenesis is absolutely dependent on its function. The pattern of ecdysone level changes and correlated changes in the wing during metamorphosis is summarized in Table 1.1.

The ecdysone level first rises 6 hours prior to pupariation as a consequence of a neurosecretory signal and then declines. During metamorphosis 3 other ecdysone pulses have been observed, at hours 10, 18 and 80 hours after pupariation (the adult fly hatches around hour 96: Bainbridge and Bownes 1988). Elongation of the wing continues after the first pulse and the hour 10 pulse coincides with full bloating of the pre-wing. Subsequently cell division resumes and wing morphogenesis is completed by hour 38-40. The hour 18 pulse maintains a 40 hour high ecdysone level in the pupae that is required for cuticle formation. During the first 20 hr cells are free to move and change shape before the adult cuticle hardens. The wing cells aggregate cuticle molecules but retain flexibility as can be seen during the unfolding of the wing

after eclosion. It could be reasoned that metabolism of cuticle formation would affect shape of individual cells or the whole wing but present lack of information prevents any statements on the matter.

In conclusion, ecdysone surfaces as a non-specific coordinator of events during metamorphosis. It remains to be seen if imaginal discs notify their readiness to receive the signal or if ecdysone pulses are given regardless of tissue susceptibility. Likewise it is not obvious how ecdysone triggers the dramatic events seen during wing eversion but it is obvious that radical cell-shape and basal matrix changes coincide with the events. Current genomic analysis of ecdysone responses are thus unlikely to shed light on the eversion phenomena. However they may prove valuable for analysis of how standing levels of ecdysone induce further refinements of adult structures by cell division and differentiation (White *et al.* 1999). The further morphogenetic refinements of the wing blade during later stages of pupal development will be discussed in the context of vein development.

Table 1.1 Timing and concentrations of ecdysone pulses during wing development.

Time (at 25°C)	Ecdysone levels	Wing events	Pupal stage
24 hr – mid 3rd instar	Basal		
~ - 6 hr	⇑ titer		
~ - 3 hr	Remains high	Appendage elongates (most cells in G2 arrest)	
0 hr pupariation	- " -		P1
2 hr	⇓ titer		P2
4 hr	Basal	Appendage elongates and evert	P3
8 hr	- " -	Continuous thoracic adult epidermis forms (most cells in G1 arrest)	
10 hr	Slight rise and fall		
12 hr - head emergence - true pupation	Basal	Wings bloated and elongated	P4
14 hr	- " -	Cell division resumes	
18 hr	⇑ titer		
20 hr	Remains high		
24 hr	- " -	Refinement of structure	
36 hr	- " -		
40 hr	- " -		
96 hr – Eclosion	The third ecdysone pulse comes at 80 hr, with no consequence for wing development	The wing surfaces separate just prior to eclosion, ready for inflation of the wing	

Vein development and differentiation

The initiation of wing veins can be traced back to events in the imaginal disc of 2nd instar larvae, when expression of *en* distinguishes the posterior and anterior compartments. Numerous papers covering this topic have been published in the past 25 years and the aim of this section is to briefly summarize the results to date. A more comprehensive review is provided by Held (2002). Prior to presenting the molecular and genetic factors involved, a closer look at the final product is appropriate. The adult wing has 5 classified longitudal veins, named L1 through L5 (Figure 1.1). Classical nomenclature of entomologists is different and was designed to assist comparisons of veins across taxa, see Figure 2.1 in Chapter 2. In addition to those 5 veins two crossveins are present in the wing, conveniently named anterior (1) and posterior (2) crossvein. As earlier suggested, the veins do not sit in the middle of the cellular field, but bulge either dorsally or ventrally. The bulging of some veins also switches surfaces as they reach the distal part of the wing.

Veins L2 and the proximal part of L4 bulge on the ventral surface while veins L3, L5 and distal part of L4 bulge on the dorsal surface. Both crossveins bulge dorsally. Wing veins across class Insecta exhibit bulging to some extent, but there is also considerable diversity in the vein structures themselves. The simplest forms are invaginations or condensations in the intervein tissue while the more complex forms, including those of *D. melanogaster* are hollow tubes made up of distinct cell types. Two additional anatomical properties should be considered. First the veins are the only living parts of the wing in adult individuals. They provide nutrients and oxygen for the sensory bristles located on the wing margin, and on vein L3. An interesting extension of this role is seen in color-winged *Drosophila's* where the veins transport the pigment precursor to the wing (True *et al.* 1999). Secondly, veins grant support to the wing and their location and rigidity may impact aerodynamic properties of the appendage (Dickinson *et al.* 1999).

An important feature of the compartmentalization of the wing blade is clonal restriction. This term refers to the observation that cells belonging to a particular intervein region will not escape the region and neither will its descendants. The governing rules are such that vein and intervein cells do not mix, and neither do cells of different intervein regions. Held (2002) argues that this may in part be mediated by differences in cell adhesion properties. Clonal restrictions and synchronious cell divisions led Garcia-Bellido and de Celis (1992) to propose the "Entelechia" model to account for how cellular populations realize developmental coherence. The cells of intervein regions progress through cycles of cell divisions and later apoptosis to fulfill the structure, by a combination of positional information and interaction between cells that aid them to interpret the global environment. The best candidates for this positional information are the EGFR pathway and Serum Response Factor, as judged by the behavior of mutant

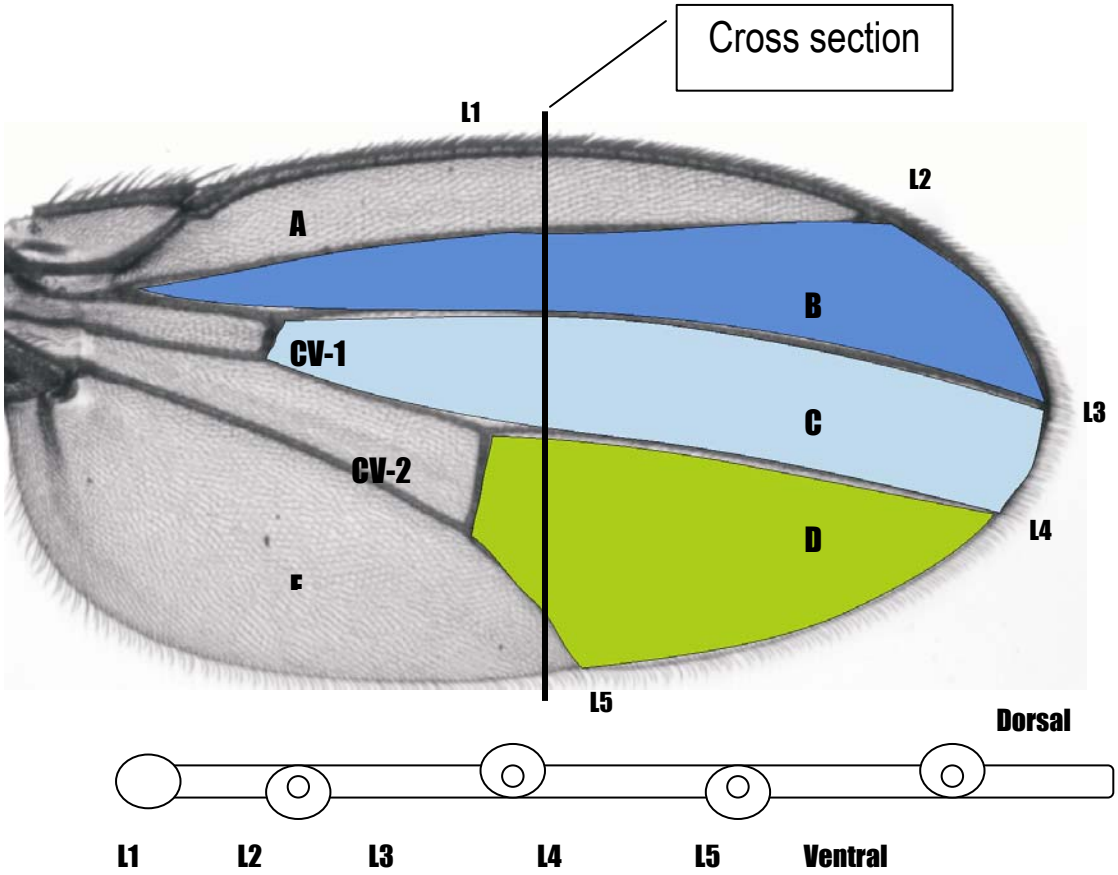


Figure 1.1 Venation patterns in the *Drosophila* wing, with anterior side facing up and posterior down. The five major veins are L1-L5, and the two crossveins (CV 1-2) are implicated and as are the 5 major intervein regions (A-E). The lower cartoon demonstrates the characteristic bulging of veins on the wing surfaces.

clones of key loci compared to their twin clones in the disc (Resino, *et al.* 2002). A distinct drawback of the model is that it harbors no information about the realization of shape. Simple mathematical models of development can generate stripes (von Dassow *et al.* 2000), dots (Meir *et al.* 2002) and Dicteostelium-like structures (Hogeweg 2000) but nothing similar in complexity to wings has been realized. Salazar-Ciudad and colleges (2000, 2001a, 2001b) have explored the topological space of small network modules in a field of cells of known dimensions. The interaction events can either happen internally or by connections between cells. The properties they investigated were sensitivity to initial conditions, effects of size, and buffering capacity of the modules. In the light of our discussion of wing development then it is interesting that only a subset of network topologies exhibited size dependence. A comparison of network topologies, for instance between the networks documented in the wing and the ones constructed by Salazar-Ciudad *et al.* is not meaningful for small networks because the standard descriptors need large datasets (Jeong *et al.* 2000). Given the wealth of molecular information on gene action in wing development then a bottom up modeling of wing shape may be due.

Vein morphogenesis

Following eversion the wing undergoes further changes in form, starting with extensive flattening. The basal sides of the dorsal and ventral surfaces come into proximity and start producing the novel basal lamina of wings. Coinciding with the start of the true pupal stage the structure inflates, separating the two surfaces. The ballooned wing is maintained in this form for 6 hours, until the 18th hour, when an ecdysone pulse facilitates it's flattening. Waddington (1940) noted a pattern to the wing contraction, with the cells of the wing margin fusing to give a "hollow sac with a thickened seam". In the following hour the sac collapses, from the distal and later the proximal end, leaving vesicles in the middle. During the next eight hours venation appears. L3, L4, L5 appear as persisting cavities while L2 appears to grow from the proximal part. From hour 28 to 45 the wing later expands and the veins become further refined and nerve axons and trachea extend along the veins. It appears that the final structure was being refined and ready to go but those familiar to *Drosophila* know the wings have different appearance in freshly hatched individuals. For the period of the last 30 hours of the pupal stage the wing is compacted by disassociating the dorsal and ventral surfaces and elaborate folding patterns. After eclosion the inflation and contraction has to be repeated to bring about normal looking wings.

Vein and intervein cells are distinguishable early on the basis of phases of the cell cycle. Fain and Stevens (1982) showed that prepupal cells are mainly arrested in G2 after DNA replication in larval stages. Intriguingly cells of the wing disc enter G2 in specific order in 3rd instar larvae. Cells of the wing margin and along the L3 vein primordia, roughly corresponding

to the anterior-posterior boundary, are the first to enter G2 arrest. Note that these cells are on the boundary between anterior and posterior part of the wing, as marked by expression of signaling molecules. Following pupariation, cells corresponding to veins L2 and L4 also stop replicating and when divisions and further replication resumes in the wing following the prepupae period it follows the same sequence, suggesting the importance of those structures for wing development. These patterns highlight the role of the organizing centers, the wing margin and anterior-posterior boundary, in progression of the developmental system in the wing. Studies of the effects of vein mutants on cell division and cycling in the wing imply that cell division is dependent on vein primordia (Diaz-Benjumea *et al.* 1989, Diaz-Benjumea and Garcia-Bellido 1990). A progressive pattern of cell divisions from veins into the inter-vein field can be interpreted as if the veins provide a cell-division inducing signal.

Anterior-posterior boundary establishment

The previous section outlined the molecular mechanism of dorsal-ventral compartmentalization and now I proceed to a similar description of the anterior-posterior patterning. The first known molecular signal that polarizes the wing blade is expression of *en*, which is exclusive to the presumptive posterior side of the imaginal disc[4]. Engrailed is a transcription factor of the homeodomain family and its expression is sufficient to specify a distinct lineage of posterior cells, unable to intermix with anterior cells (Vincent 1998). Establishment of the anterior-posterior boundary by engrailed is mediated by its activation of *hh* expression. The direction of signaling is achieved by the inability of *en* expressing cells to respond to hedgehog, a classic endocrine signal (also termed “for-export-only” by Bier 2000). Only cells just anterior to the boundary are able to respond to hedgehog, taking the role of an “anterior-posterior organizer” which controls the patterning and positioning of veins. The morphogen hedgehog has a fundamental role, by activating expression of several target genes, including *patched* (*ptc*), which encodes a subunit of the hh receptor; *fused* (*fu*), a kinase affecting hh signaling; and *collier* (*col*), which is a transcription factor. Hedgehog thus modulates its own signaling in addition to eliciting cellular changes, a mechanism also exhibited by the other signaling pathways including EGFR/Ras. Hedgehog proceeds to recruit the TGFα and EGF signaling pathways to vein development, by activating the expression of their ligands, *dpp* and *vein* respectively. Full appreciation of the initial placement of the veins demands an involved examination of the molecular and cellular specifics of Hh mediated signaling.

[4] Engrailed provides posterior identity to the whole disc, not only the wing blade.

Hedgehog defines the central organizing region

A primary concern here, as in many developmental studies, is how a morphogen gradient is interpreted. Extensive work on *hh* signaling portrays two features as responsible for turning a continuous signal into discrete decisions. First, *hh* signaling leads to nuclear localization of the heavier form (155 kD) of the transcription factor Cubitus-interruptus (Ci-155). The normal form of the protein is bound in cytoplasmic complexes associated with microtubules. In the absence of Hh, protein kinase A (PKA) mediates proteolysis of the protein yielding a truncated form (Ci-75) that localizes to the nucleus. This truncated form cannot activate Hh target genes and some evidence suggests it actively represses their expression. The sharp drop in Dpp concentration more anteriorly can be contributed to the different effects of Ci-155 and Ci-75 on *dpp* expression. The second parameter that helps refine the responses to Hh is a kinase encoded by *fused*. Loss of *fu* causes cytoplasmic retention of Ci-155, arguing for its role in facilitating nuclear localization of the active form of the transcription factor. Studies have also shown that Fu is actived by phosphorylation, but only in response to high levels of Hh. The consequence is a second threshold where sharp differences in Ci-155 nuclear levels can be observed. Concentration differences of nuclear Ci-155 in those two fields of cells leads to differential expression of Hh target genes, where *ptc*, *col* and *en* are expressed at high Hh levels but *dpp* in response to lower concentration. Those cells are the anterior-posterior organizer that will guide vein development. This is an active field of research and the molecular details of this machinery are being worked out (see Held 2002 for full review). The main conclusion regarding *hh* signaling pathway is that the molecular complexity of the pathway enables a continuous morphogen gradient to be translated into at least three discrete cellular states. The integrity of these distinct cell populations is further increased by feedback loops, as the two following examples demonstrate. First, as mentioned earlier hh activates the expression of Patched, an important part of its receptor. This makes hh responding cells even more responsive, generating positive feedback iterating the developmental decisions taken in the cells. The second example concerns the posterior boundary of dpp expression. Recall that Hh signaling turns on expression of *dpp*, *col* and *en*. The refinement of *dpp* expression is mediated by the transcription factors Col and En, which both repress *dpp* transcription in cells receiving high dose of Hh. Thus only cells seeing intermediate levels of Hh will express Dpp, with the posterior boundary enforced by *col* and *en*, and the anterior one by the negative form of Ci (Ci-75) as previously described. Hh also affects dpp signaling by down-regulating the dpp receptor *thickveins* (*tkv*) through the product of the *master of thickveins* (*mtv*) gene. The molecular function of *mtv* is still a puzzle but it harbors motifs common to transcription factors and nuclear proteins suggesting a direct regulatory role (Funakoshi *et al.* 2001). In brief, Hh acts via

complex intracellular machinery and feedback-loops to define clear cell populations that will later give rise to either vein or intervein tissues.

Besides *dpp*, the second key player in vein differentiation is the EGF/Ras pathway, and recent studies have shown that Hh also mediates its activities. It has been known for a while that Hh activates transcription of *vein* (*vn*), a locus encoding a ligand of the EGF family. However only cells experiencing low Hh can respond to the signal. This is manifested by mutations in the fused kinase that facilitates Hh signaling. Loss of fused function leads to shrinking of the width of intervein region C, in extreme cases leading to the fusion of veins L3 and L4. The molecular mediator of Hh dependent repression turns out to be Col, which renders the cells of the A-P organizer insensitive to Vn levels by repressing *EGF-receptor* transcription. Loss of *col* leads to ablation of intervein region C, which separates veins L3 and L4, giving the appearance of the two veins intertwining hence the older name for *collier*, *knot*. This for-export-only signal guarantees that only cells outside the A-P organizer can respond to Vn and start differentiating as veins. Note that mutations in two loci involved in Hh signaling, *fused* and *knot*, have serious shape change phenotypes. Furthermore, increase in hedgehog signaling by elevating the cofactor cholesterol causes displacement of veins in a panel of wild type lines (Birdsall *et al.* 2000). It is therefore an interesting candidate for wing shape.

Beyond veins L3 and L4

While the tale of the A-P organizer and L3 and L4 vein determination is still unfolding, combined with our understanding of the role of the EGF/Ras pathway, we have a compelling picture of vein determination. Before I describe the latter pathway in more detail, the remaining parts of the vein determination picture should be assembled. It is markedly patchier then our understanding of the establishment of the A-P boundary and L3 and L4 identity but several generalities can still be appreciated.

Of the genes discussed previously only *rho* can be regarded as pure venation gene in the sense that it is only expressed in vein primordia. Another universal feature of vein primordia is the absence of *blistered* (*bs*) activity. *bs* encodes the *Drosophila* Serum Response Factor homologue and is required for development of the intervein tissues. Other genes show specific expression patterns in the primitive veins: *Delta* is for instance found in all veins except L2. Likewise, the genes *caupolican* and *araucan* are only expressed in odd numbered veins suggesting their dependence on dorsal specific signals since L3 and L5 bulge out dorsally. The presence of sensory organs on vein L3 similarly explains the restricted expression of the proneural genes *achaete* and *scute* in L3 primordia. Those genes serve as markers for identifying these vein primordia in the developing disc but are considered determinants of positioning. While L3 and L4 locations may depend on Hh, dpp and EGF signaling, it is not clear

how the locations of the two remaining veins, L2 and L5, is established. L5 development may depend on the function of *abrupt* but experimental dissection is still in the pipeline, leaving the question open. The picture of L2 vein initiation is a bit clearer, providing a connection to the A-P organizer. Long-range effects of Hh are mediated by dpp, which diffuses anteriorly and triggers expression of several genes including one of its own receptor *thick veins* (*tkv*). In the context of L2 placement the target *spalt major* (*salm*) is of key importance, for the vein forms along the anterior border of its expression domain. *salm* is a transcription factor expressed in a broad domain, corresponding to intervein region B (it is also expressed in IVR D), suggesting its role is similar to that for *col* in intervein region C. There is genetic evidence for *salm* acting through the two related neighboring genes *knirps* (*kni*) and *knirps-related* (*knrl*) to induce L2 vein differentiation. The enhancer of *kni* was recently dissected experimentally and a module capable of driving expression corresponding to the L2 provein region characterized. Ectopic expression of *EGFR* and *vein* by this promoter led to extra vein formation around L2 and L5 (Lunde *et al.* 2003). It is also interesting that *kni* and *knrl* encode proteins of a steroid-hormone receptor family stressing that in the evolution of fly wings numerous pathways have been adopted to pattern the structure.

Here I have only mentioned a fraction of the 400 loci known to impact wing morphology (as summarized on www.Flybase.org, the internet server for *Drosophila* genetics and biology). This number is an underestimate as recent a microarray study uncovered 50 previously uncharacterized genes as being upregulated in the wing vs. the surrounding imaginal disc tissue (Butler *et al.* 2003). The patterning and differentiation of tissues are developmental events that are not always easy to distinguish. This holds for fruitfly wings even with the moderate complexity of the structure and only two major cell types, the vein and intervein tissues[5]. Previous sections have mainly focused on the patterning of the wing, down to the placement of veins, but now questions about cellular differentiation become important. The vein cells in particular must undergo considerable modifications to become rod-like support structures. The explicit set of loci required for vein differentiation has not been characterized but it is apparent that three main pathways, represented by the canonical members, *EGFR*, *dpp* and *Notch* are required. Similarly, the intervein tissue requires at least the *Drosophila* Serum response factor (SRF) homolog, encoded by *blistered* (Montagne *et al.* 1996) and later *EGFR.* As the thesis concentrates on analysis of the effects of polymorphisms in *EGFR* on wing shape, the focus will be kept on vein formation, and the dual role of *EGFR* in patterning of veins and realization of intervein tissue.

[5] There are other cell types in the wing, primarily extensions of the nervous system (bristles and connecting axons) or cells of the hemolymph (that permeates along veins).

The canonical EGF/Ras pathway

Studies on human cancers and cell culture as well as genetic dissection of eye development in *Drosophila* and vulval patterning in *C. elegans* converged on a pathway that is currently one of the most extensively studied signal transduction cascades. The general components of the pathway are as follows; a membrane bound receptor with tyrosine kinase activity, src-homology proteins that mediate the signal, the G-protein complex of which Ras is the most famous member, and a series of three Map kinases of which the most downstream one localizes to the nucleus and mediates the signal. A schematic summarizing the core part the cascade is represented in Figure 1.2.

Signaling through the pathway is initiated by ligand binding to a receptor leading to dimerization of the receptors. Analysis of the crystal structure of the Human EGFR suggests this is mediated solely by interactions of the receptors (Ogiso et al. 2002). Dimerization is followed by auto- and trans-phosphorilation of the intracellular domains of the receptor, making them attractive binding sites for docking proteins and signaling mediators like SHC and Src. Those in turn activate the GTPase Ras which phosphorylates the MAPKKK triggering phosphorilation of MAPKK and MAPK. The most downstream kinase then transduces the signal, by acting on other cytosolic proteins or by being transported into the nucleus where it affects transcription (see general reviews, Lewin 1997, Held 2002). While the EGF/Ras pathway is very well conserved in arrangement between eukaryotes, a marked difference in the level of pathway complexity can be seen, for instance between flies and humans. It has been known since 1998 that the human cascades has four receptors (Olayioye *et al.* 2000), while flies have only one, and 8 different ligands have been isolated from humans while only 5 are known in flies. The numbers of G-proteins, phosphatases and kinases that mediate and modify the signal are a lot less accurate. A fuller contrast between the human and *Drosophila* Ras pathways can only be achieved once the genomic sequences have been mined and the findings experimentally substantiated. For instance a fifth ligand for *Drosophila* EGFR was recently described and named *keren* (also known as *gritz*). Similarly, there are six *rhomboid* genes in the genome that appear to divide the function of cleaving the EGF-ligands throughout development (Wasserman *et al.* 2000).

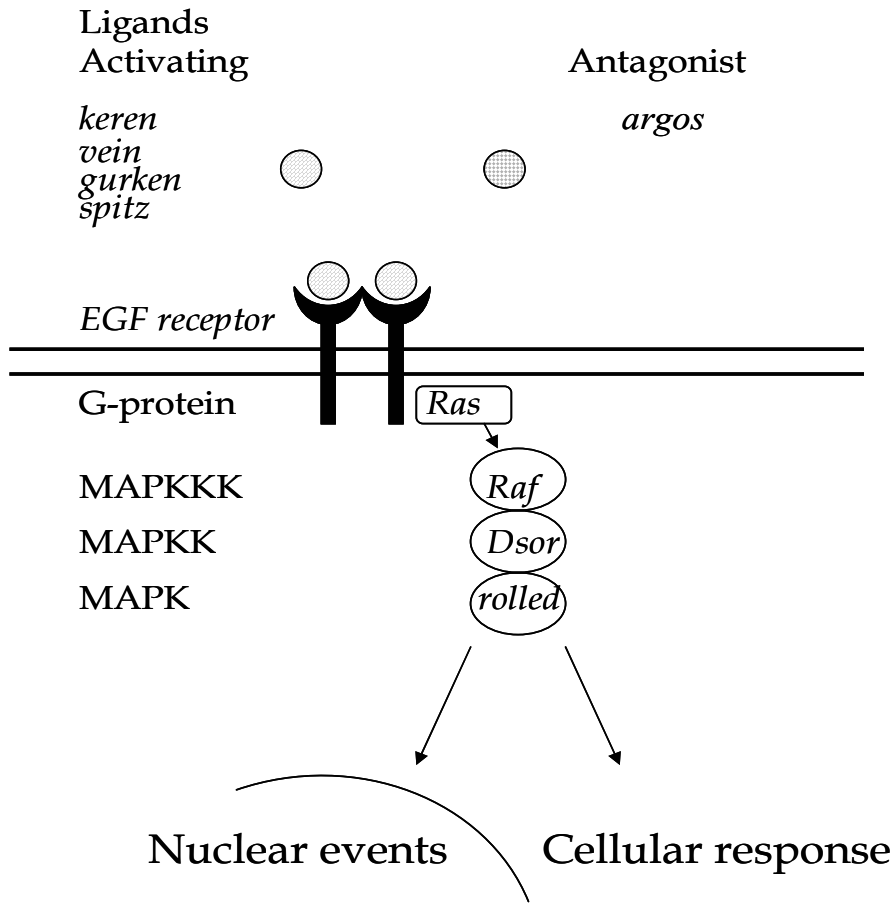


Figure 1.2 Core elements of the EGFR pathway in flies. There are four activating ligands, encoded by the loci *keren, vein, gurken* and *spitz*, and one antagonist *argos.* EGF receptor binds the ligands and dimerizes. Signal is transduced by docking molecules (not shown) to Ras. Ras in turn activates the MAPK cascade encoded by *Raf, Dsor* and *rolled.* The last kinase then mediates the signal, either by translocating to the nucleus or by promoting cytosolic events.

That explains how loss of the canonical *rhomboid* gene in *D. melanogaster* has only effects on subset of the tissues where EGFR is required. The *rhomboid* genes were first described in *Drosophila* and have later been discovered in humans and are conserved in bacteria as well. The conservation the bacterial rhomboid was demonstrated by its capacity to complement a loss of the *Drosophila* locus. The role of the *rhomboid* family proteins in mammalian systems awaits experimental confirmation (Shilo 2003). Two other members of the core pathway are represented by multiple loci. The Sevenless receptor can substitute for EGFR and does so in subset of cells during eye-development. No further roles for Sevenless have been characterized, and the question is still open about the evolutionary history of the locus. Second, the G-proteins are encoded by three *Ras* genes. They all appear important as the proteins are all invariant in a sample of 27 *D. melanogaster* alleles. The absence of protein polymorphism is at odds with low to normal levels of synonymous changes and polymorphisms in the non-coding regions (Gasperini and Gibson 1999). Other core MAPK components are represented by single *Drosophila* genes and seem to be experiencing purifying selection to a comparable extent as the Ras genes (Riley *et al.* 2003). Determination if auxiliary components are orthologous between humans and flies is complicated by rapid sequence divergence. The fact remains that the EGF/Ras pathway operates as a robust signaling cascade that has undergone considerable divergence in function and composition during evolution. Two modes of evolutionary analysis of this or comparable cascades might be very interesting. A full phylogenetic analysis of all components of the pathway is long overdue. Second, a study of genetic variation of components of this pathway across the tree of life may prove to be very insightful.

The biological functions of EGFR signaling are almost too numerous to list. Both EGFR and Ras were first identified as oncogenes in humans; that is activated forms of the loci correlated with onset of cancer. Consequent interest in the pathway led to the determination of the biological roles of EGFR across the animal kingdom. Specifically, in *Drosophila* several roles in have been attributed to the cascade, including cell growth, survival, differentiation, migration (and other shape changes) and patterning of tissues (Held 2002, Shilo 2003). A major conundrum for investigations of signaling pathways is how can the same cascade elicit such an array of responses? The simple answer is that cells are a combinatorial environment in the sense that no two cell types share the same composition of expressed genes or cellular histories. The gene expression differences will be manifested in differences in proteins, lipids, and glycosylation. Similarly the history of a cell or a lineage, can lead to molecular memory mediated by methylation, phosphorylation, and acetylation, (Lewin 1997, Gerhart and Kirschner 1997). The complex answer is that for each pathway there will be a set of molecular agents that will make one outcome of a signal more probable then another. The emphasis is on probable, as the signaling events rest on stochastic motion of molecules (McAdams and Arkin 1997). In the case of the EGF/Ras pathway then the successful mediation of a signal is not an

instantaneous decision taken by a single cell. It is more like a deliberate evaluation, taken over time often by a population of cells. For instance it has been demonstrated in human cell lines that biologically relevant signaling through EGF/Ras pathway requires continuous and stable activity of the pathway components (for details see review by Carpenter 2000). Thus the cascade harbors resilience towards background noise, but is also able to respond promptly when required. Modulation of signaling can be mediated by several molecular mechanisms. These modifications can both be results of direct interactions between the molecular components or a consequence of some form of genetic or environmental dependence, for example through nutritional or physical attributes. For example the EGF/Ras signaling not only relies on core cassette components but, as studies on human cell lines demonstrate, the effects of receptor trafficking on signaling parameters. Specifically, the receptor is observed in 3 defined regions of the cellular membrane, the caveolae domains, regular membrane and in clathrin coated pits just prior to endocytosis. Remarkably the receptor is active in all of these membrane locations just as in the early endosome vesicles. A numbers of proteins and pathways, ranging from cholesterol binding proteins and phosphoinositides to Ras-like G-proteins have been noted to affect EGF signaling (Carpenter 2000). Other potential mechanisms for mediating EGF signaling include the three MAP kinases, which enable signal amplification, a set of negative and positive scaffolding and regulator proteins affecting the cascade directly, and finally loops involving transcription of pathway components or antagonists (Simon 2000, Rebay 2002).

Thus at the molecular level there are numerous ways segregating variation in the cascade components could impact phenotypes. Natural selection on phenotypes influenced by a cascade will shape the molecular variation in genes of the pathway. A review of recent studies on the molecular evolution of the *Drosophila* EGF/Ras pathway will follow after its structure has been described.

EGF/Ras pathway in *Drosophila*

Analyses of *Drosophila* development have repeatedly implicated roles for EGF based signaling through the Ras pathway. The relationship between Ras and EGFR seems considerably tighter in *Drosophila* than in humans, where the MAPK cascade can respond to several other receptors besides the EGFR, and activation of the receptors is mediated by other cytosolic cascades like (PI3K). There is only one known exception in *Drosophila*, where the Platelet derived growth factor (PDGF) and vascular endothelical growth factor (VEGF) are needed simultaneous to EGFR activation to guide border cell migration (Duchek and Roth 2001). So for all practical purposes the EGFR receptor and the Ras cascade can be considered a coupled unit. One of the first *EGFR* alleles identified was called “faint little ball” because of the severe early

developmental phenotype (Clifford and Schüpbach 1989). The locus was later identified as contributor to oogenesis, photoreceptor determination, vein development, bristle patterning and recently border cell migration (Duchek and Roth 2001). The cascade is also required for axon guidance in the developing nerve system (zur Lage *et al.* 1997) and may also play a role in adult brain (Botella *et al.* 2003).

EGFR is represented by two transcripts, differing by 5'-exons which contains the putative signal peptides and parts of the extracellular domain. The expression of the transcripts is dynamic throughout development, starting before cellular boundaries form in the egg[6]. There is a considerable amount of maternally loaded transcript, but the turnover of the protein must be rapid since the "faint little ball" null mutations have a lethal early phenotype. The dynamics of EGFR expression throughout the *Drosophila* development have been documented (Lev *et al.* 1985, Schejter *et al.* 1986 and Kammermayer and Wadsworth 1987). The two transcripts show largely overlapping domains of expression, with the only difference being that the RA transcript has extended duration in the adult (Lev *et al.* 1985). In the malaria mosquito, *Anopheles gambiae* similar patterns of expression were detected, with sustained transcription in adults (Lycett *et al.* 2001). Interestingly, there appear to be multiple splice variants being differentially expressed throughout mosquito development. It is not clear if these represent alternate 5' exons as in *Drosophila*, as the cDNA sequences only include the core part of the gene. Survey of the released *A. gambiae* genomic sequence led only to positive identification of one 5' prime exon with noticeable sequence similarity to *D. melanogaster* (Palsson, data not shown). There are two possible explanations for the observed protein size variants. There could be more alternate 5'-exons that have yet to be identified from the genomic sequence. Also either of the two novel introns in the *A. gambiae* EGFR may allow alternative splicing.

Activation of the pathway can be monitored by quantifying the phosphorylation of intracellular components, for instance Raf or MAPK. The domains of activation throughout *Drosophila* development are clearly restricted and appear tightly regulated (Gabay, *et al.* 1997). The hypothesis has been that even though EGFR has dynamic expression, most of the regulation of signaling is mediated by cytosolic factors rather than at the transcriptional level. If this was the case then one would expect analysis of the effects of segregating variation in EGFR to implicate protein variants rather than polymorphisms that alter attributes relating to mRNA production and stability. A direct contrast would of course have to weigh in the number of polymorphisms belonging to each category, which is highly asymmetrical as I show in chapter 3.

Genetic analysis of EGFR alleles in *Drosophila* determined a level of functional independence among the protein domains. The receptor has two extracellular Cystein-rich

[6] The first 12 divisions in the Drosophila egg happen only at the level of nuclei. After twelve divisions the nuclei travel to the periphery of the egg and cell membranes start forming from the surrounding membrane.

repeats, a single membrane spanning region and a receptor tyrosine kinase domain that includes the catalytic capacity on most of the phosphorylation sites. The effects of alleles can be characterized by testing for their ability to complement characterized null mutations (Raz *et al.* 1991). Clifford and Schüpbach (1989, 1994) tested 32 EGFR alleles against a major mutation of the locus and also by testing all possible combinations of alleles and assessing their capacity to complement. Most of the EGFR alleles fell into 3 well defined classes and the remaining were lumped into a "leftover" class. The first major class included alleles affecting all developmental processes uniformly, and can be considered general loss of function alleles. The second class included four alleles impacting primarily embryogenesis, and the three alleles in the third class retain oogenic capacity. The fourth class includes eight alleles that impact certain developmental processes more severely than others. Clifford and Schüpbach (1994) proceeded to determine the molecular lesions for 24 of these alleles, all of which turned out to be changes in the protein. The class 1 alleles altered the protein mainly by creating truncated products, but there were also two changes in the receptor tyrosine kinase domain. Both of these latter ones are temperature sensitive and result in loss of vein material. Of the remaining lesions, two of the class four mutations affected wing development. Interestingly those were the only two lesions in the second cystein right repeat of the ligand binding domain. Those lesions might therefore impact the efficiency of ligand binding in the wing.

These results support the notion that the EGFR serves multiple roles during development. More strikingly they highlight the complexity of functions mediated by the receptor, as mutations in different domains can have distinct consequences on particular processes. Thus predictions about the effects of allelic variants in the locus are difficult, even in the context of a particular phenotype like the wing shape. One caveat with the Clifford and Schüpbach's experiments is that major alleles are derived from a range of sources so they may differ by more than the lesion alone. The modifying capacity of different genetic backgrounds is less appreciated in the wider *Drosophila* community, but Clifford and Schüpbach (1989, 1994) managed to clean up one major modifier that had led to misclassification of an allele.

EGFR in wing development

The epidermal growth factor receptor is required for at least four distinct functions in the wing disc. Early expression of the dorsal/ventral specifying gene *apterous* depends on EGFR signaling. The cells expressing *ap* later become independent of the receptor as the cellular field expands. The Iroquois Complex (Iro-C) on the other hand shows the sustained dependence on EGFR and Ras signaling during subsequent development (Zecca and Struhl 2002). Iro-C is required in the cells of the notum. EGFR also has two distinct functions in the wing proper. The main role is in vein cell proliferation and differentiation as the receptor is activated in the vein

primordia. Loss of *EGFR* function leads to loss of vein material, with the distal ends of veins L5 and L4 and L3 being first affected. A more comprehensive knockdown of the pathway by removing one copy of both *EGFR* and *rhomboid* leads to complete failure of vein formation. Similarly, *Ellipse* gain of function alleles induce extra veins in the intervein regions (Lesokhin *et al.* 1999). The earliest known molecular indicator of vein formation is *rhomboid* (*veinlet*) mRNA in stripes along the distal-proximal axis. Rhomboid protein functions in the Golgi system to process the spitz precursor, turning it into an active ligand for EGFR as previously described. Genetic analyses indicate that the precise expression of veinlet may be initiated by a low level of EGFR signaling triggered by the vein ligand or possibly keren (Simcox 1997, Shilo 2003). Unprocessed spitz has no biological activity in the wing. Interestingly the newly discovered ligand keren which has closest structural resemble to spitz does however seem to undergo low level autocleavage. Therefore, both keren and vein (which does not require cleavage to become active but is inherently weaker activator) may act as the early inducers of the pathway. The activation of the receptor will be taken over by the more potent ligand spitz once rhomboid production kicks in[7]. Consistent with this model is the observation that rhomboid expression can be induced by EGFR signaling in some tissues, including the wing (Shilo 2003). Thus it is safe to conclude that part of the control on EGFR activation in the wing seems to be at the level of ligand processing.

Experiments also suggest that *EGFR* and *veinlet* enter an early autocatalytic loop that helps establish the vein primordia (Martin-Blanco *et al.* 1999). Thus EGFR/Ras signaling plays an integral part in establishing veins and ensuring integrity of the differentiated cells. The latter function is shared with *Notch* signaling components. Finally, after the vein primordia have been specified then EGFR becomes active in the intervein cells. This polarization of the receptor signaling pattern is a puzzle but the gene product is may be required for proliferation in the intervein cells. The expression pattern of the EGFR transcript demonstrates this shift. During larval development the expression in uniform across the wing field, but 8 hours post puparium formation a sharp reduction in vein tissues is seen (Sturtevant *et al.* 1994). This drop in EGFR signaling in vein primordia is required for *dpp* expression and progression of vein development. Martin-Blanco *et al.* (1999) further argue that low level activation of EGFR in intervein cells during the pupal period is a consequence of a balance between the weak activator vein and the antagonist argos (Schweitzer *et al.* 1995). Exact biological purpose for this mild but sustained activation has not been determined but it may provide a general survival or growth signal to the intervein cells, as is the case with *dpp* (Moreno *et al.* 2002). This hypothesis is supported by results from induction of Ras clones in the wing, where the cells lacking Ras1 stopped growing

[7] The fourth activating ligand *gurken* is tightly controlled and seems to have restricted biological role, functioning nearly exclusively in oogenesis.

and as a population underwent more apoptotic events (Prober and Edgar 2000). Monitoring of cell cycle progression suggest that EGF/Ras signaling affects the S/G1 transition through Cyclin E and thus promoting cell divisions as well.

There are two kinds of veins on the *Drosophila* wing, longitudal and cross-veins. The longitudal veins develop first and the crossveins are established in three later steps (Marcus 2001). First the cross-vein class of genes defines an area of vein potential by a very broad stripe of expression. This domain is refined by the TGF-β related signaling genes *dpp* and *gbb*. Third, the full realization of cross veins is implemented by EGFR signaling in similar way as before. It has yet to be determined if EGFR is just needed for establishment of the cross-vein primordia and then turned off or if signaling through the pathway prevails throughout vein differentiation. The results of Martin-Blanco *et al.* (1999) on longitudal veins discussed above suggest that EGFR signaling should go through a biphasic transition as the *dpp* pathway takes over in the vein primordia to facilitate vein material formation.

Together the four distinct roles of EGFR in wing development and the spatial and temporal dynamics of its expression in vein and intervein tissue constitute a complex picture of the role of the locus in patterning the wing. Because of these multiple developmental roles of EGFR a clear *a priori* expectation of the effects of EGFR on wing shape is difficult to formulate. Part of the problem is our limited understanding of how shape of developing morphologies is achieved. That question will be revisited in a following section.

C. Evolution of insect wings

The material for morphological radiations must have been available in ancestral species as segregating variation in genes contributing to development. Insect wings have undergone major transformations from the days of the dragonflies, while the structure has been fairly stable in the Sophophora lineage for about 50 million years. Within the group wing traits have proven of little value for phylogenetic reconstruction (Powell and De Salle 1995). The ancestral stages are exemplified by rich, almost grid-like, vein patterns. While reconstruction of the evolutionary history of ancient adaptation (Marden and Kramer 1994, Hasenfuss 2002) is clearly complicated some general features are acknowledged. Insect wings must have originated outgrowth, either from legs or primitive gills. The formation of a usable flying apparatus must have been accompanied by structural modifications creating a flat appendage capable of generating lift if flapped. Once the creature took to the skies then selection may have shaped the wing blade, reduced and redistributed mass, supporting structures (veins) and flexibility all important for flight (Dudley 2000). Refinement of the intervein regions must have been an important feature of early vein development. Most of the later evolutionary changes have been in terms of placement, physical properties and number of veins. This sequence of events creates a parsimonious model of the evolution of regulation of wing development that is consistent with our knowledge of the molecular mechanics as detailed in the previous section. At the genetic level most of the regulatory apparatus is concerned with refinement of the veins, while a single locus, *SRF-blistered* is the primary regulator of intervein fate. According to this model intervein formation in a primitive insect was first established by the *blistered* locus. Later, as selection began modifying the specifics of vein development, several other loci may have been co-opted, resulting in the multilayered hierarchy of genes known in fruit fly wing development.

This begs the question of how conserved is the genetic network guiding vein formation? This question was addressed indirectly by (Abouheif and Wray 2002), who studied polyphenism in wings among several ant species. They started with the hypothesis that the major loci known to control *Drosophila* wing development would also play a role in ants. They surveyed expression of several key regulatory loci in the hierarchy in winged or wingless casts of several species. The earliest operating locus surveyed *Ultrabithorax* operates as a repressor of wings in *Drosophila*, and *extradenticle*, which has a function in the non-wing part of the wing blade. The four remaining genes they surveyed were *engrailed, wingless, scalloped* and *spalt*, all of which operate in compartmentalization of the wing, with only *spalt,* which responds to *dpp* signaling, having a role in vein formation. Their results support the hypothesis that the most upstream and core parts of the wing developmental hierarchy is conserved. Regarding polyphenism, they noted that changes in several junctions of the hierarchy were correlated with wing being lost or partially developed. For instance only the expression of *spalt* was disrupted in wing primordia of

the soldiers of *Pheidole morrisi*, while none of the above mentioned genes were expressed in a second cast of the species. The developmental morphologies of the wing primordia in the two casts did not hint at this dramatic difference. Analysis of more species confirms that evolved suppression of the wing program can affect multiple components in the hierarchy. Further elucidation of the exact molecules responsible for these effects will be very interesting. These results have a direct bearing on the discussion about stabilizing selection and cryptic variation. This implies that while structures do remain stable the underlying genetic system is retained relatively intact. The alternate hypothesis is that the regulatory network constituting stable traits is free to drift, and can thus over time undergo both quantitative and topological changes (von Dassow 2000). In the case of wing development in insects the hierarchy seems to be well maintained even between the two high order evolutionary groups Diptera and Hymenoptera.

The conservation of loci involved in vein-intervein formation was not addressed by Abouheif and Wray (2002). Insight into the potential conservation of these loci can be gained from two studies, one using purely developmental genetic methods and the second rested firmly in the domain of quantitative genetics. Recall that the transition from a uniform cellular field to defined vein-precursors takes place in the larval stage. Biehs *et al.* (1998) described how the initially broad domains of expression of the vein determining loci get refined into distinct stripes, most probably by a combination of effects from EGFR and the Notch lateral inhibition pathway. They also noted that early pro-vein patterns emerge in the intervein regions but are suppressed as the "true" veins get substantiated. Bier (2000)[8] postulated that these are molecular rudiments of an ancestral pattern characterized by a richer venation pattern. He supported this hypothesis with the observation that in mutant backgrounds extra vein material appears non-randomly in the intervein regions. Similar suggestions were made by Thompson (1974a) who studied the buffering capacity of wild type backgrounds against the major mutations including *veinlet, blistered* and *net*. He documented more carefully the placement of extra vein material in the *blistered* backgrounds and proposed a similar pattern of suppression of the provein potential. More recently Fletcher and Thompson (2001) also document this pattern in selection lines sensitized by a hairy mutation. Both Thompson and Bier postulated that these provein potentials may reflect an ancestral prepatterning mechanism that is suppressed in modern day *Drosophila* and related species.

The evolutionary stability of the venation arrangement in the *Drosophila* clade argues for a predominant role of stabilizing selection in molding wing parameters. This may be the reason for the evolutionary stability of the hierarchy of wing regulatory genes. There also seems to be a puzzling level of conservation of the capacity to create veins in regions of the wing that

[8] Bier 2000 was published within few months of Chapter 2, the GDE paper. Bier used the review to elaborate on the findings of his lab (Biehs *et al.* 1998) and to put it into the context of the ancestral venation pattern.

have not had veins for 100 million years (Powell and De Salle 1995, Stark *et al.* 1999). One explanation is that the vein formation capacity reflects an autonomous potential of the wing cells to create vein material that is uncovered when the most potent suppressor of vein formation *blistered* is removed. The active role of EGFR and Dpp signaling in promoting vein formation argues for a more elaborate picture, presumably one where the competing regulatory genes orchestrate the vein formation potential of the wing cells.

Standard toolkit for analysis of shape

The pioneers of *Drosophila* wing shape analysis attempted to summarize shape in an intuitive manner (Thompson 1974b, Curtsinger 1986, Cowley and Atchley 1990). But not all intuition are the same, which stemmed from the lack of coherence about the fundamental issue of shape description. In addition to distances between two points or landmarks, researchers have used ratios, angles and even counts to capture shape (Bookstein 1996a) Comparison between studies is complicated by the fact that very few parameters are measured in common. The richness of parameters include for example, length of veins L2 and L3 (Thompson 1974b), length of L3 from crossvein to tip (Curtsinger 1986), angular offset of paired distances (Weber 1992, Weber *et al.* 1999, 2001), wing area (James *et al.* 1997), and wing length (Imasheva *et al.* 1994, 1995). Tools for capturing and describing shape variation in adequate mathematical terms have only been developed in the past 30 years. Historically this breakthrough traces back to D'Arcy Thompson (1961) who provided a geometric framework to summarize shape, and demonstrated how evolutionary divergent forms might arise by simple transformation of geometric shapes. This obviously held great promise for the study of evolution, both at the level of phylogenetic relationships but also for documentation and analysis of morphological variation within species. Full realization of this promise required the integration of two disciplines of mathematics, multivariate statistics and geometric theory. Multivariate statistics was integrated first, as a way to capture axes of variation in numerous descriptors of shape. The advantage of multivariate analysis is the capacity for collapsing information and quantitative nature of the metrics. The disadvantage is the lack of a clear way of tracking back from a multivariate descriptor to a corresponding change in the "real" data of interest. That problem was solved when morphometricians adopted the geometry of Kendall shape space and Thin Plate Splines. Kendall's shape space is the only natural solution for the mathematical comparisons of shapes (Bookstein 1996, Rohlf 1996). The second novelty is the Thin Plate Splines which are a method to account for local as well as global changes in shape. The new paradigm of morphometrics, called the generalized procrustes analysis proceeds in four steps.

1. The centroid size of every specimen is calculated, (centroid size is the squared distance from all landmarks to the geographic midpoint of each specimen). Each specimen is scaled to the mean centroid size to remove the effects of size.

2. Specimens are rotated with procrustes superimposition to achieve optimum correspondence between homologous landmarks. The process involves pairwise comparisons between specimens that are iterated until the data converge.

3. The aligned specimens are then subjected to orthogonal projection in tangent space. Several methods are available but the best refined package is Rohlf's (2002) Thin Plate Spline (TPS) Relative warp analysis software. These procedures capture the global and local differences in landmark configurations with standard multivariate statistics.

4, Finally the results are represented graphically, most commonly by transformation grids in which two extreme shapes are projected onto each other and the changes are captured as alterations in the grid.

A key advantage of modern morphometrics tools is that variation in shape is extracted from the scored variables in an unbiased manner. The experimenter naturally decides on a structure to study and the exact landmarks to score, but then the algorithms of morphometrics are applied to extract the axes of variation. Preconceptions about the meaning or importance of shape do not factor into the process (Bookstein 1991). The only exception is if a prior knowledge argues for division the landmarks into categories.

This new protocol of morphometrics is a decade old and has been utilized to address an array of biological problems from phylogeny (Fink and Zelditch 1995) to fluctuating asymmetry (Klingenberg and McIntyre 1998) with well over 100 publications a year using the methods (Adams *et al.* 2003). There are also additional developments, for instance the extension to three dimensional data, semi-landmarks to capture shape of surfaces and ways to account for missing landmarks. Still the adoption of the protocols has been slower than anticipated in the *Drosophila* community where recent reports are still utilizing simpler ad hoc metrics. I believe that general application of these methods will be beneficial. A standard for which landmarks to scored may be an excessive proposition but a general language for landmark identity and potentially a unified repertoire for landmark data could facilitate cross-talk and allow direct comparisons between experiments. That could lead to a synthesis on the development, genetics and evolution of the *Drosophila* wing.

Microevolution and function of wings

The *Drosophila* wing and the mouse mandible are prime examples of structures that have been used in investigations of the developmental, quantitative and molecular biology of form (Atchley and Hall 1991, Leamy *et al.* 1998, Klingenberg 2002). The specific questions being addressed about *Drosophila* wings include; developmental constraints or integration (Cowley and Atchley 1990, Klingenberg and Zaklan 2000), the shape of reaction norms (Moreteau *et al.* 1995), phylogenetic G-matricies (Galpern 2000), stability of development (Klingenberg and McIntyre 1998), phenotypic plasticity (Bitner-Mathe and Klaczko 1999). However the main focus has been on the use of wing length as a proximate parameter for body size (Imasheva *et al.* 1995, de Moed *et al.* 1997, Cortese *et al.* 2002), particularly in the relation to clinal variation (James *et al.* 1997). The relationship between clinal variation in wing length and body size is only understood superficially. Imasheva *et al.* (1994) report that wing length is positively correlated with temperature and that the effects can mainly attributed to posterior part of the wing. Rand also noted similar response in the posterior region in when populations of flies assimilated to extreme temperatures in the laboratory (personal communication). Partridge and coworkers however report the reciprocal pattern between body size and wing length in Australian, South American and African clines, where larger flies at higher latitudes have been observed (see James *et al.* 1997 for review). Moreover, Long and Singh (1995) found a non-monotonic cline in the US, with smaller flies at the extremes. In this case then the wing cline was uncoupled from the size cline casting doubt on the generality of the pattern. The disparity of those observations could be explained by different metrics and as such highlights the need for application of the morphometrics toolkit. One must however remain skeptical of the relationship between size and wing length or shape unless consistent evidence acquired with the same methods is provided. Currently such data are not available but extensive work by Partridge and coworkers suffices for now (James *et al.* 1997). While several theories can account for clinal variation, then identical clines replicated on several continents argue very convincingly for the role of selection (Gilchrist and Partridge 1999). Recent analysis of freshly collected strains from the Australian locations surveyed molecular markers along with the phenotypic attributes. The phenotypic cline was reproduced but the molecular markers did not exhibit a genome wide correlation with the cline (Gockel *et al.* 2001). However five of the nineteen markers showed significant clinal differences. Gockel *et al.* (2001) confidently argue that these results suggest the role of selection in maintaining the wing area cline since the Australian population does not show evidence of structure that might alternately explain the phenomena. I believe this conclusion is too optimistic as our understanding of the genomic signatures of population subdivision and stratification are still rudimentary. Yes we can reject the hypothesis that major stratification is causing the pattern, and the most convincing alternative is natural selection.

The most interesting cline in wing shape was observed in *Drosophila subobscura.* The species was exclusively found in the old world and there was a clear clinal variation in the length of the wing. The discovery of transplanted individuals of the species in South America in 1978 put evolutionary biologist on guard and they have carefully monitored the population as it spread north.. Initial studies implicated no phenotypic cline in the new North American population but it was firmly established in a sample collected in 1998 (Huey *et al.* 2000). The rapidity of the phenotypic response suggest very high selection differential, even for these subtle differences. The most striking observation was that even though evolution proceeded to achieve the analogous shapes on both continents, it did so by different trajectories. The wings of the old world stocks were extended by displacements of landmarks in the basal portion of the wing while the North American wings were elongated by movement of distal structures. Calboli *et al.* 2002 proceeded to test for and found a comparable cline in South America. And again the developmental basis, cell size and number, were contingent while the phenotypic evolution was predictable. Those observations along with the conservation of venation pattern over millions of years suggest the role of stabilizing selection maintaining the integrity of the structure. While the wing retains living sensory neurons its main functions are considered to be flight and courtship.

As highlighted above veins provide support for the wing and are critical for its performance as a flying apparatus. For instance the thickness of veins diminishes from the proximal to the distal end, ensuring flexibility of the distal part of the structure (Dudley 2000). This is clearly demonstrated by loss of signaling by the EGFR and dpp pathways. They always lead to reduction in vein material at the distal tip of the wing, and progressive loss of signaling leads to further shortening of the veins. Interestingly vein rudiments linger in the proximal regions in most mutant combinations (de Celis 1997). Ennos (1988) preformed biomechanical analysis of wings and postulated the number, strength, and location of crossveins being primary determinants of wing rigidity. Of these two properties only the placement of crossveins is amenable with the tools applied in the current thesis to investigate shape. Variation in vein thickness could potentially be extracted from the images but would require a leap in automatic data extraction from images. In conclusion the literature of biomechanics of flight highlights the importance of veins in determining the function of wings (Dudley 2000) in conjunction with the regulated application of the appendage (Dickinson *et al.* 1999).

Males vibrate their wings in order to produce a species-characteristic song during courtship of the female. This maneuver is one step in the multi-step mating process that is also affected by the genes of the circadian clock (Kyriacou and Hall 1980). Four main components of the male courtship song have been characterized: sine song frequency, interpulse interval, intrapulse frequency (= carrier frequency), and wing cycles per pulse (Barnes *et al.* 1998). The shape of the wing may well affect those attributes but a systematic analysis has not been attempted, mainly because quantification of these components needs to be conducted on living

individuals during courtship. Recent advances in video-imagery and processing that are being exploited to study flight (Fry *et al.* 2003) in untethered specimens could be applied without major adjustments.

Developmental integration and modularity

The two main developmental questions regarding shape are, how is the final shape realized and how independent are the constituents of shape? The findings of developmental genetics demonstrate the specific cellular and molecular events leading to the patterning of the *Drosophila* wing. Some processes when investigated with these tools have direct consequences for shape while other seem to less important. It is possible that the loci implicated by "sledgehammer" genetics will have insignificant effect for wild type individuals. A parallel analysis by classical and quantitative genetics may be the best approach to the consistently elusive phenomena of shape.

The first question is how is shape realized? One hypothesis states that it is a simple function of cell number, size and shape. Alternatively shape is achieved by a higher order regulative patterning mechanism? Members of our laboratory, Birdsall *et al.* (2000) addressed this question by studying the quantitative genetics of wing shape and cell number in a panel of inbred lines in response to environmental stress. They noted that size was primarily affected by sex, with a small genetic component. The shape measures were more strongly affected by the genetic component and were stable to large deviations in wing size due to sex or temperature. Cell number in two intervein regions (C, and D) had a smaller genetic component. But more intriguingly they showed differential dependence on sex and temperature. The cell numbers along the middle region (C) of the wing were affected by temperature but cell number in the D region depended more on sex. A similar experiment was conducted by De Moed *et al.* (1997) on fewer lines but additional environments (temperature and food). They concluded that the cell numbers and sizes are not the most important determinant of wing length, but that a combination of their effects is. Paraphrased, the overall wing shape has stronger genetic component than the properties of the cellular populations comprising the wing. Several major mutations affecting cell size have been identified, but the interesting conclusion is that these have only minor effect on the size of the organism (Su and O'Farrell 1998). Similar results have been seen for mitotic clones which seem to accommodate differential growth to fill the shape (de Celis 1998). Even prevention of cell cycle progression does not alter shape drastically (Weigmann *et al.* 1997), nor does increasing the number of cell cycles (Neufeld *et al.* 1998).

The question about relative role of integration or modularity in the wing touches on the mechanistics of development and the phenotypic and genotypic variation in the appendage. Disentangling the two may prove elusive, as the genotype unfolds by the process of

development to yield the adult structure. The question will concentrate on variation at each level and the relations between the axes of variation, i.e. variance-covariance matrices. Constraint is a term to summarize lack of independence in phenotypes, due to evolutionary history, developmental mechanisms or low dimensionality in the pool of allelic variation. Constraints are thought to be rooted in the mechanism of development (Gould 1977). One can also argue that constraints may originate because of the pleiotropic relationships between segregating alleles. Lately these patterns of dependence or independence are discussed in terms of modularity. This term carries a more mechanical meaning as it can refer to a minimal enhancer element, the proteins of a signal cascade, collections of cells in development or just body parts of an adult.

The molecular details of gene action argue for a degree of independence between genetic determinants in the anterior and posterior part of the wing, and also for unique combinatorial signal establishing individual veins. Those facts propose directly the hypothesis that development is compartmentalized and may proceed independently in parts of the cellular field. However the fact that the same loci are used repeatedly over the wing, and that some genes give phenotypes across the whole structure, proposes a clear alternative. The most logical solution is that both theories will be correct to a yet determined extent. Quantitative genetics provides another framework to address the question of modularity and integration. Direct artificial selection can be used to test for constraints in morphological structures. Beldade *et al.* (2002a) selected for increased and decreased size of two eyespots in the forewing of *Bicyclus anynana*, in a coupled and uncoupled selection scheme. The coupled selection, that is for increase in both eye spots simultaneously, yielded a highly significant response. Interestingly, application of opposite selection forces on each eye spot also resulted in changes in size, though the magnitude was smaller. When the results were summarized by each eye spot and selection pressure then the coupling vs. uncoupling schemes did not differ in their capacity to mold the variation for spot size. This suggests that at the level of segregating genetic variation the two spots are unconstrained. While butterflies and dipterans are evolutionary divergent these results suggest a reasonable level of coupling of the available variation in the structure. Thompson (1974a, 1974b) investigated the capacity of natural alleles in *D. melanogaster* to suppress major venation mutations both by selection schemes and by direct crosses. He found that a fraction of the modifiers had general effects that extended over the wing blade while another subset affected certain regions specifically. More recently Fletcher and Thompson (2000) sensitized the bristle formation in the wing by crossing a mutation in *hairy* nd selected on the exposed variation. They found some regional specificity in the capacity to suppress and enhance the bristle phenotype. The perturbance also generated extra vein material in intervein regions B, D and E, that showed the same results. They further noted an interesting coupling of these developmental defects, as excess of sensory bristles was

accompanied by increased extra vein material. It is interesting in this context that a polymorphism in *hairy* is significantly associated with bristle number (Robin *et al.* 2002).

Klingenberg and McIntyre (1998) provided a more direct analysis of the patterns of variation in the wing. They pioneered the use of morphometrics to study the stability of development by the proxy of fluctuating or directional asymmetry. Sticking only with the alignment features of the toolkit they did not proceed to extract principle or relative warps as their interest is in developmental noise quantifiable in pairs of structures. Their main result was that the covariances in landmarks observed between individuals where indistinguishable from the covariances for fluctuating asymmetry. This suggests that the same developmental processes or allelic differences contribute to both phenomena. Theoretical work supports this lack of need for specific "stability" loci (Klingenberg and Nijhout 1999). Klingenberg and Zaklan (2000) applied the same morphometrics framework to ask directly about developmental integration in the wing. They conducted two kinds of analyses: regular TPS based analysis to survey the variation in the wing; and a partial least square method that computes covariances across the anterior/posterior boundary, to test modularity directly. Pervasive integration was observed as both techniques gave very similar results, proposing that the main trajectory of variation in wing shape is involved with the overall shape of the structure. Thus in the case of the *Drosophila* wing then integrated patterns of variation predominate but still leave room for localized region specific effects.

A promising new technique has been proposed by Mezey *et al.* (2000) who described a QTL study on mouse mandibles. Their method tests directly for non independence of the correspondence between QTL's and designated shape variables. The projection ramus was found to be distinctly modularized in agreement with the quantitative developmental genetic models of Atchley and Hall (1991). This method has yet to be applied to *Drosophila* data, but a QTL study by Zimmerman *et al.* (2000) can be considered a partial step. We chose to investigate wing shape by breaking it up into intervein regions and testing for segregating variation affecting those independently. Some of the individual parameters derived from each region were correlated but the QTL profiles for each trait were distinct. This argues that regardless of the integration we can identify at the molecular level distinct contributors to each region. It is possible that the integration is a consequence of processes like developmental logic, pleiotropy or epistasis.

Quantitative genetics of *Drosophila* wings

Numerous quantitative genetic analyses of wing parameters have documented a naturally occurring genetic component. The early analysis of Waddington (1957) suggested that a great deal of natural genetic variation, both standing and cryptic, was for aspects of wing

development. Further elaboration of the evolutionary genetic basis of shape was conducted by Thompson (1974a, 1974b, 1975). He studied the capacity of natural alleles to suppress or enhance known vein mutations. Most of the clearly polygenic effects behaved in a cumulative manner arguing for a substantial additive component. However the rapidity of response in his selection experiment and a plateau in repeated experiments suggest that some alleles of large effects were segregating. Those early selection experiments and the elegant selection work of Weber (1990a, 1990b) further supports polygenic basis of the traits. Gilchrist and Partridge (1999) conducted crosses amongst extreme lines from three independent wing shape clines and conducted generation mean analysis to determine the relative role of additive, dominance and maternal factors. There are some reservations about the utility of the generation mean analysis as a technique (Carrillo and Gibson 2002), mainly because of how the model is constructed and its overall significance estimated. Taken at face value the results of Gilchrist and Partridge (1999) are consistent with earlier notions of complex inheritance, with suggestive contributions of epistasis, maternal and Y chromosome effects. The preponderance of evidence suggests strong polygenic inheritance for wing shape that could be dissected with appropriate techniques.

Advances in QTL mapping led to three recent papers studying the quantitative genetics of wing shape. Weber *et al.* (1999, 2001) created two panels of recombinants from the second and the third chromosome derived from his selection lines (Weber 1992). Those studies had increased resolution compared to earlier studies to estimate the number of wing shape QTL's differing between the parental lines, and were also used to assess quantitative genetic properties of the QTL's such as testing for epistasis. The results implicated numerous loci on both chromosomes as contributing to shape variation. Those QTL acted predominantly in an additive manner while epistasis was detected for a subset. Analysis of shape and size of *Nasonia* wings yielded several major QTL, and prevailing epistasis (Gadau *et al.* 2002). Their design took advantage of the haploid state of *Nasonia* males, which boosts the statistical and genetic power to detect epistasis over studies in diploid organisms (Lynch and Walsh 1998). While additive variation seems to be prevailing then those studies suggest significant contribution of more complex inheritance to variation in wing morphology. In our QTL analysis we utilized the morphometrics toolkit as discussed in the previous section (Zimmerman *et al.* 2000). Two designs and pairs of study populations were used. One panel was a set of recombinant inbred lines established by Nuzhdin, Pasyukova and MacKay (1997). The other was a backcross design with inbred lines chosen because of their extreme wing shapes. The parents of the first panel did not differ drastically in wing shape. Size and shape were not correlated and were affected by separate set of QTL's. 35 QTL's affecting shape were identified, mapping to 23 locations. Most off the effects were additive and of moderate effect. This is entirely at odds with the eight size QTL's which all showed dominance. The results from

Weber *et al.* (1999) and Zimmerman *et al.* (2000) are in good agreement one the polygenic nature of wing shape even if the parameters were quite distinct. QTL studies generally have low resolution as each peak covers somewhere on the order of 100-1000 genes. Randomization procedures can be applied to test if a specific subset of loci is over or under represented under the QTL peaks in a given study. By this logic Zimmerman et al (2000) implicated vein-determining loci as they were over represented under the QTL peaks. One QTL mapping to the same region as *EGFR* was found to contribute to variation in the anterior part of the wing (IVR-B). These results provide a piece of evidence for the role of allelic variation at *EGFR* in wing shape variation in the wild. Further independent experiments are needed to validate this result.

Synopsis

Here the concentration is on identifying naturally occurring variants contributing to phenotypic differences. I chose to investigate the shape of the *Drosophila* wing and explore its utility for mapping of quantitative trait loci to individual nucleotides. Wing shape unfolds by the action of many characterized genes in a complex developmental fashion, involving patterning and major rearrangements of the appendage during pupation. Previous work has shown that shape has polygenic basis and can be disentangled by quantitative genetic analysis, In this thesis I have built on these result and tested for segregating variation at several wing development loci, providing an alternate test of the effects of genes implicated by QTL analysis. Chapter 2 has already been published and here I only modified the formatting for coherence. The data set is revisited in Chapter 4 to address explicitly the effect of *EGFR* on wing shape. In the third and fourth chapters I proceed to investigate respectively the patterns of nucleotide variation in the vein-determining locus *EGFR* and its relation to shape parameters.

Chapter 2

Quantitative developmental genetic analysis reveals that the ancestral dipteran wing vein prepattern is conserved in *Drosophila melanogaster*

Abstract

Quantitative complementation tests provide a quick test of the hypothesis that a particular gene contributes to segregating phenotypic variation. A set of wild-type alleles is assayed for variation in their ability to complement the degree of dominance of the quantitative effect of a loss of function allele. Analysis of 15 loci known to be involved in wing patterning in *Drosophila melanogaster* suggests that the genes *decapentaplegic*, *thickveins*, *EGFR*, *argos* and *hedgehog*, each of which are involved in secreted growth factor signaling, may contribute to wing shape variation. The phenotype of one deficiency, *Df(2R)Px2*, which removes *blistered/Plexate*, is also highly sensitive to the wild type genetic background and at intermediate expressivity reveals six ectopic veins. These form in the same locations as a projection of the ancestral pattern of dipteran wing veins onto the *D. melanogaster* wing. This atavistic phenotype indicates that the wing vein prepatterning mechanism can be conserved in highly derived species, and implies that homoplasic venation patterns may be produced by derepression of vein primordia.

Authors: Palsson A. and Gibson G.

Published: Development, Genes and Evolution. 2000 December 210 (12): pages 617-622

Introduction

One of the major issues in the study of evolution and development is to determine the relationship between changes in regulatory gene expression that distinguish higher taxonomic levels, and variation at the species level. The basic conundrum is that if a genetic change that distinguishes say a butterfly from a dipteran wing is introduced into one of the species being compared, it is generally predicted to result in a maladaptive phenotype. Despite the presence of genetic variation that could potentially soften the deleterious effects of and hence increase the probability of invasion of a macromutation (Mackay and Fry 1996; Gibson *et al.* 1999), population geneticists generally downplay the contribution of saltationary genetic changes, particularly in animal evolution. Two models that might account for marked changes in the expression of regulatory genes are (i) that the differences observed between orders result from the gradual accumulation of subtle differences at the species level, and (ii) that significant evolutionary transitions involve genes considerably downstream in a genetic hierarchy, and that changes in regulatory genes occur at a later time, without dramatically affecting the phenotype. It is thus important to ask the question whether variation in regulatory genes affects morphology within modern day species.

While interval mapping has become the standard method for identification of regions of the genome that affect quantitative traits, its resolution is too low to locate candidate genes with confidence, and a new approach known as quantitative complementation testing (Long *et al.* 1996; Mackay and Fry 1996; Gurganus *et al.* 1999) has been proposed as a quick test for the possible involvement of known genes. Whereas a significant difference in mean phenotype of heterozygous (+/-) and wild type (+/+) individuals across a range of genetic backgrounds indicates that a mutation affects a trait, the demonstration that a set of wild type alleles differ functionally requires a test of the interaction between genotypic classes. That is to say, if there is significant variation in the difference between ($+_i/+_t$) and ($+_i/-$) for a set of $+_i$ alleles measured in siblings carrying a common tester allele ($+_t$) or mutation (a *Deficiency* or strong loss of function allele), then the $+_i$ alleles may vary in their degree of dominance, which is a quantitative complementation test.

In practice, a set of isogenic lines carrying different wild type alleles of a candidate gene are crossed to a common inbred line carrying the mutation over a marked tester ($-/+_t$) in replicate, the trait is measured in multiple individuals of each genotype, and analysis of variance performed to test significance of the genotype by line interaction term. Graphically, interaction effects are illustrated by crossing of line means in a plot of the mean phenotype of each line in the two backgrounds. If lines do not cross, the mutation has the same effect in each genetic background and hence the wild-type alleles do not vary in their mean effect. Here we use this

technique to provide preliminary evidence that genes encoding morphogens with known roles in patterning and differentiation of placement of wing veins also contribute to subtle variation for components of wing shape.

Materials and Methods

Fly crosses

All crosses were performed at 25°C with flies grown on standard cornmeal supplemented with live yeast paste. Stock numbers from the Bloomington stock center are indicated in Table 2.1. Six wild type lines were chosen to cover a broad range of wing phenotypes, and included Oregon R, Russian 2b, two Ann Arbor inbred isofemale lines (AA3 and AA18), and two inbred isofemale lines from Kenya and South Africa (W6 and W29: Zimmerman *et al.* 2000). For all comparisons, the mutation-bearing stock was first crossed either to an inbred CyO/*PmSp*, CyO/*Pm* or *In(2LR)EN*/*Gla* stock, or to a TM6,*Ubx*/*Sb* stock, and then individual males of the genotype -/*Pm*, -/*Gla*, or -/*Ubx* were mated with individual virgin females of each wild type line to obtain F2 siblings carrying either the mutant chromosome or the tester chromosome. Flies were reared at a density of 50 to 100 larvae per 10 ml vial. Each cross was performed in duplicate, and five flies of each sex and genotype per cross were chosen at random for dissection and measurement of both wings.

Wing Measurements

Hand-dissected wings were simply mounted between a glass microscope slide and cover slip, and TIFF images were immediately captured using a SPOT camera attached to a Nikon Eclipse microscope at low magnification. The images were then analyzed with M. Rasband's NIH/Scion Image software downloaded from http://www.scioncorp.com, on a Dell Dimension PC, by capturing the XY coordinates of landmarks at the junction of wing veins and/or the wing margin (Fig. 2.1A). A common file containing the coordinates of all 480 wings (2 replicates x 2 sexes x 2 genotypes x 6 lines x 5 flies x 2 sides) were analyzed using F.J. Rohlf's program TpsRelw Version 1.17 (downloaded from http://life.bio.sunysb.edu/morph) that performs a Procrustes transformation and computes relative warps. Each mutation was analyzed separately, and consequently the warps obtained are independent of those for every other mutation, and in general capture different aspects of shape variation.

Analysis of Variance

The distribution of each of the six relative warp scores for each wing according to Genotype (*G*), Line (*L*) and Sex (*S*) was studied by three way ANOVA with the following model:

$$Y = G+L+S+G\times L+G\times S+S\times L+G\times S\times L+R(G\times S\times L)+E$$

where all effects were considered fixed, and the error term includes within and between individual variance (which were generally of similar magnitude). *S* tended to be significant, but interaction terms involving *S* were not. The terms of interest for this study were thus the overall effect of *G*, and the $G\times L$ interaction term. Type III sums of squares were computed using Proc GLM in SAS.

Results and Discussion

Quantitative effects on wing shape

Most genes affecting wing development have been characterized on the basis of the homozygous recessive effects of mutations on venation or overall wing shape. To test whether 14 such genes also have a quantitative dominant effect on shape in particular regions of the wing, the Procrustes-transformed landmark coordinates that define intervein regions B, C and D (see Fig. 2.1A) were subject to relative warp analysis followed by ANOVA. Relative warps are a highly sensitive morphometric measure (Boosktein 1996) that parse local aspects of intervein region (IVR) shape, such as breadth near the margin or relative length of the crossvein. The measures are not significantly affected by size differences, and hence are almost invariant to the effects of sex and temperature on wing size (Birdsall *et al.* 2000). For wing shape, the first two relative warps for each IVR (W1 and W2 in Table 2.1) captured over 85% of the phenotypic variance. With the exception of one warp for each IVR of *rhomboid*, *messy* and *vein*, significant differences between mutant hemi- or heterozygotes ($+_i/-$) and wild-type ($+_i/+_t$) siblings were observed (data not shown). Thus each of the mutation-bearing chromosomes show a quantitative difference associated with the number of wild-type copies of the gene of interest. This result confirms the inference from QTL mapping studies that mutations in a large number of genes can potentially affect subtle aspects of wing shape (Weber *et al.* 199; Zimmerman *et al.* 2000).

Support for the hypothesis that segregating variation at a particular locus affects a trait requires a much more stringent test, such as the quantitative complementation test. Table 2.1 indicates the significance of *p*-values associated with the genotype by line interaction term in the ANOVA for each mutation tested against up to 6 different wild type lines. Since six different

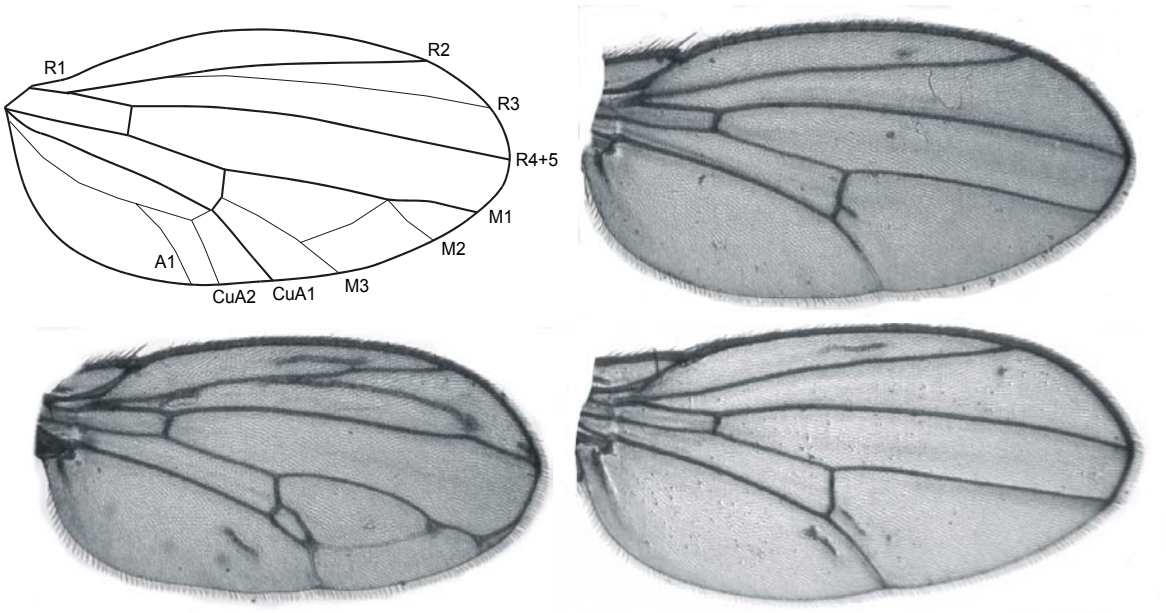


Figure 2.1. Ectopic vein formation in *Df(2R)Px2*/AA18 males. A. Outline of a typical *D. melanogaster w*ing, showing Comstock and Needham (1898) terminology (L1, R2+3, R4+5, M1, CuA1) and common developmental genetic usage in brackets (L1, L2, L3, L4, L5). The three intervein regions scored in this study are shown: landmark coordinates were captured at the junctions of veins, crossveins, and the wing margin (4 points for IVR-B and IVR-D; 5 points for IVR-C: see Birdsall *et al.* 2000). B. Projection of the location of ectopic wing veins in *Df(2R)Px2*/AA18 males onto the standard wing shape, based on extrapolation from 20 wings similar to those shown in (D, E, and F). Other genetic backgrounds show a range of variation from complete repression of ectopic vein formation, to severe blistering, but veins that do form are consistent with this pattern. C. Projection of the ancestral wing venation pattern onto the *D. melanogaster* wing, after Stark et al's interpretation of the *Protoplasa fitchii* pattern. The similarity with B is remarkable, differing only in the absence of R5, a connection between A1 and CuA2, and possibly the posterior crossvein between CuA1 and M3 (though a vestige of this may be seen in wing E).

traits (two warps for each of three intervein regions) were measured for each mutation, a significance level of 0.01 was chosen as a conservative indicator that wild type alleles differ in their complementation of the mutant wing shape defect. This results in rejection of the null hypothesis of no effect for three loci for IVR-B, four loci for IVR-D, and seven loci for IVR-C. For the remaining loci, there is no evidence that wild-type variation has a quantitative effect on wing shape.

Neither *wingless* nor *engrailed* emerged as good candidate modifiers of wing shape, despite the overall effect of mutations at these loci on all three IVRs. Consequently, the *Sternopleural* allele of *wingless* on the *PmSp* marker chromosome is unlikely to be responsible for interaction effects detected with other second chromosome loci. Similarly, the loci encoding the putative EGFR ligands *vein, spitz* and *gurken* as well as the co-factor *rhomboid* can be excluded as good candidate modifiers of wing shape in our sample of six wild type *D. melanogaster* lines.

The central and anterior portions of the wing, represented by IVR-C and IVR-B respectively, may be affected by variation in TGF-β activity, as both *dpp* and *tkv* show similar effects on both warps of these regions. The *EGF Receptor* also gave a positive result in these wing regions, as well as in the posterior IVR-D. Two different EGFR mutations were tested against two different tester chromosomes, and significant interaction terms were detected in all four cases. Since statistical power studies of quantitative complementation tests have not been performed, it is not clear whether the observed differences in significance level are real, and hence whether there is allele-specificity to the interactions. Significant results for the repressor *argos* provide further support for the involvement of the EGF pathway in quantitative regulation of wing shape. In IVR-D, two different hypomorphic alleles of *hedgehog* and a mutation and Deficiency affecting *elbow* had strong interaction effects. In most of these cases, the significance of the interaction term is clearly attributable to one or two of the six lines, as visualized by the crossing of line means in the plots shown in Fig. 2.2 B, C and D.

There are two major caveats to quantitative complementation tests that must be considered. Ideally the test should be performed after introgression of just the candidate mutation into a common wild type tester background by repeated backcrossing so that as little as 5% of the genome is tested (Mackay and Fry 1996), rather than a whole chromosome as here. As a screening method, and dealing with homozygous lethal mutations, this is impractical. Our version of the quantitative complementation test must thus deal with the possibility that either the mutation-bearing chromosome or the marked tester chromosome (for example, *PmSp*, or TM6) also carries a mutation that affects the trait. The latter is controlled to some extent by utilizing the same tester chromosome for several mutations. While significant $G \times L$ interactions may be due to polymorphisms other than the identified mutation, negative results exclude the wild type alleles opposite the mutation as a source of quantitative variance and are thus useful

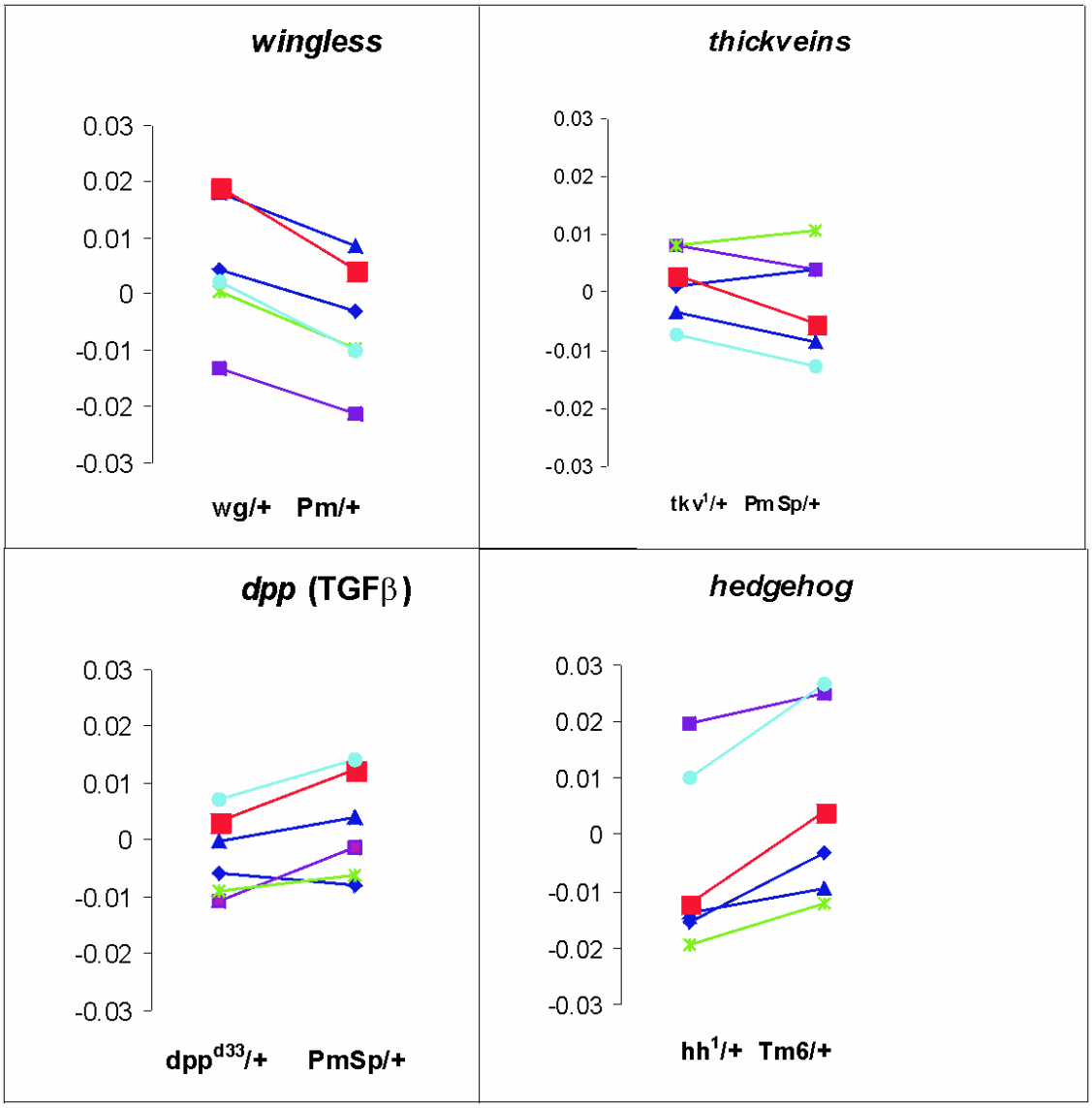


Figure 2.2 Plots of line means from quantitative complementation tests. The symbols represent the mean relative warp values in mutant and tester genotypes from 4 experiments. In each case, a significant genotype effect is indicated by the non-horizontal lines joining means. A. The lines for IVR-D, warp 2 in the *Df(2L)J-H* (*wg*$^-$) cross are nearly parallel, indicating the absence of any genotype by line interaction effect ($p = 0.35$). B. Two lines (A18 and W29, green crosses and blue diamond) show an increase in relative warp 1 for IVR-C over the tester relative to the *tkv*^{1} mutant, whereas each of the other lines show a decrease. Consequently, there is crossing of line means, which indicates a genotype by line interaction effect, which is significant from the ANOVA ($p = 0.0004$). C. Similarly, IVR-C warp 1 for *dpp*^{d33} shows a significant interaction effect ($p < 0.0001$) due solely to A18. D. For *hh*^{1} IVR-D warp 1, lines A3 and Oregon R produce the significant interaction ($p = 0.005$).

Table 2.1 Significance of genotype by line interaction terms from ANOVAs of relative warps.

Candidate Gene	Allele	Stock	Tester	IVR-B		IVR-C		IVR-D	
				B1	B2	C1	C2	D1	D2
decapentaplegic	*Df(2L)dpp^{d33}*	Bellen	*PmSp*	*	**	***	*	.	.
thickvein	*tkv^{1}*	B-427	*PmSp*	**	.	***	.	.	.
wingless	*Df(2L)J-H*	B-1357	*Pm*	.	.	.	.	.	.
engrailed	*Df(2R)en-A*	B-190	*PmSp*	.	.	.	.	.	.
			Gla	.	.	.	.	.	.
EGF Receptor	*Df(2R)Pu-D17*	B-2606	*PmSp*	.	***	.	.	***	**
			Pm	.	**	***	.	***	.
	Egfr^{f2}	B-2768	*PmSp*	**	.	.	***	.	.
			Pm	.	**	.	.	.	.
rhomboid	*rho^{ve-1}*	B-628	TM6	.	.	.	.	.	.
vein	*vn^{10567}*	B-P1749	TM6	.	.	.	.	.	.
spitz	*spi^{1}*	B-1859	*PmSp*	.	.	.	.	.	.
gurken	*Df(2L)N22-14*	B-2892	*PmSp*	.	.	*	.	.	.
	grk^{2B}	Schüpbach	*PmSp*	.	.	.	.	.	.
argos	*argos$^{\Delta 7}$*	B-1004	TM6	.	.	.	**	**	.
ventral veinless	*vvl^{sep}*	B-822	TM6	.	.	.	.	.	.
messy	*Df(3R)ry615*	B-3007	TM6	.	.	.	***	.	.
	mesA^{1}	B-4279	TM6	.	.	.	.	.	.
elbow/wb	*Df(2L)osp29*	B-3078	*PmSp*	.	.	***	***	**	**
	elB^{9}	B-4743	*PmSp*	.	.	.	.	**	.
hedgehog	*hh^{1}*	B-450	TM6	.	.	***	.	**	***
	hh^{2}	B-3376	TM6	.	.	.	.	*	*

* $0.01 < P < 0.005$ ** $0.005 < P < 0.0005$ *** $P < 0.0005$. non-significant

for screening candidate genes from further study, and for fine-structure mapping using overlapping deficiencies (Gurganus *et al.* 1999).

The second caveat concerns interpretation. The inference that a significant *G*×*L* interaction term indicates complementation of the degree of dominance of the mutation by wild type variation opposite the lesion is parsimonious. The most obvious alternative is that interactions are produced by epistatic interactions between wild type alleles anywhere in the genome, and the mutation. Fine structure QTL mapping suggests that epistatic interactions make little overall contribution to wing shape variation relative to the additive genetic variance, but that they may nevertheless be prevalent, tending to cancel one another out (Weber *et al.* 1999). The dominance and epistasis models cannot be distinguished with the current experiments. Whether the significant interaction terms are due to dominance or epistatic interactions, our results are nevertheless consistent with an ability of the regulatory genes *dpp*, *tkv*, *EGFR*, *argos*, *elbow* and *hedgehog* to contribute to standing variation for wing shape. As with bristle number, which has been shown to be modified by wild type variation in genes involved in neurogenesis (Mackay 1996), wing shape appears to be modified in part by genes identified by classical genetic methods.

Atavistic Venation

It was not possible to score wing shape in most crosses involving *Df(2R)Px2*, due to the highly variable penetrance and expressivity of the appearance of ectopic veins and wing blisters covering up to two thirds of the wing blade. This deficiency removes cytological bands 60C6 to 60D9, uncovering the *SRF*/*blistered* locus, which encodes a transcriptional repressor of vein differentiation (Montagne *et al.* 1996) and has previously been shown to have venation and blistering phenotypes (Roch *et al.* 1999). The pseudoallelic locus *Plexate* is also removed by this deficiency. Wild type genetic backgrounds clearly affect the phenotype of *Df(2R)Px2* hemizygotes, and blistering is much more severe in females than males (data not shown). One particular combination, AA18 / *Df(2R)Px2* produced a genetic balance in males (Fig. 2.1D-F) that allowed us to extrapolate the positions at which ectopic veins tend to form (Fig. 2.1B). We were able to stabilize this phenotype to some extent by backrossing *Df(2R)Px2* into AA18 for three generations, with selection for ectopic veins but lack of blistering. Two replicates of this introgression gave similar responses as documented in Table 2.2, including the appearance of a fraction of females that show the same phenotype. In these lines, the frequency of short vein fragments also increased, though there was no consistent pattern to these and they are considerably less frequent that the 6 ectopic veins indicated.

The Drosophila wing is highly derived and differs from the plesiomorphic condition through the loss of at least a half dozen veins (Comstock and Needham 1898-99; Stark *et al.* 1999).

Table 2.2 Percentage of *Df(2R)Px2*/AA18 wings showing ectopic veins [1]

		Male		Female	
Vein 2	A18	Intro1	Intro 2	Intro 1	Intro 2
A1	95	69	75	74	76
CuA2	90	69	75	22	46
R3	63	43	37	15	38
M2	58	35	45	30	32
M3	58	31	52	11	22
distal cv	68	25	23	7	19
N	19	67	65	27	37

[1] A18 refers to F1 progeny of the cross of *Df(2R)Px2*/SM5 to the near isogenic line A18. Intro 1 and Intro 2 refer to replicate 3 generation introgressions of the deficiency into A18 with artificial selection for ectopic veins and against wing blistering.
[2] See Figure 1 legend for vein identities. N is the number of wings scored.

Numerous authors have homologized the remaining veins as summarized in Fig. 2.1C, and it is often assumed that the longitudinal veins represent fusions of two adjacent ancestral veins after the loss of intervein tissue. In recent years, analysis of the location of ectopic vein tissue in mutants such as *net* and *plexus* has led to the alternative proposal that several veins are simply repressed, failing to form at boundaries of gene activity that still exist in Drosophila (Thompson 1974; Sturtevant *et al.* 1997, Biehs *et al.* 1998, Bier 2000). Our analysis of *Df(2R)Px2* provides direct support for this conclusion as each of six ectopic veins that form lie in positions where they would be expected if the ancestral condition is simply projected onto the Drosophila wing (Stark *et al.* 1999;Fig. 2.1C), so the phenotype should be regarded as atavistic. The ectopic veins include a distal crossvein in IVR-D, and five ectopic longitudinal veins. The only consistent exception is the lack of evidence for an extra vein primordium in the central region of the wing, although evidence for its presence can be seen in certain *plexus* mutants (Thompson *et al.* 1980).

It is not obvious why a disused prepatterning mechanism for vein formation would be conserved over one hundred million years (Powell and De Salle 1995), unless it plays an integral part in some other aspect of wing morphogenesis. Whatever the reason, its persistence and the observation that atavistic vein phenotypes can be produced by single mutations, implies that the evolution of homoplasic wing patterns may not be uncommon in dipterans. In addition to describing the mechanisms of phenotypic change, developmental studies should thus also contribute to a better understanding of the general tempo and mode of morphological evolution.

Acknowledgments

We thank Trudy Mackay for introducing us to the quantitative complementation testing method, and Hugo Bellen, Trudi Schüpbach and the Bloomington Stock Center for supplying fly stocks. This work was supported by a Fellowship from the David and Lucille Packard Foundation to G.G.

Chapter 3

Nucleotide variation and linkage disequilibrium in *EGFR* in three populations of *Drosophila melanogaster*

Abstract

As a step in my study of the relationship between nucleotide polymorphisms contributing to natural variation and the evolutionary forces molding this variation over evolutionary time, this chapter describes the distribution of sequence variation in 245 alleles of the *EGFR* of *D. melanogaster* from three populations. The protein is nearly invariant like other components of the Ras pathway, with only 5 high frequency variants segregating in the ~1400 amino acid protein, thereof 4 in a putative signal peptide of alternate 5' exon 1. Fixed synonymous changes are also absent from this exon, resulting in rejection of neutrality by a MK test. The other alternate exon is invariant between species and has only 3 rare replacements segregating. Other tests of deviation from neutrality were not significant after correcting for experiment wide tests. However, several indications that weak purifying selection reduces the neutral mutation rate include: (i) a relatively slow rate of protein evolution; (ii) a skew towards negative Tajima's D values particularly in non-coding regions; (iii) excess of local negative deviations in Fu and Li's F* and D* (iv) absence of large indel polymorphisms at high frequency; (v) fixation only of short indels in reference to *D. simulans*; and (vi) the restricted length distribution of a microsatellite in promoter 2. Linkage disequilibrium as summarized by r^2 falls rapidly with distance along the locus and does not differ between two North American populations from North Carolina and California. *D'* extends further along the locus, due mainly to coupling of rare sites to older polymorphisms. Comparison of both the spectrum of allelic frequencies and number of private alleles between the North American populations and a Kenyan sample highlighted distinctness of the Kenyan population. The two North American samples were essentially identical with the exception of handful of sites along the locus which showed high F_{ST}, all of which are in pair-wise linkage equilibrium.

Introduction

Determination of the forces that shape the evolution of individual loci is a step towards evaluation of the genomic distribution of functionally important polymorphisms. Are polymorphisms having significant effects on phenotypes predominantly located in highly conserved regions of a gene, or are a significant fraction found in regions of relaxed selection? The distribution and the fitness effects of these quantitative trait nucleotides within a locus will affect the extent and significance of linkage disequilibrium (LD) between sites, and reciprocally the level of recombination will influence the distribution of QTN's. LD is also of considerable importance for the practice of mapping traits by association to scored markers in one or more study populations. It follows that stratification of populations can impact both the patterns of nucleotide variation and levels of LD with consequences for surveys of variation and association or linkage mapping. Here I describe the molecular evolution and population genetics of the Drosophila Epidermal Growth Factor receptor gene (*EGFR* or *DER*), which evidence suggests contributes to variation in wing shape and eye phenotypes (Palsson and Gibson 2000, Zimmerman *et al.* 2000, Polaczyk *et al.* 1999), as a step towards illuminating the relation between functional variation and selection.

The question of how selection operates on larger cohorts of interacting proteins like signaling cascades has received attention recently (Jeong *et al.* 2000, Olsen 2002, Nijhout 2002, Riley et al 2003). Is the intensity of selection related to the placement of a gene in a functional hierarchy, or will purifying selection be strongest on proteins occupying key positions in intracellular pathways, like p53 which forms a hub in the network governing cell division (Vogelstein *et al.* 2000)? Analysis of proteins involved in inflorescence decision and floral development in *Arabidopsis thaliana* is consistent with the selection primarily preserving the upstream components (Olsen *et al.* 2002). Support for the latter hypothesis was provided by studies on the Ras/MAPK pathway in D. melanogaster (Riley *et al.* 2003). A particularly striking observation is the absence of protein polymorphisms in the three Ras genes (Gasperini and Gibson, 1999). The signature of purifying selection will presumably be most intense at the level of protein polymorphism, but might also impose constraint on non-coding DNA by requiring stringent regulation of protein level through transcription and translation (Ludwig *et al.* 1998, 2000, Bergman and Kreitman 2001, for a review see Ludwig 2002).

DER, the product of the EGFR locus, is one of eight Receptor Tyrosine Kinases (RTKs) in flies. These proteins are membrane bound receptors that form a dimer upon ligand binding, resulting in trans- and auto-phosphorylation of specific tyrosine residues. This activates an intracellular signal mediated by Ras and a cascade of kinases resulting in, for example, cell division, differentiation or migration. DER and Sevenless are considered the key receptors for the Drosophila MAPK cascade (Rebay 2002; Held 2002) with the former having a wider range

of defined functions and more extensive expression during development and adulthood (Lev *et al.* 1985, Schejter *et al.* 1986 and Kammermeyer and Wadsworth 1987). EGFR encodes 6 exons, and has two splice-forms differing by 2 alternate 5'-exons (Figure 3.1). The alternate exons are spaced 24 kb (exon 1 – the RA transcript) and 3 kb (exon 2, transcript RB) upstream of the common exons, 3 through 6. The expression of DER is spatially and temporally dynamic (Lev *et al.* 1985, Schejter *et al.* 1986 and Kammermayer and Wadsworth 1987) but the regulatory regions remain uncharacterized, as all previously characterized mutations impact the protein coding region (Clifford and Schüpbach 1994, Lesokhin *et al.* 1999).

Two features of molecular variation within a locus have considerable bearing on the practice of genotype-phenotype mapping. First, the pattern of Linkage Disequilibria between single nucleotide polymorphisms (SNP's) within a locus or genomic region affects the process of mapping quantitative trait nucleotides by association to scored markers. In practical terms, how much information does a typed marker cede about linked, and putatively contributing, polymorphisms? This is best addressed with extensive genotyping along a candidate locus or region, facilitating contrasts either between populations or loci (Remington *et al.* 2001). Second, the degree of population subdivision has implications for mapping complex traits (Pritchard and Rosenberg 1999, Remington *et al.*, 2001). Historical relatedness can be captured with the F_{ST} parameter, which assesses deviations in allele frequencies between populations (Wright 1969, Weir 1996, Weir and Hill 2002). An alternate approach has been developed by Pritchard and Rosenberg (1999), where sub-populations are identified by clustering subjects into groups within which LD is minimized and violations of Hardy Weinberg equilibrium are avoided.

With an eye toward using the Drosophila system to investigate some of the statistical genetic complexities involved in mapping human disorders or agricultural traits in structured populations, I undertook a study of the molecular evolutionary and population genetic parameters of *DER*. This was done by analyzing the sequence variation in 10.5 kb of DNA, corresponding to protein coding and flanking regions in 250 alleles of from 3 populations, 2 North American and an African population. Molecular evolutionary parameters are used to investigate the evolutionary forces acting on coding and non-coding regions of the locus. A detailed description of LD within the locus allows assessment of the independence of segregating sites. Analysis of population subdivision in terms of the frequency and distribution of polymorphisms was studied with contrasts between the 3 populations. Finally, I investigate the relationship between linkage disequilibrium and population subdivision in the *EGFR*.

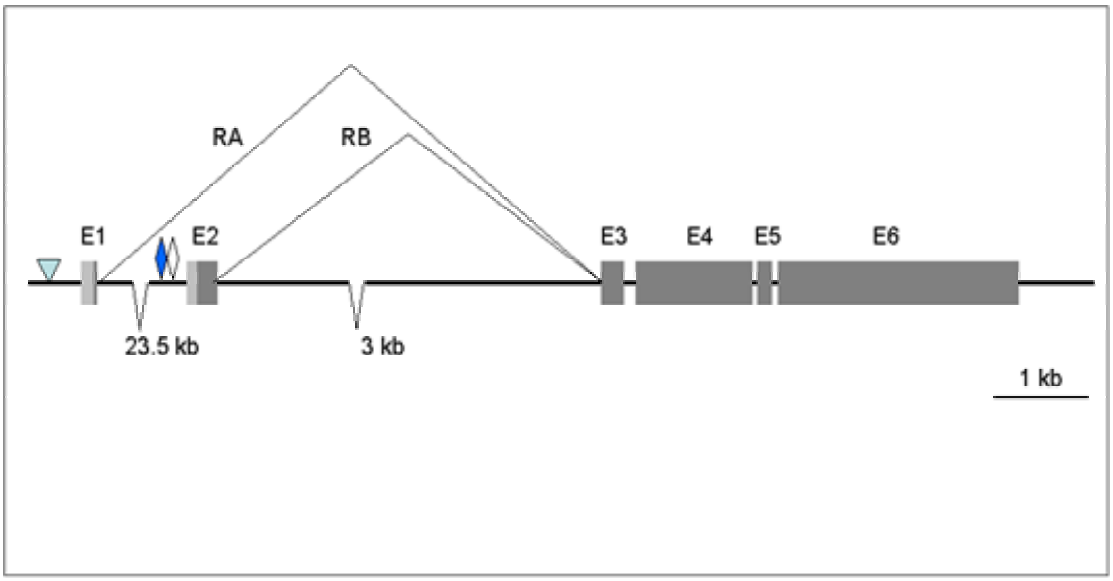


Figure 3.1. Structure of the sequenced region of *EGFR*, with the length of non-sampled regions designated below. Total sequenced regions spans 10.9 kb. Exons are designated by boxes labeled sequentially (E1-E6). E1 and E2 are alternatively spliced giving rise to the RA and RB isoforms. The lighter gray parts of E1 and W2 correspond to predicted signal peptides. Triangle: Location of Pogo element insertion in the RA promoter. Blue diamond: A putative GAGA factor binding site, see text for details. White diamond: the only microsatellite in the sampled region. Both elements are in the RB promoter.

Materials and methods

Fly handling and populations

Alleles of *EGFR* came from 3 populations of Drosophila melanogaster, two from opposite coasts of the USA and one from Kenya. The 36 Kenyan stocks were obtained from the Bowling Green Stock Center before it closed, in 1997, but were trapped by R. Woodruff around 30 years ago. 2nd chromosomes from this population carrying the candidate locus were substituted into a common genetic background, Samarkand, by way of a standard three generation cross. Sergey Nuzhdin contributed 83 lines collected in the Wolfskill orchard in Winters near Davis California in 1998, which underwent 40 generations of inbreeding (Yang and Nuzhdin, 2003). A second North American panel was established from a collection of 150 isofemale lines from a peach orchard in West End North Carolina in the summer of 2000. Sib-mating for 15 generations resulted in 70% homozygosity, as monitored by sequencing of *EGFR* and other loci (data not shown). Heterozygous lines were inbred for 5 more generations and sequenced for parts of EGFR in both parents of a sib- cross allowing us to select homozygous lines for further analysis and experiments. A total of 130 West End lines survived; 15 retained heterozygosity for a part of the locus. Hereafter, the populations will be referred to and abbreviated as Kenyan (K), UC Davis (UC) and West End (WE). For inter-species comparisons we sequenced one *D. simulans* line (WE143) collected at the same time and locality as the West End lines. In order to resolve an unusual pattern of polymorphism level and divergence in exon 1 we added several more *D. simulans* alleles (Florida CS from Marla Sokolowski, and TT01TS, NC112T, MD106TS, MD225T, described in Ballard 2000); and one *D. sechellia* line (BG-1, originally collected by Issacs and Bachi 1981 and provided by Bill Ballard). A sample of 5 isofemale *D. simulans* lines from Raleigh NC was collected by Richard Lyman in early 2003.

Genotyping

DNA was obtained from individual male fly preparations except for marker assisted inbreeding where both parents were surveyed. Regions corresponding to the 6 exons and flanking regions of *EGFR* were amplified in 6 PCR fragments (Figure 3.1). See Appendix A for primer sequence and locations. The PCR products ranged from 1.2 kb to 2.1 kb spanning 10.9 kb of the locus in 3 contiguous fragments. Promega *Taq* Polymerase was used for amplifications and Perkin Elmer Big Dye mix and enzymes for sequencing reactions. Sequencing reactions were run on Perkin Elmer 377 (Kenyan sample) and 3700 (WE and UCD samples) automated sequencers stationed in the NCSU Genome Research Laboratory. Trace-files were incorporated into the Contig-Express module of Vector NTI 5 (Informax, Boston MA, www.informax.com) for primary editing and construction of contiguous alleles for each line. Sequence alignment was conducted

with Clustal W (Thompson *et al.* 1994) and the matrix of alleles was transferred to Genedoc (Nicholas *et al.* 1997) for manual adjustment of insertion and deletion polymorphisms (indels). Each SNP and indel variant was then verified by reanalysis of trace files, and ambiguous calls were resolved or discarded. The extent of sequence coverage is ~1.5X, with PCR fragment and sequence trace overlap being 100 bp on average and 18% of the sequenced regions represented by two or more reads. This allowed estimation of PCR errors, which proved 9minimal, as 4 PCR errors were identified in the 0.9 Mb of sequence of multiple reads and overlaps. These errors were all singletons and no discrepancies were observed in polymorphic sites of higher frequency. The base calling error due to scoring was estimated to be 0.036% per polymorphic site and 0.0012% across the whole dataset.

The genotype matrix for the 257 *D. melanogaster EGFR* alleles sequenced is 74% complete, with average length of 8067 (±131) bp. There are three reasons for the incompleteness of the dataset. (i) Several West End lines were put aside during the sequencing, either because they exhibited lingering heterozygosity or went extinct. (ii) A few of the targeted regions proved cumbersome in the WE population and were therefore not sequenced as methodologically when it came to filling in the gaps or when we started sequencing the UCD population. For instance PCR amplification was only 60% successful for the whole fragment surrounding exon 2, presumably because of excessive level of SNP and indel polymorphisms. Also, the quality of sequence reads fell after running through long uniform stretches, like the C-stretch 98 bp upstream of exon 1 and A-stretches in intron 2. (iii) The Kenyan sample was considered exploratory and was not subjected to intense finishing efforts. Locations of variants are in reference to genebank entry 17571116 (Flybase number: FBgn0003731) which spans 48 kb corresponding to the *EGFR* locus and ~5 kb on either side. Alleles will be submitted to the Genebank database.

Parameters of molecular evolution

The average nucleotide diversity per site between pairs of alleles π, and Watterson's θ were estimated for point mutations (Wayne and Simonsen 1998), both with DnaSP Version 3.53 (Rozas and Rozas 1999) and Tassel (www.maizegenetics.org). Tassel handles missing data by weighting the estimates of the parameters (Ed Buckler, personal communication). Short regions where sampling was only 5-30% of the overall sample were removed from the analysis because weighting in those regions inflated Watterson's θ. Parameters were also estimated for regions of the gene either by sliding window analysis or by predefined attributes, such as exon-intron boundaries, transcribed vs. non-transcribed DNA and the functional domains of the protein. Only minor incongruencies between Tassel and DnaSP estimates were observed. Direct tests of the neutral theory using D (Tajima 1989) and the statistic of Fu and Li (1993) with and without

outgroup were conducted in DnaSP with trimmed datasets, again only on the SNP's. The significance of statistics was determined with a coalescence module in DNAsp, with recombination parameter set at 10 and sample size ranging from 100-200 depending on the region surveyed. Except significance of sliding window statistics is based on estimated variance of the statistics (Tajima 1989, Fu and Li 1993). Implementations with total number of mutations or segregating sites did not alter the results, nor did using *D. simulans* outgroup for F* and D*. Lack of replacement polymorphism and divergence complicated the application of the MacDonald-Kreitman test (1991) for the DER protein by domains, because the G-test can not cope with empty cells. Fisher's Exact and the Chi-square test can tolerate one empty cell and were used to assess deviations from neutrality for exons 1, 2 and the data from exons 3-6. The neutrality index of Rand and Kann (1996) was calculated by hand for those same regions. The HKA test (Hudson *et al.* 1987) was conducted for exons vs. introns and also in an *ad hoc* manner to contrast exon 1 and the remainder of the protein. Both tests were implemented with the tools module in DNAsp. Divergence to *D. simulans* was assessed in DnaSP, while the identity to *D. pseudoobscura* (raw contig1071) was analyzed with AVID and visualized with VISTA (Mayor *et al.* 2000, on the web at http://www-gsd.lbl.gov/vista).

Linkage disequilibrium and population subdivision

The significance of Linkage Disequilibrium (LD) was assessed with Fisher's-exact test (Lewontin 1988) implemented in Tassel with the multiple comparisons issue addressed by a randomization and retesting procedure. The squared allele-frequency correlation r^2 indicates the level of independence between polymorphic sites, and D' expresses the relative value of D, the deviation in frequency of heterozygotes, in reference to the maximum (or minimum given the sign) D possible for the specific sites tested (Langley *et al.* 1974, Weir 1996). The signs of D and r are usually considered arbitrary, except if the loci are oriented into coupling and repulsion phase according to the two allele frequencies (Langley *et al.* 1974). DnaSP was used to estimate the D and r parameters for individual regions of the locus. The distribution and relationship of D and r with distance were analyzed with Proc TTEST and Proc REG in SAS. All LD analyses were performed on SNP's and Indels separately with both classes revealing similar patterns. LD estimates were computed for datasets from the North American populations separately and on a combined dataset. The cutoff was 10% for the minor allele with minimum sample size per site at 30 for the UCD and 50 for the WE population, reflecting the sizes of the initial samples. This was done because tests of LD are incapable of detecting significance for pairs of sites where one or both are at low frequency (see Lewontin 1988 for tabulation of minimum frequencies required to get significant test statistics). Comparisons between populations must also consider the direct effects of allele-frequency on r^2, D' and the

significance of the exact tests (Lewontin 1988, Weir 1996). A comparison was conducted by testing for correlation (Proc CORR in SAS) between parameters (r^2, D') by site-pairs represented in both populations. Distances between sites were not adjusted to account for insertion deletion polymorphisms as these were generally small. Due to small sample sizes the Kenyan sample was analyzed separately.

The level of population differentiation at the *EGFR* was estimated by the AMOVA feature in Arlequin 2.0 (Schneider *et al.* 2000), returning an estimate of the differences in SNP or haplotype frequency as summarized in the parameter F_{ST}. Significance of the parameter for each individual test was estimated by 10,000 permutations, and experiment-wide significance was achieved by Bonferroni correction. Subdivision was estimated at the level of individual polymorphic sites, and by sliding a window of haplotypes spanning 5 or 10 polymorphic sites along the gene. F_{ST} partitions the contribution of within population to between population variance in allele frequency, and ranges from 0 to 1. But since the procedure can yield negative estimates of F_{ST}, particularly if the true *Fst* is close to 0, negative values were adjusted to zero. The F_{ST} was estimated for all combinations of the 3 populations under study, both on all SNP's (547 total) and with rare (<.05) variants (201 sites) excluded. Locus wide significance of the F_{ST} estimates was determined with the Bonferroni, Dunn-Sidak and Hummel corrections in Proc MULTTEST in SAS 8.02 (SAS Institute 2001). Linkage disequilibrium around sites that differed between UCD and WE was extracted from LD analysis on the distinct population datasets. All sites were present in the 3 populations with the exception of site 35345 which was only present in two of 74 Californian alleles making comparison impossible. Correlations between r^2 in the populations were assessed with Proc CORR in SAS.

Results

Nucleotide polymorphism and divergence

What are the patterns of nucleotide diversity in the canonical member of the EGFR/Ras pathway? This was addressed by sequencing 10.9 kb of the ~38 kb constituting *EGFR*, focusing on the coding regions and promoter sequences while omitting two stretches, 23.5 kb of intron 1 and 3 kb of intron 2 (Figure 3.1). The general parameters of molecular diversity for the 523 di- and 24 tri-nucleotide polymorphisms are presented in Table 3.1 and are within the range for Drosophila genes (Kreitman and Hudson 1991, Powell 1996, Richter *et al.* 1997, Zurovcova and Ayala 2002).

Protein evolution

Gasperini and Gibson (1999) and Riley *et al.* (2003) showed that core components of the EGFR/Ras pathway in *D. melanogaster* is short (*Drk* and *polehole*) or devoid (*Ras1-3*) of

Table 3.1. Descriptors of nucleotide variation in regions of *EGFR* in *D. melanogaster* and differences (fixed) compared *D. simulans*.

	Length[1]	Location[2]	Segregating Polymorphisms						Differences		
	(bp)	Start/End	π	θ	TajD	Syn	Rep(rare)	Indel	Syn	Rep	Indel
5'- E1	605	5402	0.0132	0.0180	0.0212	41	-	7	13	-	7
Exon 1	153	6016	0.0079	0.0090	1.1247	7	4 (2)	1[(3)]	0	6	1[(4)]
3'- E1	492	6170 6552	0.0037	0.0111	0.1801	20	-	8	ND	-	ND
5'- E2	416	30120	0.0020	0.0259	-0.7620	16	-	9	13	-	1
Exon 2	300	30518	0.0042	0.0383	-1.0899#	21	0 (3)	0	5	0	0
3'- E2	1384	30819 32168	0.0042	0.0405	0.1581	78	-	16	27	-	8
Intron 2	2425	35340	0.0034	0.0122	-0.9577#	102	-	14	25	-	2
Exon 3	222	37757	0.0136	0.0122	0.3978	13	0 (0)	0	6	0	0
Intron 3	170	37980	0.0404	0.0480	1.0082	30	-	5	10	-	1
Exon 4	1174	38116	0.0088	0.0124	-0.1079	54	0 (1)	0	22	0	0
Intron 4	66	39291	0.0097	0.0197	-1.3993*	6	-	3	6	-	0
Exon 5	132	39358	0.0114	0.0217	-1.1061#	11	1 (1)	0	0	2	0
Intron 5	74	39491	0.0047	0.0105	-1.0660	2	-	1	7	-	1
Exon 6	2448	39561	0.0080	0.0082	-0.6577	95	0 (8)	1[(3)]	44	2	0
3'-UTR	347	42010	0.0078	0.0170	-0.6735	18	-	3	6	-	1
Intergenic	455	42355	0.0022	0.0064	-1.0886#	13	-	2	9		0
Total	10863	42804	0.0083	0.0134	-0.206	527	5(15)	70	193	10	22

1. Length indicates size of region after accounting for insertions in reference to Genebank document: 17571116.
2. Location marks the start and end (below) of each of the 3 sequenced fragments.
3. Three base indel in exon1, resulting in two tandem start codons in all but two alleles. The extra Methionine is the derived state. Nine base indel in exon 6, resulting in 4 amino-acids (Pro-Asn-Asn-Asn) being replaced by a single Histidine.
4. A Methionine codon insertion in the D. simulans lineage or loss in the *D. melanogaster* lineage.
5. Differences, in reference to a *D. simulans* allele. ND, region not sampled.

Significance of Tajimas D, by coalescence simulation. # $0.05 < P < 0.1$ and * $P < 0.05$.

Syn: synonymous or non-coding polymorphism, Rep: replacement polymorphism, rare: frequency of allele less than 0.05.

protein polymorphism while downstream (*Dsor1*) or auxiliary proteins (*Ksr* and *corkscrew*) are evolving more rapidly. In this context it is interesting to know how the key receptor of the pathway evolves at the protein level. The frequency of fixed and segregating replacement polymorphisms were relatively low in the ~1400 amino-acid *EGFR* (Table 3.2). Four of the five replacement polymorphisms observed at moderate frequency localize to the predicted signal peptide encoded by exon 1, and the remaining one is in exon 5. The four common replacements in exon 1 do not constitute a single haplotype (data not shown). The remaining 15 replacements were only present in 1, 2 or 4 alleles, and may potentially have deleterious effects as about half of them alter amino acids that are conserved between *D. melanogaster* and *D. pseudoobscura* (data not shown). Purifying selection at the amino-acid level is also suggested by the contrast to *D. simulans*, which only shows eight replacements relative to *D. melanogaster*, four of them in exon 1. Two segregating indels within coding regions were observed. An indel of 9 bp was found in exon 6 in a single Californian line (UCD61). The base change replaces part of a conserved Asparagine repeat, Pro-Asn_3, with a Histidine (site 41684, amino acid 1270). The second indel involves a methionine codon just following the initiation codon of exon 1 (site 6016) which was absent in two West End lines. Comparison to *D. simulans* shows that the change in exon 6 is obviously a recent deletion but the methionine codon change appears to be a new insertion in the *D. melanogaster* lineage. Interestingly the only fixed insertion in coding regions observed in contrast to *D. simulans* is another Methionine codon in exon 1 (amino acid position 9). The *D. sechellia* sequence is identical to the *D. simulans* allele in that regard. The net result is that the first exon in those species includes four methionine codons among the first ten codons. Those are then followed by a hydrophobic stretch and an acidic domain rounding up the predicted signal peptide. Six indel differences in exons 3 to 6 are seen in reference to *D. pseudoobscura.* Comparison of the alternate 5'-exons are complicated by the rapidity of divergence with exon 1 in D*. pseudoobscura*, which could neither be aligned by standard algorithms nor by hand. The facts that the region does contain several start codons as well as a splice recognition sequence, and that there are conserved sequence stretches in the promoter and downstream intron, suggest that the exon is present. This is demonstrated by the sequence identity between *D. melanogaster* and *D. pseudoobscura* in this region (Figure 3.2). Note the low level of identity of exon 1 vs. exons 2 through 6. This is acutely interesting considering the other alternate N-terminus is encoded by exon 2, where the coding region can be aligned without problems and the non-coding regions also contain conserved blocks (see more below).

The MacDonald-Kreitman test (1991), which tests for adaptive protein evolution, was not significant (Fisher's exact $p = 0.587918$). The possibility remains that individual domains are experiencing different modes of selection. This was assessed by applying the MK test to individual exons. This analysis is constrained by the lack of replacement polymorphisms,

Table 3.2. EGFR replacements segregating in *D. melanogaster* (A) or fixed to *D. simulans* (B).

A)	Exon	Site #	Base[1] Ancestral	Derived	Amino Acid[2] Ancestral	Derived	AA code	Frequency[3]
	E1	6019	Insertion 3 bases		-	Met	In2	D
	E1	6034	T	C	Trp	Arg	W7R	S
	E1	6058	T	C	Trp	Arg	W15R	0.07
	E1	6065	G	T	Ser	Ile	S17I	0.08
	E1	6073	A	G	Ile	Val	I20V	S
	E1	6077	G	T	Trp	Leu	W21L	0.59
	E1	6085	C	G	Leu	Val	L24V	0.15
	E2	30563	C	A	Leu	Ile	L23I	0.04
	E2	30623	G	A	Ala	Thr	A43T	S
	E2	30704	G	C	Ala	Pro	A70P	S
	E4	38924	A	T	Thr	Ser	T396S	S
	E5	39433	A	C	Thr	Asn	T543N	0.24
	E5	39451	T	C	Phe	Ser	F549S	D
	E6	39571	T	C	Phe	Leu	F566L	S
	E6	39594	A	G	Ile	Met	I573M	S
	E6	40231	T	A	Leu	Met	L786M	S
	E6	40672	A	C	Ser	Arg	S933R	S
	E6	41241	T	G	Cys	Trp	C1122W	S
	E6	41520	G	T	Glu	Asp	E1215D	D
	E6	41684	Deletion 9 bases	-	Pro-Asn-Asn-Asn	His	Del1270	S
	E6	41743	G	C	Ala	Pro	A1290P	S
	E6	41818	C	G	Leu	Val	L1315V	S

B)	Exon	Site #	Base *Sim*	*Mel*	Amino Acid *Sim*	*Mel*	AA code
	E1	6040	Deletion 3 bases		Met	-	Del9
	E1	6056	C	T	Ser	Leu	S14L
	E1	6073	A	C	Leu	Ile	L20I
	E1	6082	C	A	Ile	Val	I23V
	E1	6103	C	A	Leu	Met	L30M
	E1	6116	C	T	Thr	Ile	T34I
	E1	6127	A	T	Thr	Ser	T39S
	E5	39468	T	G	Ser	Ala	S555A
	E5	39481	T	A	Phe	Tyr	F559Y
	E6	40144	G	T	Ala	Ser	A757S
	E6	41450	T	C	Val	Ala	V1192A

1. The exact location (in exon and genebank record), and identity of the base change are listed along with the resulting amino acid replacement. There are 3 indels, with the same information.
2. The AA code designates the change and location in reference to the RA isoform, except the 3 rare replacements in exon 2 that constitute the RB isomere.
3. Frequency of the segregating variants is given in the last column, where S: singleton, D: doubleton and other sites by relative frequency of derived allele.

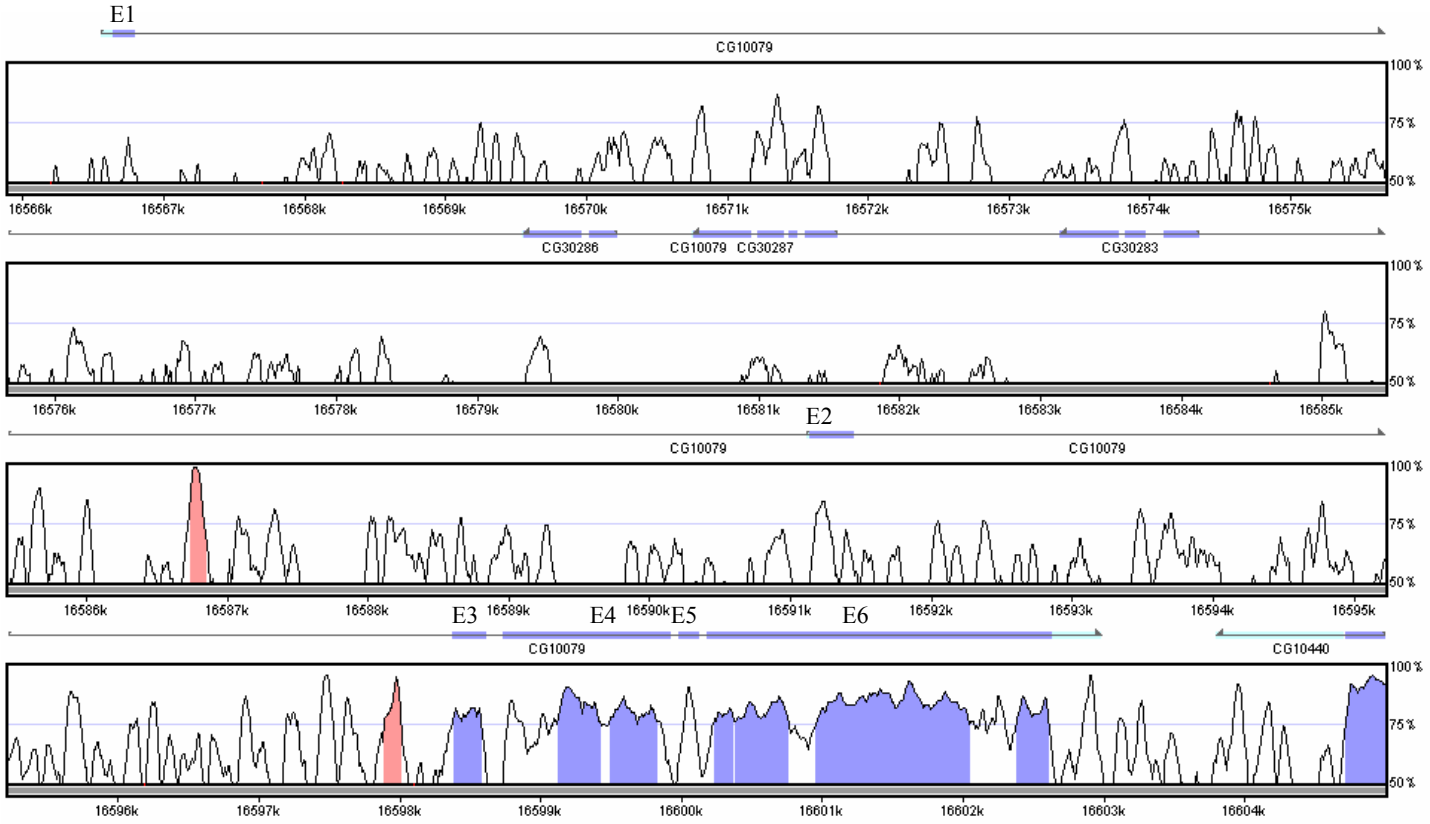


Figure 3.2. Sequence identity ranging from 50-100% in the 48 kb region including *EGFR* (CG10079) in an alignment *D. melanogaster* and *D. pseudoobscura*. The four panels represent the order from upstream (top left corner) to the downstream (bottom right corner) of the EGFR sequence in *D. melanogaster.* Intron 1 includes three predicted open reading frames in *D. melanogaster*, and 1 kb downstream of EGFR a conserved locus CG10440. Transcripts and directions are designated by arrows, exons by blue boxes and 3'untranslated region by light blue coloring. The six EGFR exons are represented by the labels E1…E6 above the blue boxes, with the alternate exons 1 and 2 being spaced away from the four main exons. Coloring under the line represents high conservation, blue for coding regions and pink for non-coding regions.

segregating and divergent, resulting in contingency tables not being computable (for exon 3, as two cells are empty) or reducing power. Also, single empty cells in exons 1, 2, 4 and 5 prevented the use of G-tests. Exon 1 was significant by exact test (one tailed p = 0.034056) and Chi square test p = 0.02371 except with Yates correction p = 0.08009. The distribution behind significance is most peculiar, six and seven replacements and synonymous respectively are segregating in *D. melanogaster* with the differences being restricted to six replacements changes. This lack of synonymous divergence is probably the reason for significance of the MK test for this exon. The Neutrality index (Rand and Kann 1996) for E1 is 0.142857. The other exons are not significant, E2 p = 0.553914 for exact one tailed and E3-6 lumped p = 0.541383 for exact one tailed. NI for E3-6 is 1.144509 and the whole locus is 0.766169. Although exon 1 appears to deviate significantly from neutral expectation, correction for the experiment wide number of MK tests by Bonferroni corrections renders it formally insignificant. Similarly the HKA test which compares the rate of evolution of exon 1 and exons 2-6 was neither significant for replacement polymorphisms alone nor for all changes in the coding region (data not shown). Also, a HKA test performed by dividing the locus into introns and exons was not significant p = 0.8919. The low level of moderate frequency protein polymorphism in the protein proper is consistent with the protein being under purifying selection in concordance with reports on other key components in the pathway (Riley *et al,* 2003, Gasperini and Gibson, 1999). However formal MK or HKA tests failed to confirm this hypothesis, suggesting that the majority of replacement evolution in Drosophila *EGFR* is neutral. Neutral drift and purifying selection are not mutually exclusive phenomena as purifying selection simply reduces the effective neutral mutation rate.

Divergence of non-coding regions

The evolution of the non-coding regions surrounding EGFR was investigated at three levels, by comparing polymorphism within *D. melanogaster*, discussed in a following section, and divergence on two evolutionary timescales. *D. melanogaster* and *D. pseudoobscura* have been separated for ~45 million years, while *D. simulans* shares more recent ancestry at 2.5 million years with *D. melanogaster* (Powell and De Salle 1995). This disparity in timescales requires distinct metrics, the level of sequence identity for long and the divergence statistic for shorter scale. The whole *D. pseudoobscura* genome sequence enabled a comparison off the entire *EGFR* locus (Figure 3.2). Three short genes of unknown function, though one shares similarity to chymotrypsin proteases, reside within the first *D. melanogaster* intron. They appear weakly conserved, if present, in *D. pseudoobscura*, while the CG10440 locus immediately downstream is clearly conserved. *D. pseudoobscura* showed 50-100% sequence identity for all of the common C-terminal coding regions, but conservation was notably reduced in the alternate 5'-exons as mentioned above. Similarity levels fluctuate along the non-coding region with several

regions exhibiting strong conservation, with two 100 bp tracts in intron 2 being 92% identical and several 20-30 bp 100% identical between the two species. Lack of divergence is also apparent in the 3'UTR and upstream of alternate exon 2. Better resolution of the sequenced regions can be seen in part A of Figure 3.3 contrasting the two divergence plots and the polymorphism levels. Note that the meaning of peaks differs between the two divergence metrics. For the sequence identity graphs, peaks indicate conservation while the opposite holds for the disparity metric and the parameters of polymorphism levels. Qualitative inspection of divergence in the non-coding regions suggest that *D. simulans* follows the same pattern as *D. pseudoobscura* as exemplified by the low level of divergence in the 3'UTR and parts of intron 2.

Silent and non-coding polymorphism

Patterns of polymorphism within *D. melanogaster* are summarized by sliding window analysis of the average number of pair-wise differences between alleles (π) as shown in Figure 3.3. The low polymorphism level in non-coding regions around *EGFR* is striking as 7 out of 9 regions have lower π than the 0.01 average for Drosophila non-coding DNA (Powell 1996). Purifying selection also influences the frequency distribution of alleles at a locus, causing an excess of low to moderate frequency variants. Tajima (1989) provides a test of this hypothesis. Negative D implies an excess of low frequency variants consistent with purifying selection, while positive values could be caused by balancing selection. The statistic for the whole *DER* is -0.206 and is not significant (p = 0.504) by a coalescence simulation based on the total number of segregating sites (Table 3.1). Gene wide departures from neutrality are rare (Powell 1996) and more fine scale analysis can reveal more interesting patterns. I broke *EGFR* up into regions corresponding to exon/intron/transcript boundaries and estimated Tajima's D for each one. The values are negative for majority of the regions analyzed (11/18), consistent with most genes in *D. melanogaster* (Powell 1996). The regions with positive Tajima's D were exons 1 and 3, parts of introns 1, 2, and the whole of intron 3. The significance was estimated for each of the predefined regions and only those with negative values approached or were significant by individual tests. Exons 2 and 5 were the only coding regions approaching significance. The non-coding regions, part of intron 2 upstream of exon 3 and the 3' untranscribed region also showed the same tendencies. However only intron 4 was formally significant by the simulation procedure (p = 0.043) and it has the most negative Tajima's D value observed -1.3993. This result was not altered by assuming absence of recombination in the region (p = 0.048). These p-values are not adjusted to count for the multiple tests, and when that is done become insignificant and the results therefore remain suggestive. Similarly the statistics of Fu and Li, and Fay and Wu (2000) were calculated for individual regions and gene wide, but none were significant (data not shown). Even finer level of analysis is achieved by calculating Tajima's D and the D* and F* statistics of Fu and Li in sliding window (100 bp wide moving 25 bp). All three

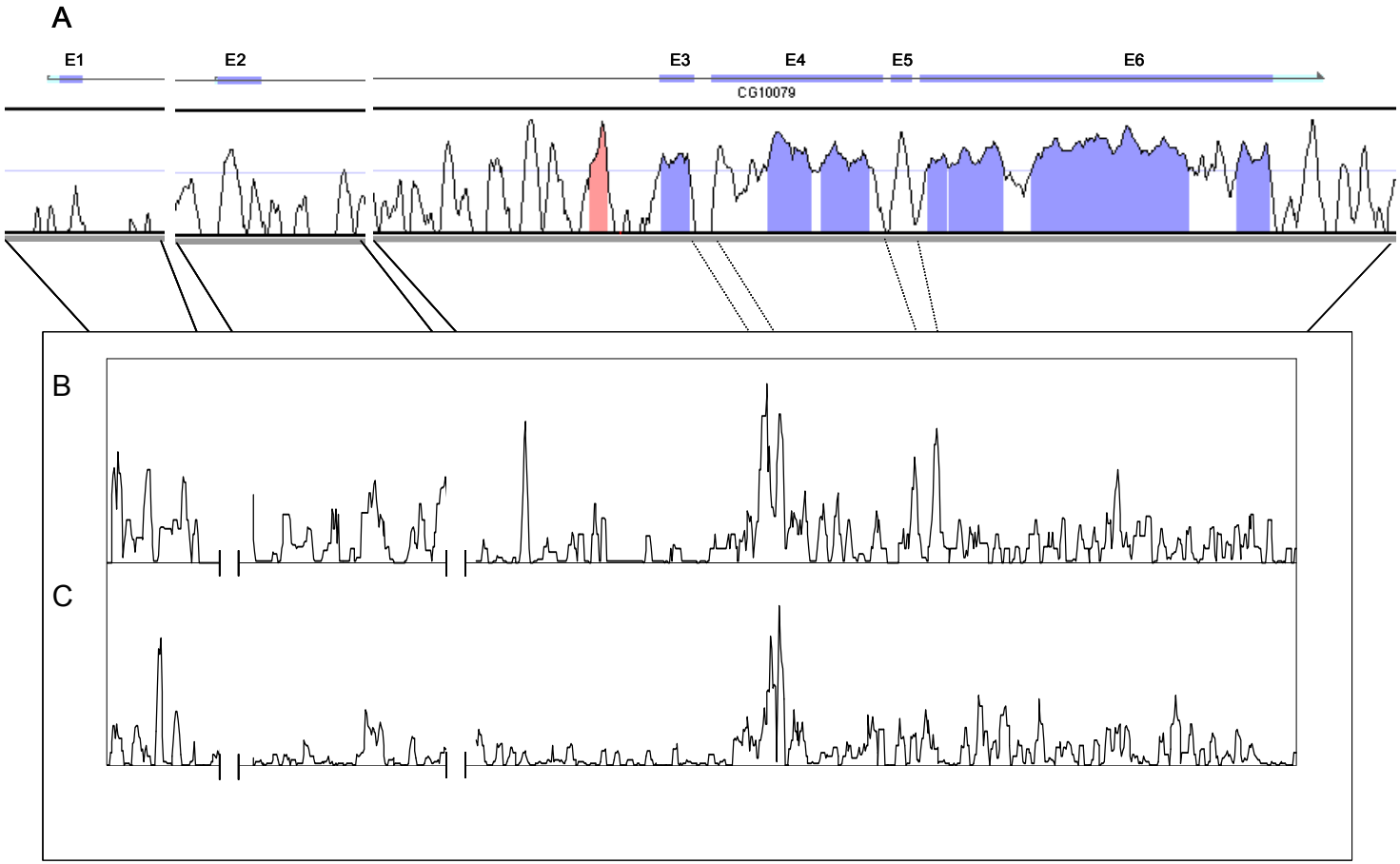


Figure 3.3. Three levels of polymorphisms and divergence along the surveyed regions of EGFR (CG10079). A) A plot of sequence identity between *D. melanogaster* and *D. pseudoobscura* genomic alignments, ranging from 50-100%. B) Divergence to *D. simulans* (axis spans 0-0.3), except for last part of region 1 due to lack of sampling and C) the within species SNP variation is represented by the parameters Pi and Theta (axis spans 0-0.15). For B and C, Parameters were estimated for 50 bp windows sliding 10 bp. Above is a schematic of the gene structure, with bold boxes representing exons and light blue the UTR. The X axis excludes parts of intron 1 and 2 not surveyed, boundary breaks indicated by black lines and introns 3, 4 and 5 by dotted lines.

statistics yielded similar results, except the latter two had greater amplitude (Figure 3.4). Twenty one of the windows have *p*-values between 0.05-0.01 and nine *p*-values between 0.01-0.001 in the 341 windows tested as estimated from Fu and Li (1993). Even though no values below Bonferroni cutoff (p = 0.000147) were observed, then the number of tests significant at 0.05 level (30) exceeds the 18 expected. It is also interesting that all *p*-values below 0.05 are associated with negative test statistics. Both of those observations are consistent with purifying selection removing high frequency variants in functionally important regions. Windows of 100 bp resolve these signatures to short segments, with the most negative values at the boundary of intron 5 and intro exon 6. Negative values are also found in intron 2 and the promoters.

The results of these tests and the Tajima's D indicate that positive and/or balancing selection are not strongly influencing variation in the *EGFR*. More qualitatively, the pair-wise differences and divergence give similar patterns along the locus, particularly for the more conserved segments like the 3' untranslated region and intron 2 (Figure 3.3). However, the correspondence does not hold for highly divergent regions, as there is a deficit of polymorphism in introns 4 and 5 in *D. melanogaster* considering the divergence, documented by the values of Tajima's D and Fu and Li's F* and D* (Figure 3.4). This is at odds with other regions with higher polymorphism levels (intron 3 and parts of introns 1 and 2). Exons 3 and 4 are separated by 1.2 kb and with LD decaying rapidly may evolve relatively independently at the population level.

Indel polymorphism in non-coding regions

Insertion and deletion polymorphisms are largely neglected in standard population and molecular evolution analysis. Since they can affect regulatory sequences, the distributions of indel size and frequency may provide independent descriptors of the evolutionary forces molding variation at a locus. Bergman *et al.* 2003 demonstrated evolutionary constraints on the length of non-coding regions by phylogenetic shadowing of several Drosophila species. In *EGFR*, indels were particularly frequent around exons 1 and 2 and exhibited a range in size of lesions (1-23 bp), with lesions larger than 4 bp prevalent at above 5% frequency (11/22). A complementary pattern of few, relatively rare and predominantly short lesions, was found in the 2200 bp upstream of exon 3, and in the 3' untranslated and untranscribed region (1/7 larger than 4 bp above 0.05). The only high frequency indels in those regions are 1 or 2 bp lesions and two large deletions in the 1500 bp upstream of exon 3, 37143 (18 bp) and 37539 (27) are extremely rare (present in 1/204 and 2/171 of the alleles respectively). Graphing frequency against indel length for these regions yields almost L shaped curves, which is in sharp contrast to the distribution of indels around exons 1 and 2 (Figure 3.5). There is a parallel lack of divergence of indels to *D. simulans*, with two 1 bp deletions being fixed in intron 3 adjacent to exon 2 and no fixed differences in the 3'UTR. Larger indels are fixed in the vicinity of exon 2 and in intron 3. The largest insert segregating was located 253 bp upstream of the start codon

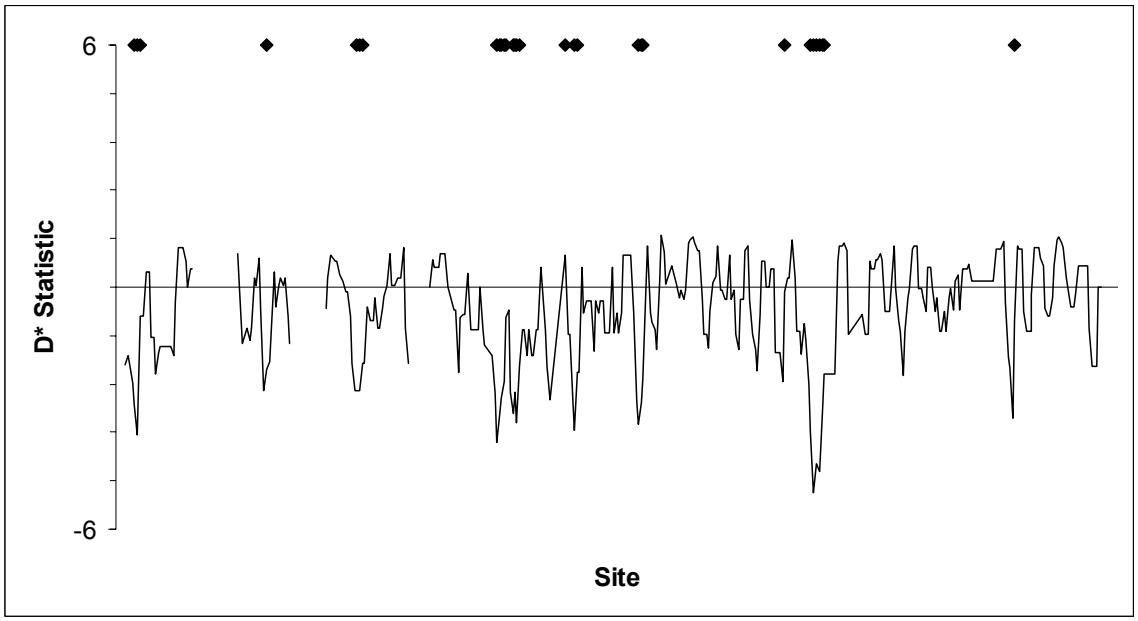


Figure 3.4. Fu and Li's D* statistic along *EGFR* (100 bp window sliding in 25 bp increments). The statistics outside of 95% confidence intervals are noted by black diamonds (above). Those are in all cases accompanying negative D* statistics. Due to heterogeneous in sampling intensity the sample size for estimates ranged between 78 (3'UTR) and 162 (promoter and exon 1). Polymorphisms overlaying indels were omitted.

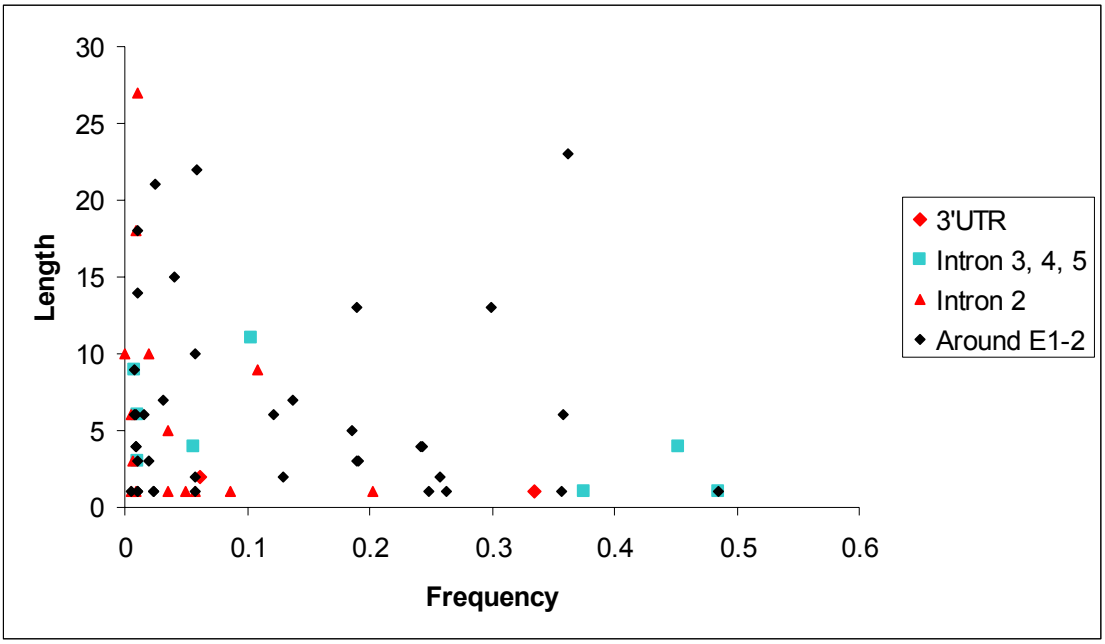


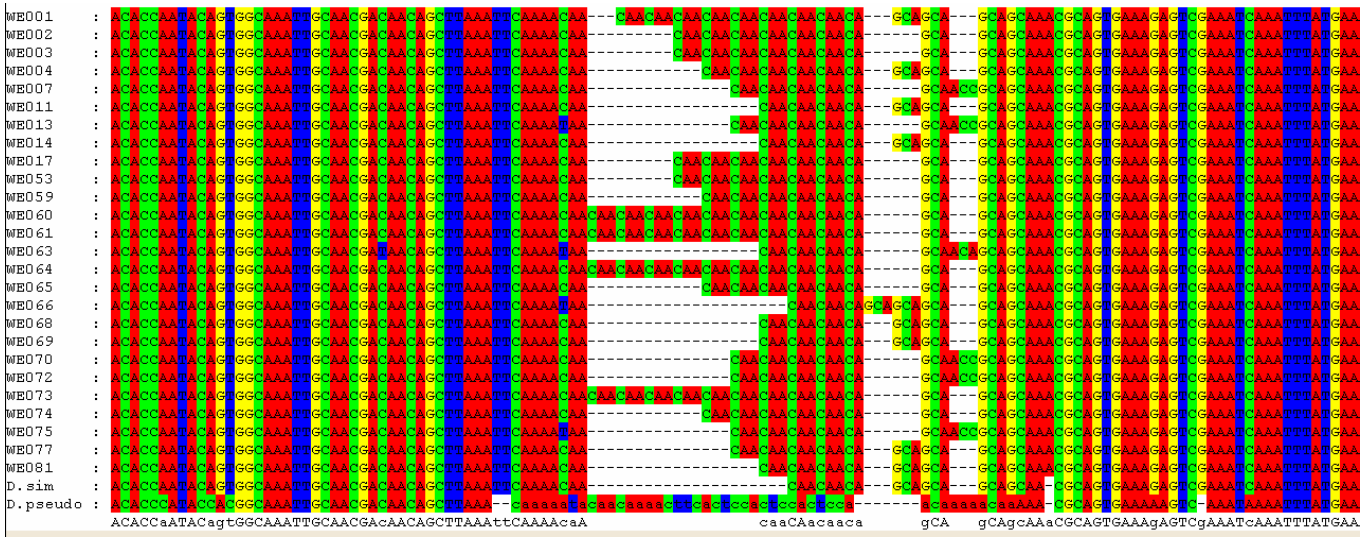
Figure 3.5. Pattern of insertion and deletion polymorphism in *EGFR* graphed by frequency and length of lesion. Indels in regions around exons 1 and 2 are plotted in black (~3.5 kb sequence), while those in the proximity of exons 3-6 are in red (7.3 kb). See Figure 3.1 and Table 3.1 for coordinates.

in exon 1 (site 5763). It is the short (200 bp) variety of the transposable element Pogo. Release 3 of the Drosophila genome contains 39 partial and 5 full Pogo elements (Kaminker *et al.* 2002). This Pogo element was found in 2 lines from the WE collection in 2000 and also in 3 of 350 chromosomes sampled from the same location in 2002 (data not shown). Previous studies are consistent with large transposon insertions interfering with expression of a locus (MacKay and Langley 1990, Dunn and Laurie 1995). Another hint of the role purifying selection in non-coding regions of *EGFR* is a complex microsatellite in the exon 2 promoter (230 bp upstream). The satellite is comprised of 3 kinds of repeats with a CAA-element being the most variable (Figure 3.6). The level of polymorphism in the microsatellite is at odds with ~30 bp neighboring regions which are almost deprived of polymorphisms and are highly conserved in *D. pseudoobscura.* Closer inspection of the length variation reveals a degree of non-randomness, as the total length of the alleles shows two predominant peaks separated by 12 base pairs. As DNA helices turn on average every 10.4 bp (Lewin 1997), this results suggest a functional constraint on the length polymorphism. According to this model, selection would preserve the orientation between the two highly conserved motifs on either side of the microsatellite by restricting its total length. Interestingly, the *D. pseudoobscura* sequence is highly conserved in the surrounding regions but has a different type of repeat. This argues for the biological importance of repeat in this precise sequence context, and also predicts the same length restrictions will apply to the *D. pseudoobscura* repeat.

Promoter of exon 2

The comparison to *D. pseudoobscura* shows that the second exon is better conserved then the first, not least in the promoter region. However, it took a coupled comparison of divergence to *D. simulans* and polymorphism levels in *D. melanogaster* to uncover an interesting conserved element in the promoter for exon 2, located just upstream of the microsatellite discussed above. The polymorphism levels for the region are low while the divergence estimates remain high, but a closer inspection indicates a degree of non-randomness. The region contains two stretches of variable di-nucleotide repeats, with C's alternating every other base, separated by a 27 base spacer. The bases between the repeats are most commonly A or T, with an occasional G (Figure 3.7). The integrity of the alternate C's is conserved, but 8 out of 9 the alternating bases are changes in reference to *D. simulans.* The integrity of the C part of the element is also maintained to an extent in *D. pseudoobscura*, though the linker region is seven bases instead of 27 (identical sites are indicated by stars in Figure 3.7). The stretch downstream of this element is observable on Figure 3.5, on the left of the microsatellite. Only one of three segregating polymorphisms in this region of the promoter, variant C30200T, is at high frequency. Interestingly this site is significantly associated (p = 0.000027 by individual test) with aspects of

A



B

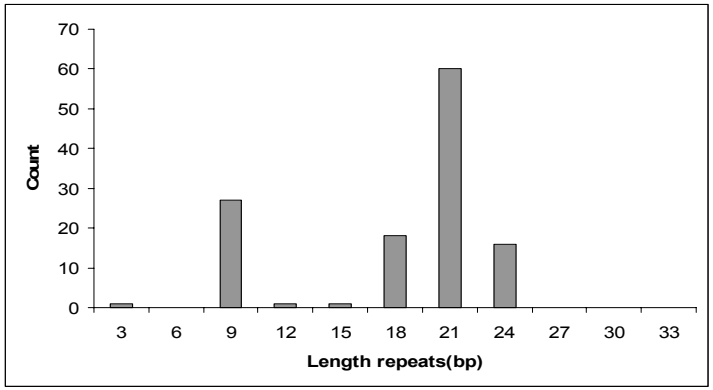


Figure 3.6. Complex micro-satellite in the promoter for exon 2 (alternate transcript RB). A) Alignment of 26 WE alleles and *D. simulans* and *D. pseudoobscura* (below). B) Length distribution of all alleles in *D. melanogaster* as summed over the three types of lesions.

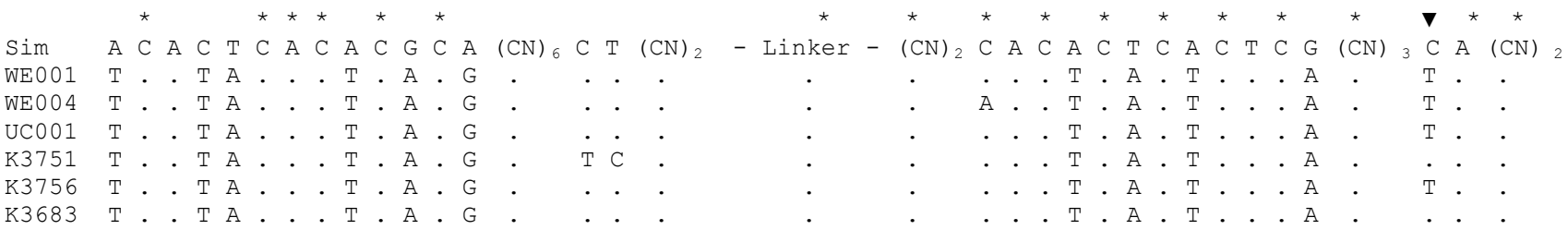


Figure 3.7. The patterns of conservation and polymorphism in the non-coding region upstream of exon 2, from *D. simulans* (Sim) and six representative *D. melanogaster* alleles. WE, UC and K are abbreviations for the population of origin. Stars (*) indicate the sites that are also conserved in *D. pseudoobscura*. CN: Stands for the dinucleotide repeats with the subscript signifying the number of repeats and N being A, G or T. Dots show the invariant sites . The linker is 27 bp in *D. simulans* and *D. melanogaster* but 7 bp in *D. pseudoobscura*.▼ indicates site C30200T that contributes to natural variation in wing shape.

wing shape, as quantified by relative warp parameter C1 which captures distance between crossveins in the wing (Chapter 4).

In combination, these results provide suggestive evidence of a functional role for this element in regulating *EGFR* transcription. Formal tests of the relation between divergence and periodicity in regulatory regions have not been devised, as the focus is on elements with clearer footprints (Dermitzakis et al. 2003). To further investigate the nature of this element I searched the TRANSFAC database (www.gene-regulation.com, registration required) with a representation of the element (C?C?C*C?C?C, where "?" represents a single wild card and "*" a stretch of any character). This database contains information about the transcription factors and sequences they have been shown to associate with in *in vitro* assays. The search returned mainly elements with contiguous C stretches, but a distinct subset preserved the periodicity of the C-element and had all been isolated as binding sites for a GAGA factor. The *Drosophila* targets with characterized GAGA binding sites included *eve*, *ftz* and *Ubx*. Database mining compares moderately with actual experiments but the results suggest that the periodic C element in the exon 2 promoter of *EGFR* constitutes an element recognized by GAGA factors.

Linkage disequilibrium in *EGFR*

The lack of independence between polymorphic sites in a gene or genome can result from close physical proximity, historical events, functional relationships or reduced recombination (as on the *D. melanogaster* 4th chromosome). For the *EGFR*, the squared allele-frequency correlation, r^2, between pairs of sites as a function of distance is summarized in Figure 3.8. The linkage disequilibrium decays rapidly: r^2 values between 0.8 and 1 are only seen between sites in close physical linkage (1 kb), and values above 0.5 within 1.5 kb, with only a handful of exceptions. Out of nearly 12,246 pair-wise comparisons between sites, only 17 had r^2 equal to 1. This number must be considered an approximation since the genotyping was not 100%. The pattern of D' drops distinctly more slowly with distance, with D' of 1 (signifying complete LD) even observed between sites in the first and last exon, separated by 30 kb. Maximum D' was primarily produced by rare variants in complete coupling or repulsion LD to other sites, but the bulk of those are neither significant after correcting for multiple comparisons nor by individual exact tests (below the diagonal in Figure 3.9). Local LD can be observed with the clustering of high values on the diagonal, but very few long range associations are significant. In particular there is no LD between the regions of the locus surveyed in the study, even if they are separated by only 3 kb. The extent of linkage disequilibrium is summarized with means, confidence intervals and median for r^2 and D', and the percentage of significant associations (Table 3.3). LD profiles are very comparable between the two North American populations, but

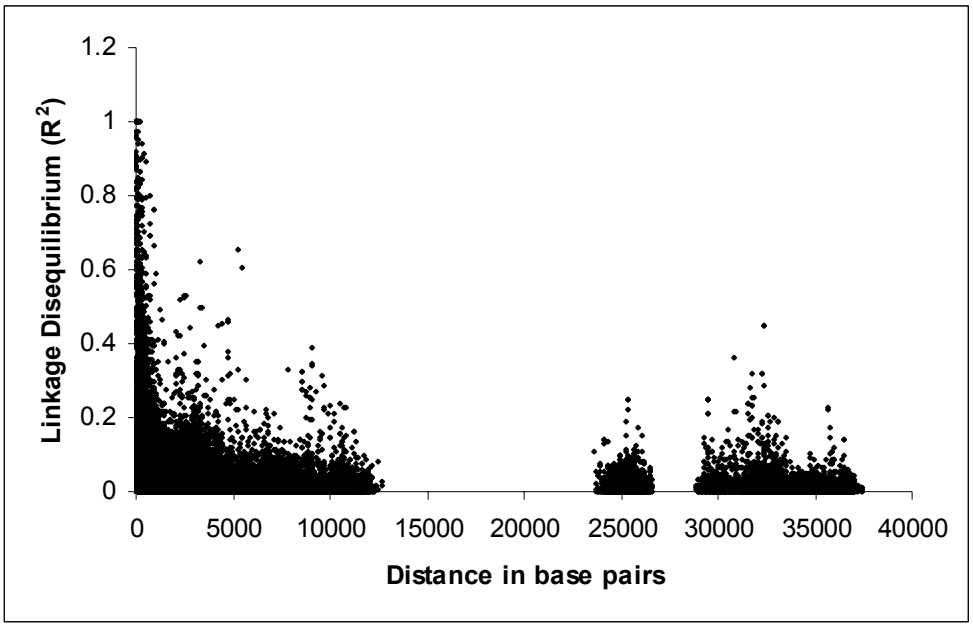


Figure 3.8. Linkage disequilibrium in *EGFR*, described by the relationship between r^2 estimated from pairs of sites and physical distance. Calculated on data pooled from the two North American populations, removing sites with frequency of rare allele less than 15% and N less than 50. The 3 regions sequenced are separated by 3, 23.5 and 27 kb.

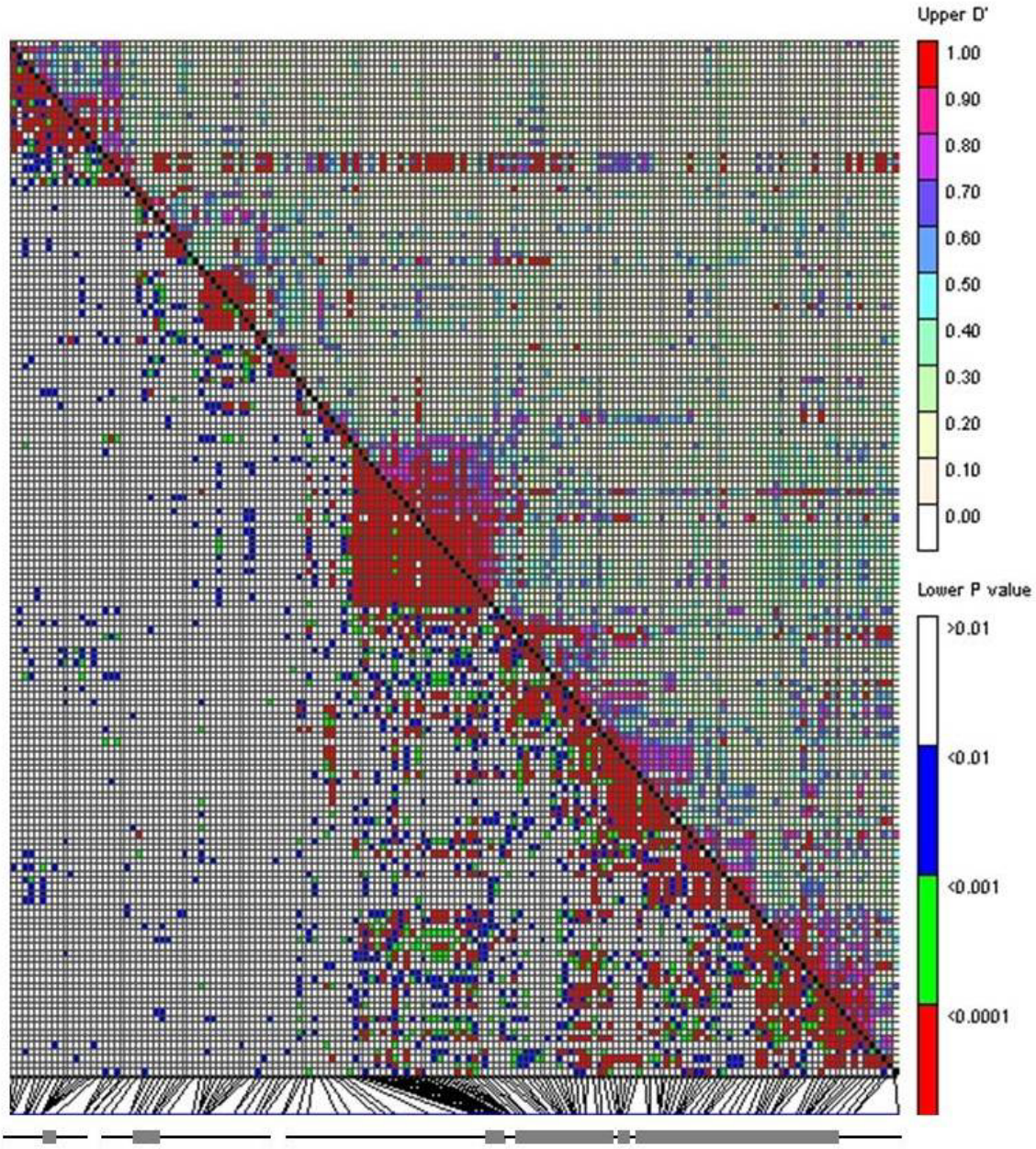


Figure 3.9. The degree of linkage disequilibrium *D'*, is above diagonal and significance of each test below (p-value of exact test vs. permuted data). Strength of the association and significance is indicated by coloring scheme, with red indicating strongest and most significant relationships. Location of each polymorphism is designated on the structure of the gene below. Calculated for the two North American populations, removing sites with frequency of rare allele less than 15% and N less than 50.

Table 3.3. Descriptive statistics LD in *EGFR* from three populations of *D. melanogaster*.

Population[1]	LD metric[2]	Mean		Median	% below 0.05 [(3)]
Kenya	r^2	0.10	±(0.0019)	0.05	
	D'	0.59	±(0.0056)	0.54	
	p-value	0.36	±(0.0037)	0.36	12.2
UCD	r^2	0.05	±(0.0020)	0.02	
	D'	0.39	±(0.0063)	0.28	
	p-value	0.30	±(0.0044)	0.29	18.4
WE	r^2	0.05	±(0.0014)	0.02	
	D'	0.35	±(0.0044)	0.25	
	p-value	0.23	±(0.0031)	0.18	30.5

1. Due to a sample size differences then the Kenya had on average 22 alleles behind each metric, and UCD and WE, 56 and 100 respectively.
2. LD profiles were calculated by population datasets for sites at 10% frequency or higher.
3. Indicates the percentage of tests with probability less than 0.05.

the Kenyan sample seems to exhibit higher values for both r^2 and D'. This could be caused by inversions segregating in the African populations. Even though the descriptive parameters are low, a sizable portion of the pair-wise LD tests are significant at the 0.05 level, 18.4% in UCD and 30.5% in the WE population. This relation could however also be attributable to the large size of my sample, as the significance of tests of LD is affected by sample intensity and allele frequency (Lewontin 1988). The concordance between the North American populations was visualized by plotting r^2 and D' for pairs of sites common to both locations (Figure 3.10). r^2 shows better correspondence between populations then D'. The Pearson correlation of r^2 values in the two populations is 0.86312, while the respective estimate for D' is 0.54393, both of which are significant at the 0.0001 level. This illustrates the overall similarity of the LD profiles in the two populations.

Langley *et al.* (1974) proposed orienting alleles by frequency, thus providing meaning to the signs of *r* and *D*. Positive values indicate coupling, for instance of the rare variant at one site with the rare variant at second site. Negative values represent repulsion phase, namely association of rare with common alleles. For the large contiguous regions of *EGFR*, these estimators are asymmetrically distributed (Figure 3.11). An excess of coupling LD is observed for the site pairs separated by 2.5 kb or more, as the slope of regression of *r* (and *D*) on distance is greater for the coupling than the repulsion phase ($p < 0.0001$).

Estimation of population subdivision

Wright's estimator of population differentiation, F_{ST}, estimates the proportion of within to between population variance in allele frequency. Instead of summarizing the populations with a single metric, I describe the pattern along the *EGFR* locus. F_{ST} values were estimated for a sliding window of 10 polymorphic sites for pairs of populations, and range from 0 to 0.13 for the two North American populations. By contrast, the range of values for the Kenyan population contrasted to WE and UCD separately gave a maximum F_{ST} of 0.28 (Figure 3.12). The significance of the F_{ST} estimates shows the same pattern, with amplitude and width of the peaks contributing to the significance. The signature of population subdivision does not appear to be restricted to one particular part of the gene. Fine scale dissection of the difference between the North American populations, by shrinking the window to 5 polymorphic sites and eventually to individual sites, reveals that the peaks of significant population subdivision along *EGFR* can be attributed to single sites. The only exception was a set of 3 sites (40428, 40458 and 40464) in exon 6. Frequencies of each SNP class are shown in Table 3.4, which indicates for example that the two alleles are in a ratio of 64:43 at site 35697 in the WE population, but 10:27 in the UCD population ($p = 0.0004$).

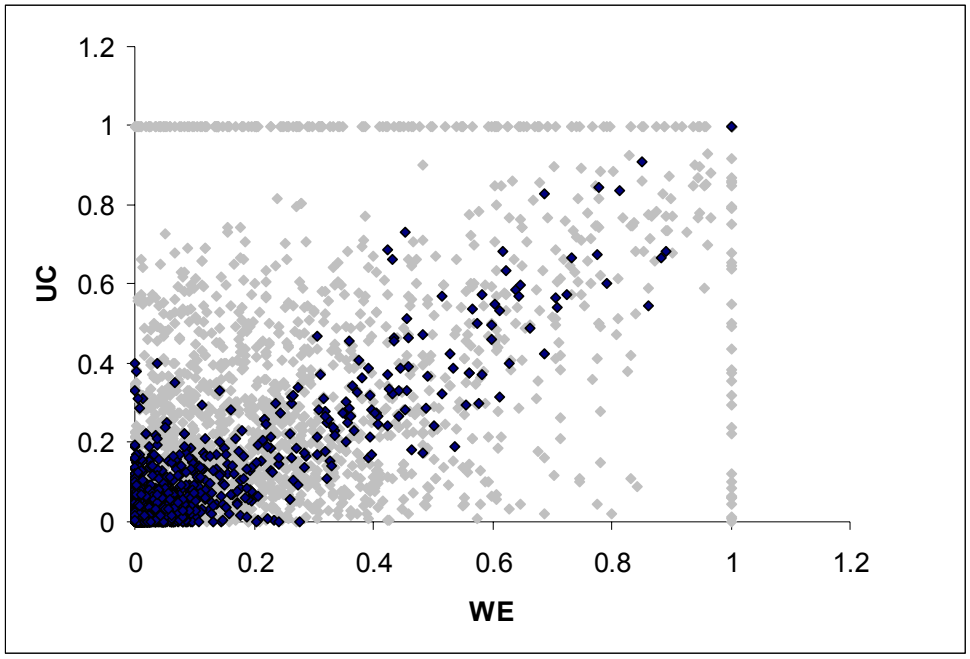


Figure 3.10. Correspondence between LD metrics (D' gray, r^2 black) in the two North American populations. WE: designates the population from West End, North Carolina, and UC the Californian population from UC Davis. The cutoff for sites analyzed where 10% frequency of minor allele and minimum count of 30 (UC) and 50 (WE) alleles in the sample. Only pairs of sites present in both datasets are reported. The Pearson correlation between the r values in the two populations was 0.86312 and the correlation between the D' values was 0.54393. Both are significant at the 0.0001 level.

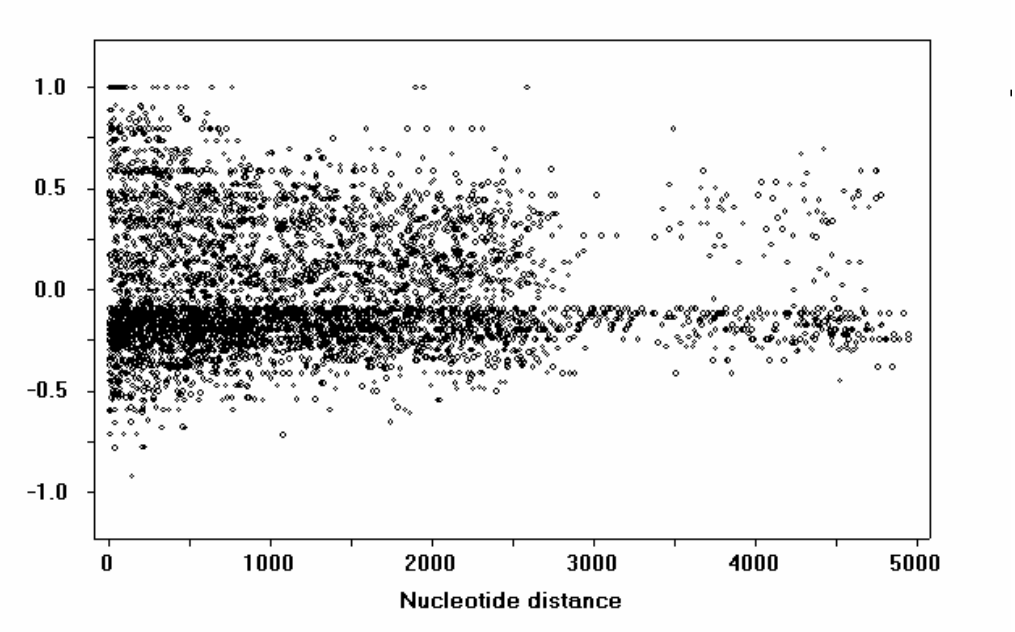


Figure 3.11. The decay of *r* in exons 4, 5 and 6 in *EGFR*, estimated on alleles oriented by frequency. Positive values indicate coupling LD while negative reflects repulsion of the minor alleles. Estimates are derived from 36 Kenyan alleles.

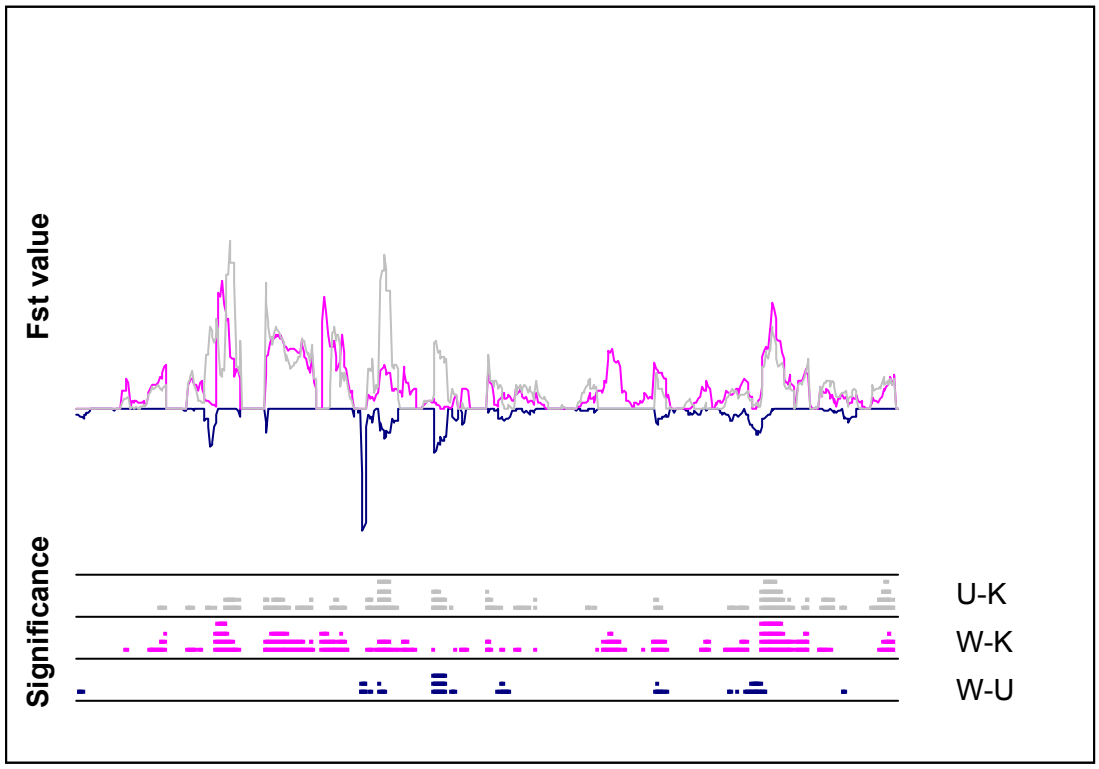


Figure 3.12. Sliding window analysis of F_{ST} along *EGFR*, estimated in pair-wise analysis between the populations (Kenyan (K) and the two North American populations, WE and UCD). The window spans 10 polymorphic sites and shifts 1 bp at a time. A. The estimate of F_{ST} of WE and UCD comparison (W-U in blue) is graphed below the X axis (the true value multiplied by -1) while the two contrasts to Kenyan (K-W in gray and K-U in red) are above. B. Significance of the F_{ST} estimates for the respective contrast along the locus, are indicated by the horizontal stack of stars, where *<0.05, ** < 0.01, * < 0.001 and **** < 0.0001.**

Exclusion of the rare polymorphisms (<0.05) yielded comparable F_{ST} and p-values for the focal sites and regions (results not shown). Experiment wide significance of F_{ST} values was assessed by correcting for the number of tests by Bonferroni and related methods. Focusing on the contrast between WE and UCD, none of the 650 F_{ST} values are significant at the p=0.05 level after correcting for multiple comparisons. A less conservative approach is to consider only the frequent sites, which reduces the number of tests to 201. Then only one site (36214) survives Bonferroni correction (cutoff p = 0.00025). However comparisons to the Kenyan sample yielded more significant F_{ST} estimates that also had more extensive distributions along the gene (Figure 3.12). Excluding rare variants from the analysis, regions represented by four "stars" in part B of Figure 3.12 are significant considering all 3 contrasts. The overall magnitude of F_{ST} was not uniform among the 3 pairs, with the North American contrast having lower estimates (average F_{ST} 0.0098 compared to 0.045 in the Kenya sample, significant by ANOVA; p < 0.0001). The population samples also differed by the total number of unique sites, also known as "private alleles" (Slatkin 1985). The Kenyan sample has 92 unique sites, almost three times as many as the UCD sample, despite the latter having two and a half times the sample size. Interestingly, the North American samples shared 50 sites, 14 of which are at or above a frequency of 0.05, suggesting considerable evolution in the lineage after isolation from Africa but before dispersal across North America. The Kenyans shared 14 sites with the UCD's, and only 4 with the WE's, reminding us of the potential variance in these estimates.

The distribution of F_{ST} values along a chromosome region or a gene can correlate with the intensity of natural selection. This could reflect weak population-specific positive selection in one or both of the populations. In the *EGFR* locus, tracking of F_{ST} for the three contrasts is not uniform along the locus, as there is an apparent deficit of evidence for population structure around exons 1, 3, 4 and 5 and in the 3' UTR. However, the significant F_{ST} values do not localize exclusively to any distinct regions of the locus.

Relationship of F_{ST} to LD

A prediction of a simple selection model is that a signature of population division should be generated by hitchhiking associated with the increase in frequency of a single haplotype. As a result, the frequency of several sites on this haplotype would be simultaneously altered, leading to LD between them. To address this, I plotted the decay of r^2 with distance from the sites that show the highest values of F_{ST} (Figure 3.13). Only the trio of sites in exon 6 (40428-40464) have r^2 values higher than 0.5 and pair-wise r^2 values greater than 0.90. Linkage Disequilibrium drops to zero very quickly on both sides of these and other focal sites. The LD decay from focal sites was highly correlated between the North American datasets and did not differ significantly in the Kenyan set (data not shown).

Table 3.4. Allele frequencies and F_{ST} parameters for a comparison between the West End and UC Davis populations.

		WE[1]	UCD[1]	Kenya[1]	WE to UC contrast[2]			Haplotypes[3] (WE)		Haplotypes[3] (UCD)	
Site	Variant	N	N	N	Window	F_{ST}	*p*-value	N	Diversity	N	Diversity
35345	A	73	72	na	1	0.10093	0.00129	2	0.29491	2	0.05259
	C	16	2	na	10	0.28514	0.00257	4	0.65702	4	0.58333
35697	T	64	10	25	1	0.17558	0.0004	2	0.48074	2	0.39445
	C	43	27	3	10	0.07428	0.01832	6	0.57543	3	0.49174
36214	G	93	43	30	1	0.13641	0.0003	2	0.28173	2	0.48685
	A	19	31	3	10	0.0765	0.0001	10	0.49761	5	0.58895
39010	C	79	67	30	1	0.08844	0.00089	2	0.46515	2	0.27219
	T	46	13	5	10	0.03822	0.00772	8	0.78178	12	0.81295
40428[4]	G	73	31	20	1	0.07552	0.00317	2	0.48065	2	0.47469
	A	49	49	16	10	0.06561	0.00257	6	0.58748	5	0.53094
40464[4]	T	54	51	20	1	0.0688	0.00406	2	0.49168	2	0.46219
	C	70	29	16	10	0.06561	0.00257	6	0.58748	5	0.53094
42023	A	90	36	23	1	0.0958	0.00158	2	0.41748	2	0.49861
	G	38	40	12	10	0.02461	0.01812	10	0.811	12	0.76247

1. Absolute counts of the SNP states at the 7 sites, listing the frequencies of in the Kenyan sample as a reference (na indicates site not surveyed in the Kenyan panel).
2. F_{ST} between WE and UCD, for individual sites and haplotypes spanning 10 segregating sites with corresponding *p*-value for each estimate.
3. Diversity and number (N) of haplotypes (alleles) per population are summarized.
4. Sites 40428 and 40464 are in nearly complete LD (D'=1, r^2=0.9, *p*-value<0.0001 by Fishers exact test). They are separated by 38 bp and contribute to the same 10 site haplotype.

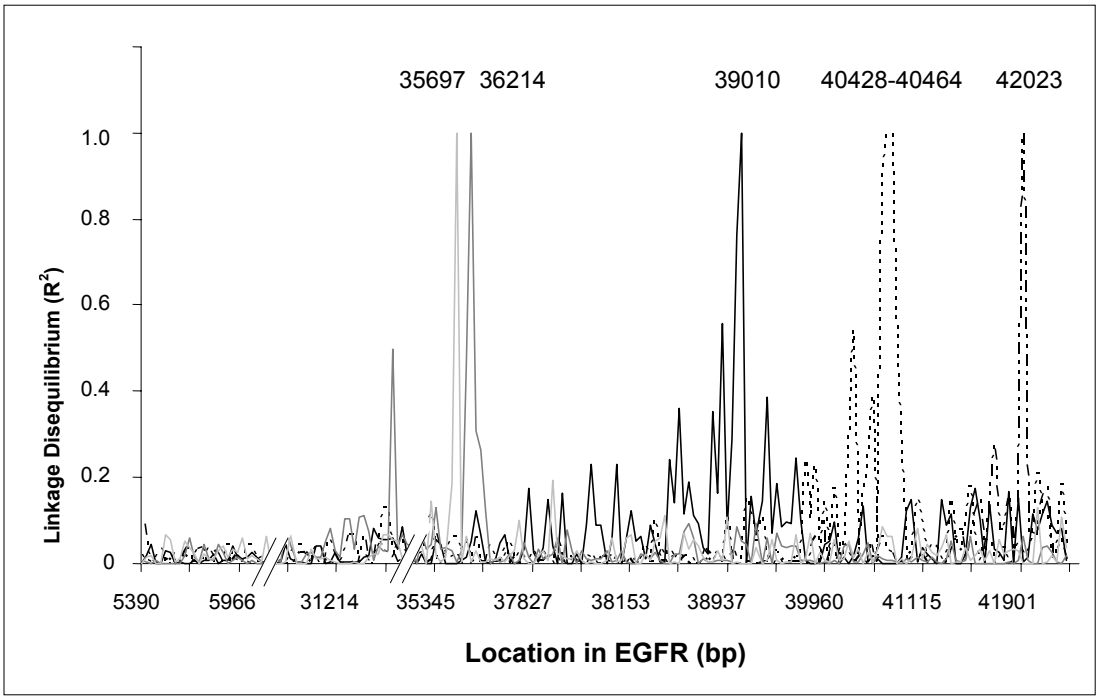


Figure 3.13. Linkage disequilibrium (r^2) in relation to 5 of the sites with significant F_{ST} differences between populations WE and UCD. Decay of LD from each of the sites graphed with distinctly shaded or patterned line. The X axis is adjusted for representing the sequenced regions. From left to right; exon 1, exon 2 and exon 3 through 6. LD estimated for the WE population alone, on all sites above 5% frequency and present in 50 or more alleles.

Discussion

Molecular evolution and population genetics of *EGFR*

Several pieces of evidence are consistent with purifying selection shaping evolution of the *D. melanogaster EGFR*. (i) At the protein level, only 1 replacement at moderate frequency is found in the main part of the protein (1400 amino acids). The 4 other common replacements are located in a presumed signal-peptide encoded by exon 1 and the 17 remaining amino-acid changes are all rare. (ii) The majority of the estimates of Tajima's D were negative, including estimates in several regions that approaching significance by reference to coalescence simulation. (iii) Sliding window analysis of Fu and Li's statistics showed only negative values significant at 0.05 level, and there is an excess of those considering the number of tests (30 vs. the expected 18). (iv) Two non-coding regions, 2.2 kb upstream of exon 3 and the 3' untranslated and untranscribed region had reduced levels of high frequency SNP's and no large indels. (v) The only microsatellite in the sequenced region does not appear to be evolving randomly, with two distinct length classes observed that could be maintained by functional constraint. Suggestion of negative selection rests on these observations but does not have formal statistical support as will be discussed.

EGFR protein evolution

Two features of EGFR protein evolution are noteworthy, the low level of protein polymorphism, and the rapid evolution of the putative signal peptide encoded by alternate 5'-exon 1. In the protein proper, one high frequency and 14 rare variants are found, while the corresponding numbers for exon 1 are four and three. Those findings are in concordance with Fay and colleges (2001) claim that around 80% of replacement changes may be deleterious. Divergence to *D. simulans* tells the same story, with six of the 10 fixed replacements residing in the same signal-peptide. These results are consistent with EGFR protein experiencing similar negative selection as observed for other key components of the Ras/MAPK pathway (Gasperini *et al.* 1999, Riley *et al.* 2003). While the MacDonald-Kreitman test was designed to test for positive selection on proteins it has some, though notably weaker, power to detect negative selection. It is therefore not entirely surprising that the MK test was not significant for the whole EGFR protein. However application of the test by exons gave significant results for exon 1, caused by rapid protein evolution. These results are however only suggestive as the test statistic is not significant if I correct for multiple tests. The Neutrality Index of (Rand and Kann 1996) documents this pattern but does not offer an explicit test. The index for the whole gene reports higher level of polymorphism within species, while excess of divergence was detected for exon 1. The nature of the NI descriptor (ratio of ratios) may leave it sensitive to chance and thus diminish its utility. The fact that the exon is hardly distinguishable in the *D. pseudoobscura*

genome, and can not be found by similarity in *A. gambiae* supports the hypothesis that exon 1 is evolving more rapidly then the rest of EGFR. But it is not likely that the exon has been undergoing positive selection throughout this evolutionary time scale. Relaxation of purifying selection, accompanied by spurts of positive selection is a more likely scenario.

The fact that all of the changes that are either segregating within *D. melanogaster* or fixed with respect to *D. simulans* are in the putative signal peptide is puzzling. The other alternate 5'-exon, encoded by exon 2, also has a putative signal peptide but there are no fixed differences and only three low frequency replacements segregating. However, the HKA test did not document significant differences in the rate of evolution of these exons. This could be caused by the shortness of the regions, as exon 1 is 153 bp and exon 2 is 300 bp. Acknowledging that the results are only suggestive of exon 1 evolving more rapidly, it is still worth asking if functional differences between the two alternate transcripts have been documented? We only have mRNA expression data available and those will not present the full picture as translation, membrane trafficking and possibly ligand recognition may be affected by the alternate N-termini. Both transcripts show largely overlapping spatial and temporal expression during embryonic and pupal development, with the notable exception being adult specific expression of the shorter transcript (encoded by exon 1), which is restricted to ganglia and the cortex (Lev *et al.* 1985, Scheiter *et al.* 1986, Kammenmayer and Wadsworth 1987). *EGFR* contributes to axon guidance during the wiring of the nervous system and Ras signaling has been implicated in several behaviors. Behavior is normally mediated on a short time scale, through neuronal signaling, and may therefore not be as sensitive to variation in transcriptional or translational attributes as developmental processes, with the possible exception of memory-consolidation, which requires gene expression. While the transcripts appear redundantly expressed during development, the patterns of replacement polymorphism in the two alternate N-termini pose the hypothesis that transcript RB (including exon 2) will have broader and more functionally constrained expression. This could be explicitly tested. Characterization of 26 *EGFR* alleles by Clifford and Schüpbach (1994) and Lesohkin *et al.* (1999) showed that the whole protein is effectively a target for functional mutations, with the exception of the alternate 5'-exons. The mutations were predominant in domains implicated in ligand binding and tyrosine kinase activity, and result in a range of phenotypic effects, from larval death to specific aberrations of wing morphology. Interestingly Lesokhin *et al.* (1999) identified 3 of the replacements in exon 1 reported here in an effort to determine the molecular lesions of the *Ellipse* gain-of-function alleles. They tested the effects of two of those (W15R and L21W) in transgenic animals and by monitoring the phosphorylation state of downstream targets (ERK). Qualitative results indicated that those sites are functionally neutral. Absence of effects of polymorphisms in a signal-peptide are not surprising, but must be considered in light of the disparity between the two alternate exons. A closer look at the dynamics of expression of the

EGFR splice variants might shed light on this. Also, the presence of multiple Methionine codons at the start of exon 1, while only one is found in exon 2 is also curious in the context of the hypothesis that exon 1 is undergoing rapid evolution.

In a broader context, two theories relating to selection on protein modules state that purifying selection will either be strongest on the most upstream (Olsen *et al.* 2002) or the most highly connected components of the pathway (Riley *et al.* 2003). *EGFR* is not the best candidate for a discriminative test between the contrasting hypothesis since it is both high in the regulatory hierarchy and holds a central place in the Ras/MAPK cascade. The centrality of EGFR is not necessarily expected *a priori* as seven other RTK's have been described in flies (Sevenless, Breathless, Heartless (FGF-receptor), JAK, Torso, and two Insulin-receptors, see Flybase for updated information). However in a very thorough review Held (2002) argues on the basis of genetic evidence and the extensive and dynamic expression of *EGFR* during development that the locus encodes the main receptor for the Ras/MAPK signal cascade in flies. The observed deficit of protein polymorphism is consistent with both hypotheses. Fraser *et al.* (2002) investigated the rate of evolution in yeast genes in relation to the number of protein to protein interactions each peptide participates in (as assayed by yeast two-hybrid screens). They observed that increased connectivity of a protein restrained its rate of evolution, providing support for the topological hypothesis.

Non-coding regions of EGFR

Analysis of the evolution of the non-coding regions in *EGFR* was conducted on two classes of data, SNP's and indels. The data are consistent with two larger regions, in particular parts of intron 2 and the 3' untranslated and untranscribed region, along with two peculiar motifs in promoter 2, experiencing purifying selection. Tajima's D is predominantly negative for these regions, but none are formally significant after correcting for multiple tests. Qualitative comparison of polymorphism levels as estimated by π and divergence to *D. simulans* and *D. pseudoobscura* is however suggestive of evolutionary conservation of these regions. The lack of significance of positive Tajima's D, Fu and Li statistics and rejection of the Fay and Wu tests argues against the role of strong positive selection in *EGFR*. The results are more consistent with purifying selection, as best indicated by sliding window analysis of D* and F* (Figure 3.4). While no *p*-value is lower than Bonferroni cutoff, there is an excess of regions with significant values (30 observed vs. 18 expected). The significance estimates were derived from sampling variance of the statistics and not by coalescence, so we can not exclude stochastic processes as a reason for the pattern. However the fact that all significant windows have negative values is also consistent with purifying selection removing high frequency variants from parts of *EGFR*. Those regions include intron 2 and short segments in the two promoters that are well conserved. The lowest D* values were in a segment of exon 6 and extended into intron 5, which

could explain the deficit of polymorphisms observed in introns 4 and 5 compared to intron 3 and the level of divergence (Figure 3.3).

The disparity in parameters suggests considerable variation in the forces that different parts of the locus experience, especially as monitored by D* (Figure 3.4). Weak and purifying selection are hard to identify, and a mixture of the two even harder. It is possible that our toolkit of molecular evolutionary analysis (Wayne and Simonsen 1998) is still underdeveloped. Also, the stochastic nature of the evolutionary process may reduce the power of the tests. For instance it was shown recently that increased LD and high frequency derived polymorphisms (Fay and Wu 2000), which are considered signatures of selective sweeps, decay rapidly following fixation (Przeworski 2002). Most of the molecular tests of neutrality are only useful for identifying strong signals, which do not appear to be the case for *EGFR*.

The second class of data, insertion deletion polymorphisms and divergence, also suggest disparity in the evolution of non-coding regions. Official tests for deviation from neutral evolution for indels have yet to be devised, (see Schaeffer 2003 for a first pass) and therefore we can only discuss notable features of data. The indels in *EGFR* enable an inspection of the patterns of polymorphism around a major developmental locus. As Figure 3.5 shows, the 2.2 kb upstream of exon 3 and the 3' UTR have no large indels above 0.1 in frequency. This pattern of exclusion argues for the role of purifying selection on non-coding regions of *EGFR*. The deficit in the 3' untranslated and untranscribed regions can not be attributed exclusively to selection on *EGFR* as only 0.8 kb separate it from the adjacent locus, CG10440, of unknown function. It remains to be seen if these patterns are sufficient to build statistics to describe the evolution of indels. The 60 indel polymorphisms detected in ~11 kb of sequence suggests that such metrics may require large datasets and only be meaningful for larger regions. Statistics utilizing indels will be data demanding as only 1 in 10 polymorphisms are indels. The fact that divergence to *D. simulans* shows the same pattern argues for the utility of phylogenetic shadowing to determine restrictions on indel evolution. Bergman *et al.* (2003) found genome wide constraints on length polymorphisms in conserved non-coding regions in *Drosophila*, lending credibility to suggestions about their functional importance. Bergman and Kreitman (2001) found indications of the same pattern, with no differences between intronic and intergenic regions. Molecular dissection of regulatory elements has also stressed the importance of spacing between sequence motifs that regulate transcription (Small *et al.* 1992, Ondek *et al.* 1988).

One interesting feature of the current dataset is a complex microsatellite in promoter 2. It is comprised of three different repeats and there are restrictions on the length of the whole element (Figure 3.6 B). Two allele classes predominate, differing by the length of a single turn on the double helix. This could not be a historic artifact as the microsatellite is comprised of 3 kinds of lesions, and the pattern only becomes apparent when the length is calculated for the whole element. The most plausible explanation for this length distribution is negative selection,

probably acting on the relative spacing of highly conserved flanking regulatory motifs. Downstream to the microsatellite is a second element which may be a putative GAGA factor binding site, discussed in depth below.

Conservation of a putative GAGA binding motif

It is generally acknowledged that selection on gene expression will preserve the integrity of regulatory regions down to composite enhancers and even individual binding sites (Ludwig 2002). In case of *even-skipped* the preservation seems to be on individual modules, as turnover of binding sites occurs over evolutionary time (Ludwig *et al.* 1998). The individual binding sites, ranging in size from 7 to 30 bp remain relatively intact. The patterns we observe in the alternating C di-repeat in the promoter of exon 2 reflect a different pattern of conservation that may be attributable to the biochemistry of its function. Database mining suggested that the element was a target for GAGA-factors (GAF) as it shares the alternating C feature with characterized binding motifs for GAF's in other Drosophila genes (*eve, ftz, Ubx*). The GAF's are highly abundant nuclear proteins which act mainly as repressor or anti-repressors by disrupting nucleosomes. There is also limited evidence suggesting that the protein tracks along with RNA polymerase II during transcription, but its main function is to operate on the histone complexes (Kerrigan *et al.* 1991, O'Brien *et al.*1995, Tsukiyama *et al.* 1994, Wilkins and Lis 1998). Drosophila contains two GAF's encoded by *Trithorax-like* and *pipsqueak*, which seem to act in a concerted manner (Schwendemann and Lehmann 2002). There is also recent evidence suggesting that GAFs may act *in trans* by linking two DNA molecules (Mahmoudi *et al.* 2002), providing a functional basis for the phenomenon of transvection (see review by Duncan 2002). It is interesting that one of the best described examples of transvection in Drosophila is *Ubx* which was shown to contain GAF binding sites. Early *in vitro* assays showed that GAF's required only GA repeats for binding, hence the name. Later a consensus sequence CT $(GA)_n$ was proposed with the average repeat length of 3.5, and considerable ambiguity allowed in the non G-part of the element. In addition, the orientation of the element does not matter for function, consistent with the alternating C pattern in the exon 2 promoter. Finally Hodgson *et al.* (2001) characterized a 70 bp GAF module in the *bithoraxoid* region of *Ubx* that has a similar organization as the element noted in promoter 2 of *EGFR*, with two GA tracts separated by a linker. Hodgson *et al.* (2001) demonstrated with site directed mutations, *in vitro* binding assays and transgenics the importance of the GA tracts for binding, but also the composite nature of the domain. These molecular details of GAF's, their detailed binding domains and location in promoters are consistent with the alternating C-repeat in exon 2 being a GAGA factor binding module.

LD and fine-mapping in flies

Mapping of complex traits relies on a high degree of dependence between a marker and contributing variants. The degree of independence of polymorphisms is assessed with the squared correlation coefficient of allele frequencies (r^2) as the *D'* statistic seems to have higher emperical sampling variance (Pritchard and Przeworski 2001). Here, r^2 was shown to be the more replicable parameter in comparisons between the two North American populations (Figure 3.10). A delineation of the relative power of these metrics needs to be assessed in light of larger genomic datasets. Linkage disequilibrium in *EGFR* decays over a kb, consistent with several *D. melanogaster* genes in regions of high recombination (Powell 1996). Significant associations were observed almost exclusively between sites in close (<0.5 kb) physical linkage. There are local blocks of LD for instance around intron 3, a short region displaying a high level of polymorphism. r^2 also captures functional independence of segregating sites or domains within a locus or the genome. Absence of LD increases the discriminating power of purifying selection and also decreases the coupling between sites that are associated with phenotypic variation. Betancourt and Presgraves (2002) demonstrated this relation between selection and linkage in Drosophila, both for adaptive evolution and codon bias, the weakest purifying selection. Consequently, for most complex quantitative traits it is essential to characterize the distribution of linkage disequilibrium in a candidate region before attempting fine-mapping. Establishment of the human haplotype map is an effort to gather genome-wide data on LD that will allow cost efficient genome screens and tailored efforts to dissect complex loci in specific regions. There is no evidence for long range haplotype structure in *D. melanogaster* and our results suggest a rapid decline in LD even within genes. Long range LD has been document for allozymes in *D. subobscura* that may reflect true functional coupling (Zapata *et al.* 2000). However, we also have evidence for short-range LD depicted in the high significance on the diagonal in Figure 3.9. Several regions of EGFR show two distinct haplotype clades over stretches of a half kb (data not shown) similar to observations by Teeter *et al.* (2000). Those could be signatures of ancient admixture, functional coupling of variable sites or part of the stochastic fluctuations in allele frequency and linkage over time. More precise quantification of these patterns and the random expectation, or survey of a second species will hopefully elucidate the conundrum. Success of fine-mapping of complex phenotypes in flies must therefore depend on very extensive genotyping or a strong functional effect of the locus.

The positive corollary is that lack of dependence will increase the accuracy of quantitative trait mapping. If there is sufficient evidence to suspect the involvement of a locus in a given trait, then an association-test based dissection of the locus has the potential to identify the quantitative trait nucleotides that contribute to the variation. This is naturally contingent on the, by no means trivial issue, of sufficiently extensive genotyping. It must be stressed that the

success of fine-mapping of quantitative trait loci in *D. melanogaster* has been considerable (MacKay and Langley 1990, Laurie *et al.* 1991, Lai *et al.* 1994, Long *et al.* 1998, Long *et al.* 2000, Robin *et al.* 2002). The paradigm has been careful gathering of independent evidence for the involvement of a locus in a particular trait, before attempting fine-dissection. Proliferation of genomic tools might make genome wide association tests in flies an at least theoretical option (Kwok 2001, Pritchard and Przeworski 2001). The definite advantages of association tests in contrast to other methods for mapping are the resolution and speed, as it depends not on recombination events in laboratory subjects to generate informative genotypes (Buckler and Thornsberry 2002). However, the results presented here suggest that if most Drosophila genes have similar patterns of LD and we depend on LD between marker and QTN for mapping then such efforts will be biased towards alleles of larger effect. They might also explain why Lai *et al.* (1994) and Lyman *et al.* (1999) implicated the same region of *scabrous* as a bristle QTL but by associations with two different markers. The two markers are in strong linkage disequilibrium and if the LD pattern in *scabrous* is similar to *EGFR* then the causative quantitative trait nucleotide is likely to be closely linked if not one of those two. These results carry a mixed message as we can anticipate higher resolution in fine-scale analysis of candidate loci, but we will have to accommodate the multiple testing problems. How these two issues relate to genome wide mapping in populations poses a challenge for modern quantitative genetics.

Population subdivision and independence of differentiated sites

Theory of *D. melanogaster* population history suggests an out of Africa model (David and Capy 1988) and are corroborated by population genetic analysis (Begun and Aquadro 1993, Schlötterer and Harr 2002, Carachristi and Schlötterer 2003). Our results are in accordance, with high F_{ST} (Wright 1969) between African and North American samples. In addition to allele frequency differences the Kenyan population had excessive number of private alleles. This has relevance for ascertainment of SNP's for characterization of population structure as discussed by Schlötterer and Harr (2002). Our results do not suggest major distinction between the North American samples for *EGFR*. A genome wide study of microsatellites from several localities in Africa, Europe and North America (including 30 alleles from North Carolina) demonstrated the predicted ancestral status of African populations (Carachristi and Schlötterer 2003) and surprisingly high diversity in North America compared to European populations. It also suggests this diversity is caused by admixture of European and African alleles exclusively in East coast populations.

Significant F_{ST} can result from population structure, where lack of migration enhances the effects of genetic drift, non-random mating and selection, or panmixia which can stratify the genome. Both are expected to impact allele frequencies genome-wide while selection has the

power to differentiate between loci and generate specific allelic deviations. The seven sites with largest F_{ST} between the North American populations are not reflecting a common selection event, as they are in Linkage equilibrium (Figure 3.13) and the LD profiles of focal sites do not differ as predicted under a selection model (Przeworski 2002). Also high F_{ST} values are not restricted to one particular part of the locus as anticipate if selection is favoring a single polymorphism. More complex scenarios of selection are possible but can not be distinguished from alternate hypothesis, including the one of chance.

Population differentiation has ramifications for fine-mapping of complex traits (Pritchard and Rosenberg 1999) as it can affect frequencies of alleles as well as patterns of LD (Goldstein 2001). Moreover sites differing in frequency can create spurious associations if phenotypes deviate significantly between populations. Studies in maize have uncovered population structure (Remington *et al.* 2001b) but relevant corrections (Pritchard and Rosenberg 1999) allow mapping to proceed. Similar progress is being made in studies of human disorders (Ardlie *et al.* 2002). In case of fine mapping within *EGFR*, deviations in allele frequencies do not affect LD or alter the mapping potential in the two populations. Finally the utility of flies to mimic human disease gene hunts is increased by similar population histories of the two species; both originated in Africa and are now cosmopolitan species with large population size and show geographic redistribution in recent evolutionary time (Aquadro *et al.* 2001). The varying degrees of population structure in *D. melanogaster* implicated by Carachristi and Schlötterer (2003) are particularly interesting and should be utilized explicitly. For instance in a comprehensive association study where genomic markers are scored along with candidate locus in populations with varying levels of population structure, for instance from Africa, Europe and East coast of the USA.

Conclusions

Here I described analysis of genotypic differences in 10.9 kb of the *EGFR* locus in *D. melanogaster* derived from three populations. Formal deviations from neutrality are not established, after correcting for multiple tests. But several observations, particularly the distribution of D* and F* along the locus, are consistent with the role of negative selection molding both protein and non-coding regions. The low replacement polymorphism level in the protein proper places EGFR firmly among the more conserved components of the Ras/MAPK cascade (Gasperini and Gibson 1999, Riley *et al.* 2003) and could both be caused by the high level of pleiotropy exhibited by the locus and its key location in the pathway. Peculiarly, alternate exon 1 has high rate of protein evolution, documented by replacement polymorphisms segregating and high divergence. This could be result of lack of functional constraint or positive selection. The lack of fixed synonymous changes compared to the *D. simulans* group could indicate the latter. We took advantage of the sequenced *D. pseudoobscura* genome to connect divergence in non-coding regions and polymorphism data in non-coding regions, and hypothesize a role for GAGA factors in *EGFR* regulation. There is however obvious room for further development of tools to quantify and describe divergence and conservation in non-coding regions, particularly in respect to distribution and length of indels, consensus or common features of elements and large scale organization of those motifs. This, along with broader geographic sampling and wider genomic surveys, would place our analysis of the molecular evolution and phenotypic effects of polymorphisms into both wider evolutionary context and enhance our chances of elucidating the functional effects of genetic variation.

Chapter 4

Test for associations between *EGFR* polymorphisms and wing shape in *Drosophila melanogaster*

Abstract

Tests of association normally rely on linkage disequilibrium between marker and causative polymorphism and therefore rarely test effects of quantitative trait nucleotides directly. Our analysis of the genetics of wing shape in *Drosophila melanogaster* led us to test for association between shape and all polymorphisms above 5% frequency in 10.9 kb of *EGFR*. The 267 common polymorphisms were tested against 18 shape parameters, and one size measure in two panels of inbred lines from North America. The association tests identified a non-coding variant (T31365C) with sex dependent effects on wing size, significant after Bonferroni corrections. The effect was found predominantly in the Californian population, but did not replicate either in recrossed North American nor in an independent sample of African lines. The most significant associations with shape affected placement of crossveins. Site C30200T disrupts a putative GAGA factor in the promoter for exon 2, and while not formally significant after multiple comparison correction in the inbred lines, was significant in both follow up experiments. Six other non-coding sites suggest marginal associations with shape, of which only C30505A was replicated, but the effects of this substitution were reversed between North American and African samples. Despite documenting stronger associations than most published studies in Drosophila, these results are only mildly consistent with the hypothesis that polymorphic sites in *EGFR* contribute to standing variation for wing shape.

Introduction

Morphological diversity in organisms past and present documents the evolutionary process and is an indicator of the relative contribution of natural selection and other forces molding the allelic pools of populations. Moreover compound morphologies can be used as models for complex diseases, both in terms of testing techniques to capture common axis of variation and in particular for the practice of mapping heritable components to individual nucleotides. Shape of the *D. melanogaster* wing is an example of a complex morphological structure that can be investigated in an unbiased manner with the new tools of morphometrics (Bookstein 1991, 1996). Moreover, studies on fruitflies exemplify how to dissect the heritable basis of natural variation in continuous phenotypes (Mackay 1995, Lai *et al.* 1998, Clark and Wang 1997), and have led to the identification of polymorphisms in particular genes associated with phenotypes (Laurie *et al.* 1991, Lai *et al.* 1994, Long *et al.* 2000, Robin *et al.* 2002, De Luca *et al.* 2003). This chapter extends on previous association studies in *D. melanogaster* by increasing sampling along three axes.

The suitability of *Drosophila* wings as a system to elucidate the inheritance of composite traits was reasoned in Chapter 1. The morphology has to be summarized with multiple parameters instead of one or few as for more conventional traits. Genotyping was achieved by sequencing the coding and flanking regions for large portions of a candidate locus, enabling a direct test of the effect of each site in those regions. These tests of association can be considered direct estimates of the effects of individual polymorphisms as there is no dependence on linkage disequilibrium between marker and QTN, except for sites bordering on the sampled regions. De Luca *et al.* (2003) applied sequencing to survey variation of a ~600 bp promoter of *Ddc* in a survey of sites affecting longevity. Here the sequenced region is approximately 17 X longer. Finally the inbred lines are derived from two distinct North American populations permitting a contrast of the patterns of associations between them, and a test of population specific effects of segregating polymorphisms. Previous efforts by MacKay and Langley (1990) also sampled two populations but the current enterprise has five times the sample size and therefore more power to detect SNP effects and their reliance on population and sex.

Wing development and EGFR

The initial step of wing blade development is the division of a cellular field into dorsal and ventral surfaces simultaneous to the establishment of the anterior-posterior axis. The latter axis serves as an organizing frontier for secretion of endocrine proteins like Hedgehog, Dpp

and Vein that lead to the formation of veins. The refinement and early differentiation of vein tissue is mediated by the *EGFR* and *Notch* pathways (Held 2002, Bier 2000). *EGFR* (*DER*) is required for at least three phases during vein formation. First the locus is needed to establish the vein primordia but later signaling is suppressed to allow *dpp* to regulate vein cell formation. Interestingly, lingering low level *DER* signaling seems to be important for growth and survival of cells in the intervein regions. The third role is in crossvein formation, in response to *dpp* and *gbb* signaling (Yu *et al.* 1996, reviews Chapter 1 and Held (2002).

Quantitative trait locus mapping of wing shape shows that natural variation in the structure is affected by a large number of genes (Weber *et al.* 1999, Zimmerman *et al.* 2000, Weber *et al.* 2001). QTL effects are mainly additive but epistatic interactions have also been reported (Weber *et al.* 2001). While QTL studies generally have low resolution, randomization procedures can be applied to test if a specific subset of loci is over or under represented under the QTL peaks in a given study. By this logic Zimmerman *et al.* (2000) implicated vein-determining loci as they were over represented under the QTL peaks. One QTL mapping to the same region as EGFR was found to contribute to variation in the anterior part of the wing (IVR-B). QTL mapping procedures only contribute one piece of evidence. Allelic variation at a locus can also be exposed by a quantitative complementation test (Thompson 1975, MacKay and Fry 1996, Lyman and MacKay 1998). The test involves a cross of +/- stocks (where – stands for a deletion or mutation of a particular gene and + designates a balancer chromosome with a wild type copy of the same locus) to two or more wild type stocks. If alleles of wild stocks differ in the capacity to complement a deficiency then it will result in a significant line by genotype term in analysis of variance. As described in Chapter 2, I applied this approach by testing for the capacity of 6 wild type lines to suppress mutations at 15 loci known to be involved in wing development. In order to reduce the effects of genetic backgrounds, multiple alleles were tested for several loci and in some cases a preliminary cross to laboratory stocks enabled contrast of the mutation to another control chromosome. The results are consistent with segregating variation at several wing loci (*dpp, tkv, EGFR, spitz, argos, elbow* and *hh*) affecting distinct aspects of shape.

There was also heterogeneity among alleles possibly due to differences in type of lesion (a combination of large deletions and point mutations) or genetic backgrounds. The specific effects of EGFR varied between the two lesions tested. The *Df(2R)Pu-D17* deletion takes out a large chunk of the genome including three loci involved in EGFR signaling or wing development. *Misexpression suppressor of KSR number 2* (*MESK-2*) is located 10 kb upstream of *EGFR* and *Protein Tyrosine Phosphatase-ERK/Enhancer of Ras1* (*PTP-ER*) 70 kb downstream, and both loci are known to have a direct impact on Ras/MAPK signaling in

flies. The third gene is *crossvein-less 2* (*cv-2*) and is required for crossvein formation, thus overlaps in function with DER. The *Df(2R)Pu-D17* deletion is therefore not an explicit test of *EGFR* function alone. That test is provided by the *f2* (also known as *1K35*) allele, a point mutation in the first cystein repeat of the ligand domain which creates an ochre termination codon (Clifford and Schüpbach 1994). The 267 amino acid peptide is incapable of normal function as the allele coordinately affects all activities of the gene, through embryonic and pupal development (Clifford and Schüpbach 1989). Tests of the big deletion indicate that parameters of intervein regions B, C and D are affected by segregating variation in the locus or linked genes. The *f2* allele offered a more conservative assessment, but still corroborated the inference that segregating variation at the locus affects regions B and C. In a broader context then the results are also consistent with the role of two other loci involved in *EGFR* signaling, the ligand *spitz* and the antagonist *argos*. Interpretation of the genotype by line term comes with two important caveats. The interaction could be caused by epistatic interactions between the mutation tested and other genes in the genome. Secondly, the interaction could be caused by allelism to other mutations on the chromosome tested. Our results must therefore be considered corroborative, not conclusive.

The direct consequences of a loss of gene function on subtle aspects of wing shape have not been carefully documented. Most loci known to be involved in wing development have been characterized on the basis of gross phenotype, manifested by their names: *wingless, vestigial, Notch* and *veinlet*. There is a subset of loci known to impact shape directly, but those effects have not been characterized with the tools of morphometrics. Here I investigate the function of *EGFR* for *Drosophila* wing shape, by reexamining data from Chapter 2 on the effects of two alleles Ellipse (E1) and *f2*. E1 is a gain of function mutation altering a conserved Alanine in position 887 to a Threonine, and increases signaling activity. E1 was identified by the elliptical shape of the *Drosophila* eye that results from abnormal photoreceptor determination. In the wing the E1 allele causes extra vein formation. The contrary effect of the loss of function allele *f2* is a reduction in vein material. The causal relationship between vein formation and shape has yet to be established, but intuitively one way to realize shape is to regulate vein development.

It was on basis on developmental genetic reasoning that I chose to investigate the variation in shape in sub-sets of landmarks surrounding individual intervein regions (IVR) (Birdsall *et al.* 2000 and Zimmerman *et al.* 2000). Each of the three IVR were found to be affected by segregating variation in distinct genetic factors (Zimmerman *et al.* 2000, Chapter 2). Other applications of modern morphometrics to insect wings (Klingenberg and Zaklan 2000, discussed in Chapter 1) have approached the whole wing blade as a

composite trait, instead of subdividing the structure. Here I applied both modes of shape analysis. This allows a contrast of the two approaches in terms of investigating insect wing shape variation, facilitating understanding about the relative contribution of standing variation to developmental integration or independence. Finally, a more comprehensive representation of the phenotype was beneficial with respect to the goal of examining the genetics of a complex trait.

Sequence variation in EGFR

Numerous tools, for instance RFLP, microsatellite, allele specific oligos (ASO), Luminex, CAPs or DNA sequencing can be used to determine molecularly the genotypic state of subjects (Syvanen 2001). Most of these techniques with the exception of sequencing have been used in previous association tests in *Drosophila*. A pilot study of nucleotide variation in the locus or region of interest is normally followed by a second phase where a subset of variants is scored in large study population.

Studies of molecular variation in *Drosophila* have documented high nucleotide diversity (average π = 0.011) on all three major chromosomes (Aquadro *et al.* 2001). The fourth chromosome carries 1% of the euchromatin and has low diversity and nearly no recombination (Berry *et al.* 1991, Wang *et al.* 2002). The high diversity suggests that thorough description of the polymorphism at a candidate locus may be achieved most cost efficiently by DNA sequencing. Furthermore linkage disequilibrium in *Drosophila* rarely extends beyond 1 kb and as mapping by association depends on LD then this suggests we need to sample variants at very short intervals along the locus. But even such sampling does not guarantee positive identification of QTN's if they are present. These facts along with the decreasing cost of DNA sequencing convinced me that genotyping by sequencing was a practical option. Studies on nucleotide variation in *Lipoprotein Lipase* in humans (Clark *et al.* 1998) that led to tests of association between a fully sequenced genotype matrix and disease (cardiac failure) can be considered a proof of the principal. Chapter 3 described the sequencing of 10.9 kb spanning the coding and adjacent non-coding regions of *Drosophila EGFR* from 210 lines. We observe normal levels of diversity on average with notable fluctuations along the locus coinciding mostly with estimates of sequence divergence both to *D. simulans* and *D. pseudoobscura* (Figure 3.2 and 3.3 in Chapter 3). This pattern serves as a baseline for questions on the relation between the descriptors of nucleotide diversity and the distribution of quantitative trait nucleotides along a locus. The second main conclusion confirms previous estimates of LD decay in the *Drosophila* genome, as r^2 drops in less than a kb and only 17 pairs of sites out of the 60,000 tested gave an r^2 value equivalent to 1. All of these are short range, the maximum distance being

150 bp between two sites in intron 3. Consequently, tests of association are expected to be able to discriminate between the effects of even closely linked polymorphisms on phenotypes.

As discussed in Chapter 3 estimates of F_{ST} for SNP's in *EGFR* do not demonstrate major differences between the two study populations. The values range up to 0.13 and are significant, but the top seven sites are all independent. Stochastic fluctuations, complex demographic history or selection are all plausible causes. Carachristi and Schlötterer (2003) sampled West and East coast (including 30 West End lines) along with European and African populations and their results support demographic models. Their Californian samples were of clear European decent, while the East Coast populations all show evidence of admixture of African and European alleles. The emerging evidence therefore supports a model with a degree of population structure which may reflect on association tests in two ways. Heterogeneity in allele frequency between populations can create spurious associations particularly in the case of phenotypic disparity. Sites exhibiting high F_{ST} values were monitored carefully in the following analysis. Second, population differentiation may contribute to dependence of allelic effects since the genomes may differ in frequency at multiple loci. That can be assessed explicitly in the analysis of variance model used to test for associations.

Hypothesis

The aim of this chapter was to establish wing shape in *Drosophila* as a model multidimensional trait, and to elucidate the potential contribution of polymorphisms in a candidate locus to wing shape. Explicitly, I tested the hypothesis that nucleotide polymorphisms in the vein-determining gene *EGFR* contributes to natural variation in wing shape. The approach was to couple high-throughput genotyping and unbiased extraction of the phenotypic variance in two panels of inbred lines of *D. melanogaster*. This will combine the power of controlled quantitative genetic experimentation in *Drosophila* with robust statistical analysis of shape, leading to exact tests of the traits dependence on polymorphisms in *EGFR*, sex and population of origin.

Materials and Methods

Fly stocks and husbandry

Inbred lines came from two North American populations of *D. melanogaster* (Chapter 3). A total of 80 Californian and 124 North Carolina lines were used to test for associations. Flies were reared on 10 ml standard cornmeal medium at 25ºC on a constant light/dark cycle. Larval density was controlled to yield 50 - 100 flies per vial, of which 10 of each sex were randomly picked for phenotyping. A total of 3 replicate vials per line were sampled in independent blocks, bringing the number of flies scored per line and sex to 30. The exceptions were 30 lines with poor fitness. Of this set we included only lines with a minimum of 10 individuals per sex scored from replicate vials. Handling and digitizing of wings follows our earlier protocol (Birdsall *et al.* 1999, Zimmerman *et al.* 2000 and Palsson and Gibson 2000: i.e. Chapter 2). 3-6 day old individuals had their right wings excised at the hinge by micro-scissors. Wings were arranged on a microscope slide and carefully wedged under a cover slip. Each wing was digitized within 48 hours at constant (4x) magnification with a Spot camera (Diagnostic Instruments Inc.) mounted on a Nikon Eclipse E800 microscope. Images were saved in TIFF format and recorded to CD's for storage and shape analysis.

Analysis of wing shape

Wings where analyzed in Scion Image software version Beta 4.0.2 (Scion Corporation http://www.scioncorp.com). Shape was captured by the location of 9 landmarks (Figure 4.1) at the junction of the veins and the wing margin, all being Type 1 landmarks (Bookstein 1991, 1996b). All 9 landmarks are used as a basis for shape parameters W1-W9. Coordinates demarcating individual intervein-regions (IVR), for instance landmarks 1, 2, 3 and 9 identifying IVR-D, comprise three separate datasets (IVR B, C and D). Subsets of X and Y coordinates corresponding to 3 defined inter-vein regions were extracted in Microsoft Excel (Figure 4.1). Similarly, computation of the length of the wing (trait L1), from landmark 6 to 8, was also implemented in Excel. Finally size of inter-vein regions was calculated by standard geometric equations (Bonic 1971) from the landmarks of the corresponding regions. This yielded the area measures (B-Area, C-Area, D-Area and T-Area), where T designates Total, the sum of area of regions B, C and D. Shape parameters were estimated in TPS-Relw package version 1.2 (Rohlf 2002) available online at http://life.bio.sunysb.edu/morph. The two step procedure initially scales and rotates the

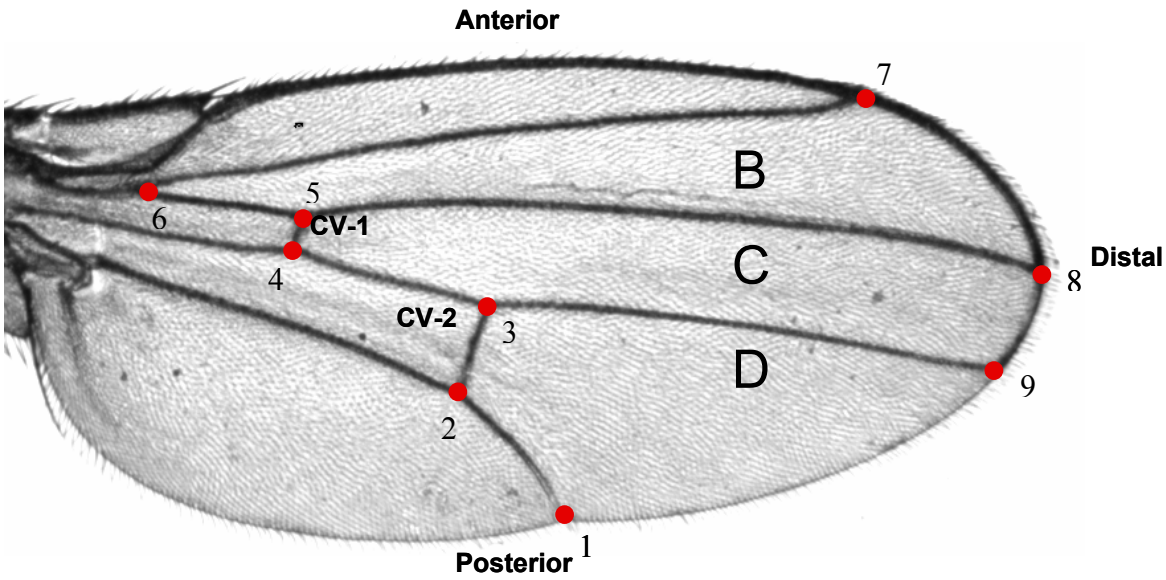


Figure 4.1. Stereotypical Drosophila wing, with the 9 landmarks indicated by red dots. The capital letters represent the 3 intervein regions analyzed. Cross-veins (cv-1 and cv-2) and anterior, posterior and distal portions are indicated. The proximal region (opposite of distal) includes the condensed and thick vein like material of the hinge.

specimens and then captures the axes of variation by a method akin to Principal Component Analysis. The first step is a General Procrustes Superimposition of the landmark data that effectively removes size. Centroid size is calculated for each specimen and they are then scaled to the mean centroid size of the dataset. Pairs of specimens are rotated to minimize the summed square distances between them. The whole procedure is iterated until the minimum distance over all landmarks and specimens is achieved (Dryden and Mardia 1998). Consensus shape for the dataset is calculated after rotation and alignment, and the adjusted coordinates for each specimen subjected to the second procedure, Thin Plate Spline (TPS) analysis. TPS offers a unique solution to D'arcy Thompson-type deformation graphics and accounts both for uniform and non-uniform changes in shape. The uniform component summarizes global changes to the grid, for instance shear, which refers to changes of angles that leave lines parallel. Local changes in shape are summarized by the non-uniform component after calculating the energy matrix required for the distortion or bending to align all the specimens to the consensus. The relative warps are essentially weighted principal components of the variation in landmarks and are therefore orthogonal descriptors.

The major aim of multivariate statistics, like relative warp analysis, is to reduce the dimensionality of datasets. We previously used 2 or 3 parameters per intervein region (B, C and D, see Chapter 2, Palsson and Gibson 2000 and Zimmerman *et al.* 2000) and chose here to study also 9 parameters for the whole wing (W metrics). While relative warp analysis extracts (2 x N – 4) parameters of shape, where N is the number of landmarks, we need to decide on how many warrant further study. Hatcher (1994) outlines four rules to aid decisions on which parameters to analyze. These include the eigenvalues criteria, a "scree" test, percent variation accounted for and the interpretability of parameters. The most stringent test is the Kaiser criterion, which builds on the fact that eigenvalues correspond to units of variance captured, with each of the observed variables contributing 1 unit. As the aim is to reduce dimensions then it is intuitive to dismiss components accounting for variation less than what is contributed by a single observation. The rule is easily applicable with the exception of decisions about components with eigenvalues just below 1. Only 11 of our eighteen shape parameters pass this test (Table 4.1). The "scree" test demands visual inspection of the decay in eigenvalues with a downwards slump in the slope used as criteria to discard all consecutive components. The distribution is not uniform with a particular bend between warps 6 and 7 for the whole wing (Table 4.1, graph not shown). The third criterion uses the proportion of variance accounted for to select components, either by selecting minimum percentage per component or cut-off for a cumulative portion explained. This method is subjective in the sense that the experimenter must decide the percentage values

Table 4.1. Summary of relative warps from intervein regions B,C and D, and the whole wing.

Trait	Eigenvalue	Explained (%)	Cumulative (%)	RWS
B1	1.91	58.5	58.5	0.9986
B2	1.23	24.2	82.7	0.9907
B3	0.95	14.5	97.2	0.9918
C1	1.68	62.5	62.5	0.9992
C2	0.89	17.5	80.0	0.9816
C3	0.67	9.9	89.9	0.9665
D1	2.54	54.6	54.6	0.9997
D2	1.55	20.4	75.0	0.9914
D3	1.42	17.1	92.1	0.9862
W1	1.94	29.7	29.7	0.9860
W2	1.68	22.4	52.1	0.9739
W3	1.31	13.5	65.6	0.9693
W4	1.11	9.8	75.4	0.9571
W5	0.99	7.8	83.1	0.9607
W6	0.91	6.6	89.7	0.9821
W7	0.65	3.3	93.1	0.9892
W8	0.49	1.9	94.9	0.9527
W9	0.44	1.5	96.5	0.9509

The dataset has total of 12,531 wings from the two populations.
Eigenvalues, of each wing shape parameter as derived from TPS-Relative warp analysis.
RWS: Relative warp stability, calculated by Pearson correlation between specimen trait values estimated from the inbred panel (12,531 wings) to trait values estimated as part of the 40,626 wings scored in the lab.

used. We previously chose parameters explaining 10% individually and between 80-90% cumulative. By this criterion then factors W5-W9 should be disregarded. The fourth rule concerns interpretability of the components in question (Hatcher 1994). Discussions about the meaning of components in multivariate statistics center on factorial vs. principal component analysis, and underlying models of causes. Analysis of wing shape with Relative warps is effectively a Principal Components based analysis and does therefore not rest on assumptions of explicit causes. It is however unique in terms of capturing shapes of actual structures, where the extremes can be inspected and evaluated. Figures 4.2 to 4.5 depict consensus configurations for the extreme 5% at either end of the distribution for each of the 18 shape parameters demonstrating the alterations in shape. For instance B1 in the anterior part of the wing affects the length of wing margin between landmarks 7 and 8, and W8 which almost exclusively describes widening between two landmarks in the distal part of the wing. The relative warp analysis proceeds by dividing up the variation in landmarks and will by default eventually describe small localized events like those captured in parameters W7-9. The conundrum is deciding if those small components, with eigenvalues less than 1, that fail the scree test and account for less than 10% of the variance deserve further investigation. The clear interpretability of these shape metrics argues in favor of including them in the analysis. Also the exploratory nature of the experiment provides one reason to include these parameters, as the goal is to explore our power to distinguish the heritable constituents of composite traits. But inclusion of minor components also comes with the penalty of added dimensions, affecting experiment wide thresholds of significance.

Finally a key feature of the relative warp analysis is the dependence of estimated parameters on the dataset under study. This can potentially identify sample specific axes of variation, particularly in smaller experiments. Such potential bias was investigated by comparing correlations (Proc CORR in SAS) and absolute values of relative warp scores for individual specimens of inbred lines as the traits were calculated for the total dataset and a subset of data. I also calculated the trait values for the specimens after including ~28,000 more wings scored in the lab over a 5 year period. A benefit of computing common axes for the current dataset and the complementation experiment described in Chapter 2 is that the relative warps will have identical meaning in both experiments. This is not the first study that collapses a complex dataset with the tools of multivariate statistics before testing for associations. Long *et al.* (1998) approached the phenotypic space by collapsing the bristle counts in four genetic backgrounds into principal components before testing for association. Dissection of wing shape proceeded similarly, except that the geometric roots of the relative warp analysis are tailored to capturing variation in shape and can therefore be interpreted in that domain.

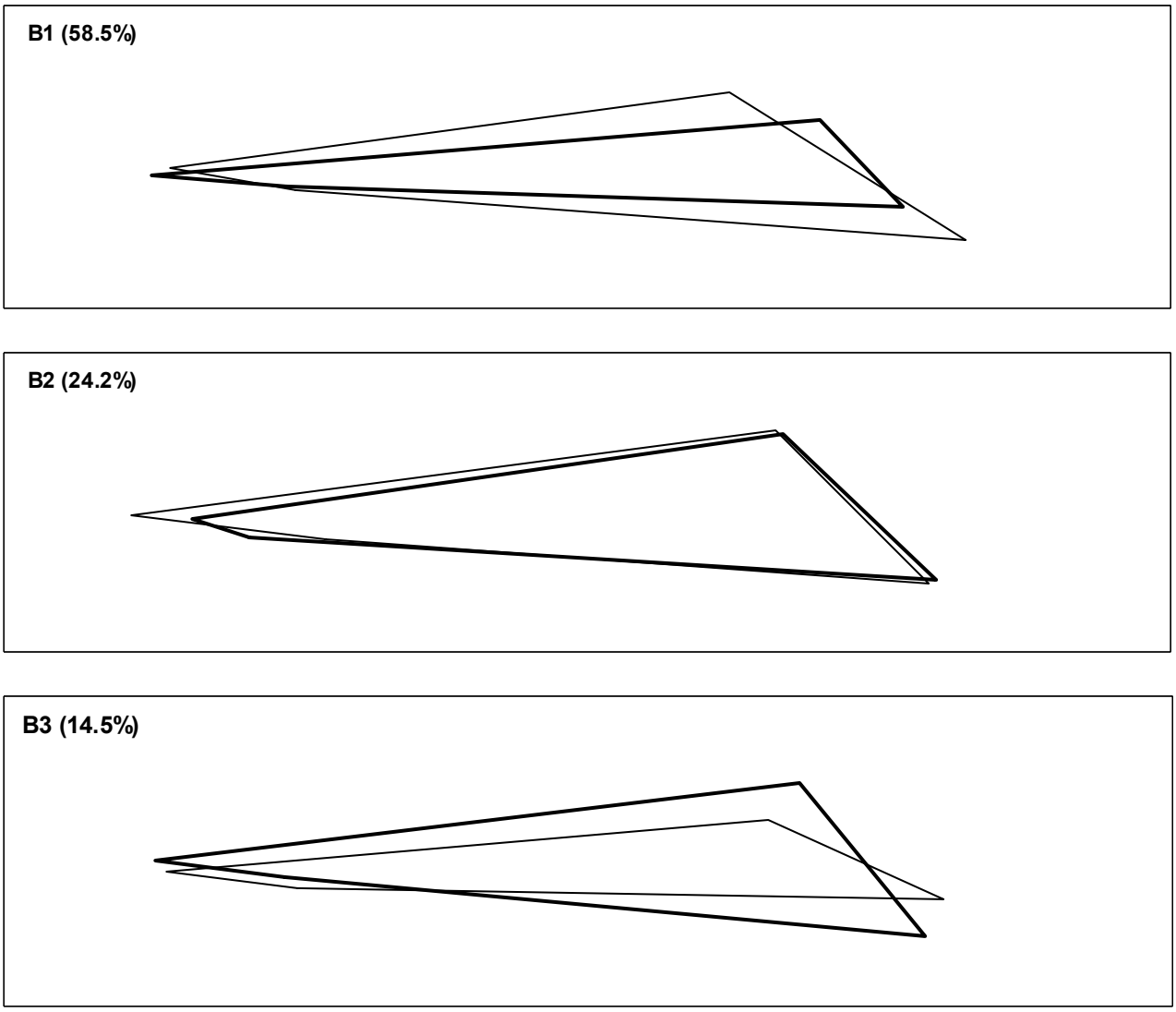


Figure 4.2. The consensus configurations for depicting the extreme 5% at both ends of the shape spectrum for intervein region B (B1-B3). Thin lines correspond to positive values and bold to negative. The proportion of total variance in landmark data explained by each warp is reported in brackets.

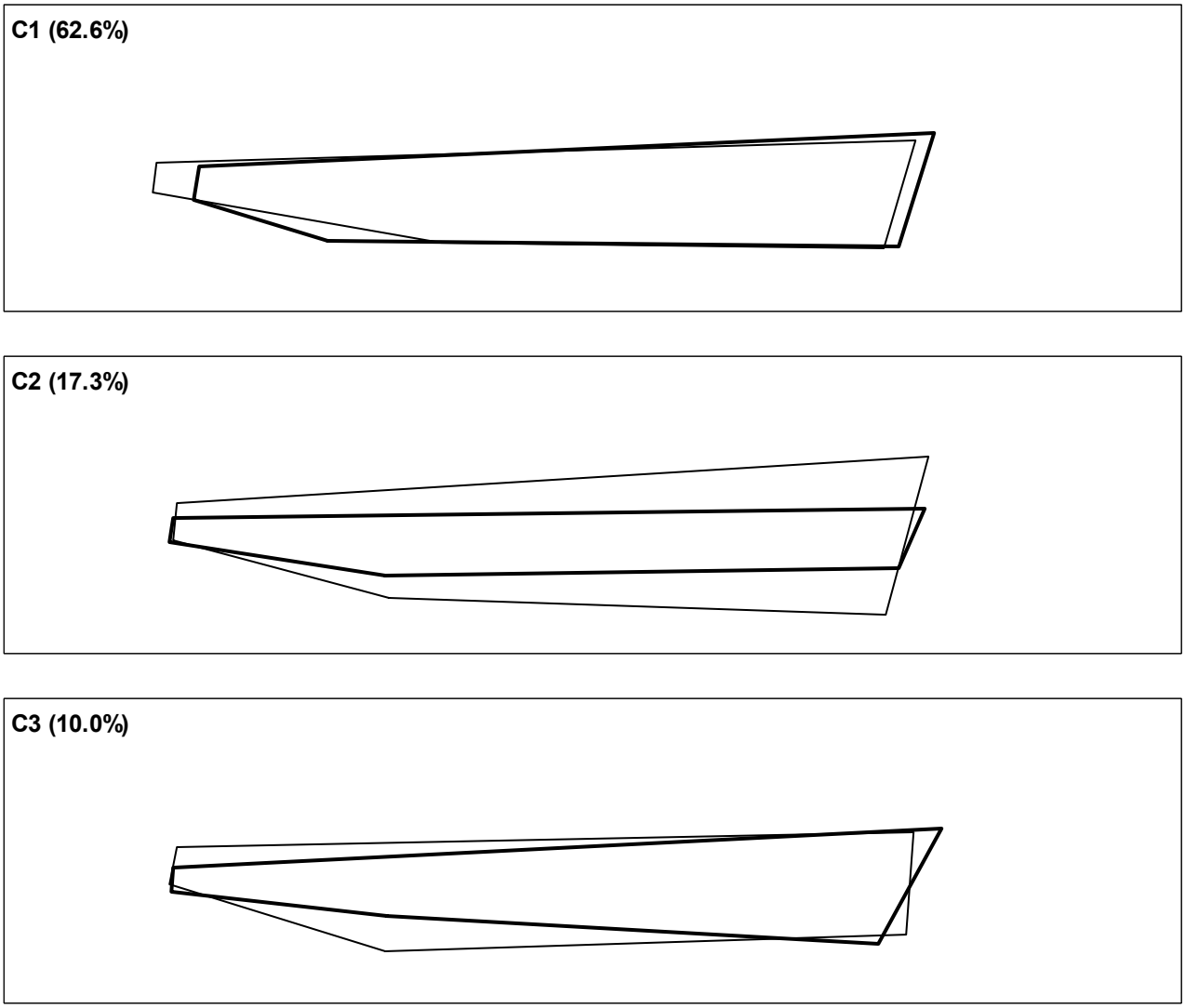


Figure 4.3. The consensus configurations for depicting the extreme 5% at both ends of the shape spectrum for intervein region C (C1-C3). Thin lines correspond to positive values and bold to negative. The proportion of total variance in landmark data explained by each warp is reported in brackets.

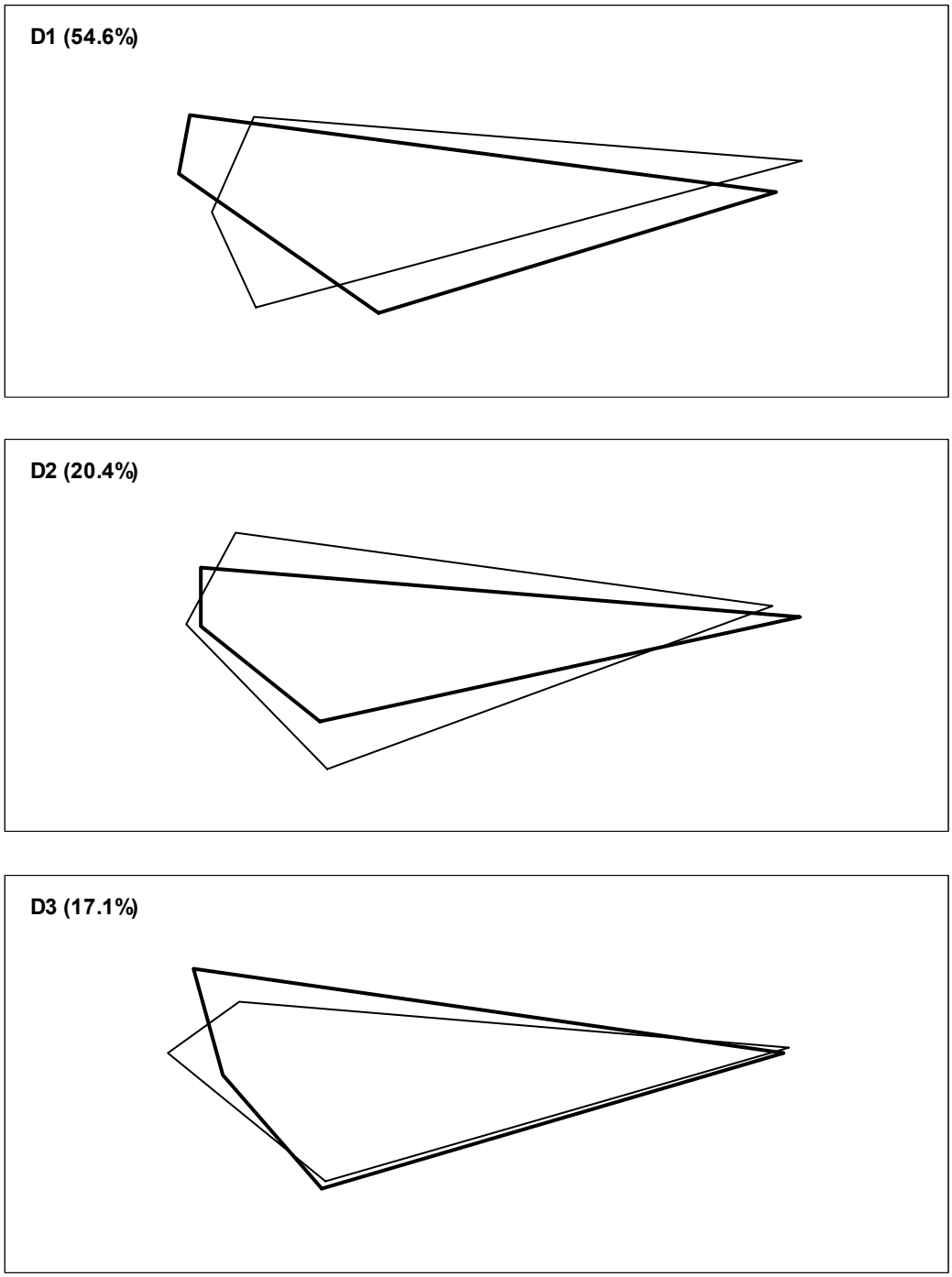
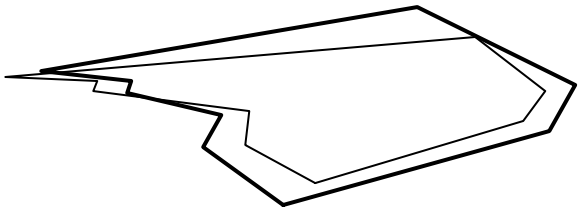
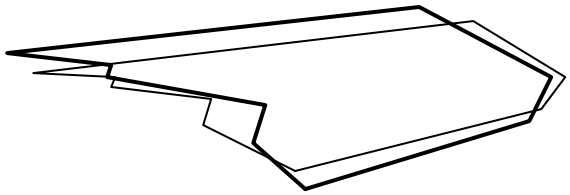


Figure 4.4. The consensus configurations for depicting the extreme 5% at both ends of the shape spectrum for intervein region D (D1-D3). Thin lines correspond to positive values and bold to negative. The proportion of total variance in landmark data explained by each warp is reported in brackets.

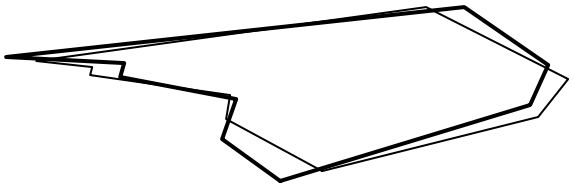
W1 (29.7%)



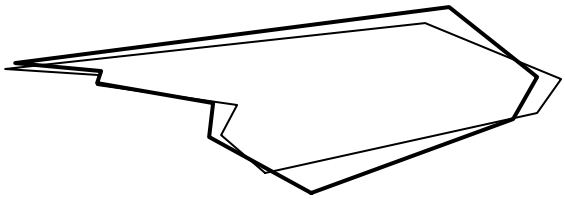
W2 (22.4%)



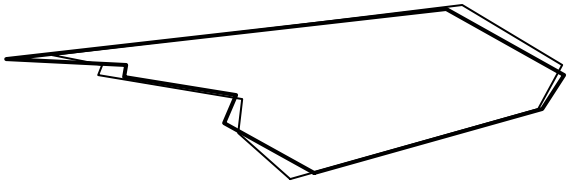
W3 (13.5%)



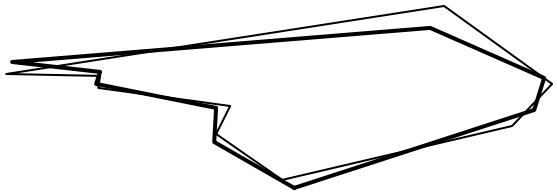
W4 (9.8%)



W5 (7.8%)



W6 (6.6%)



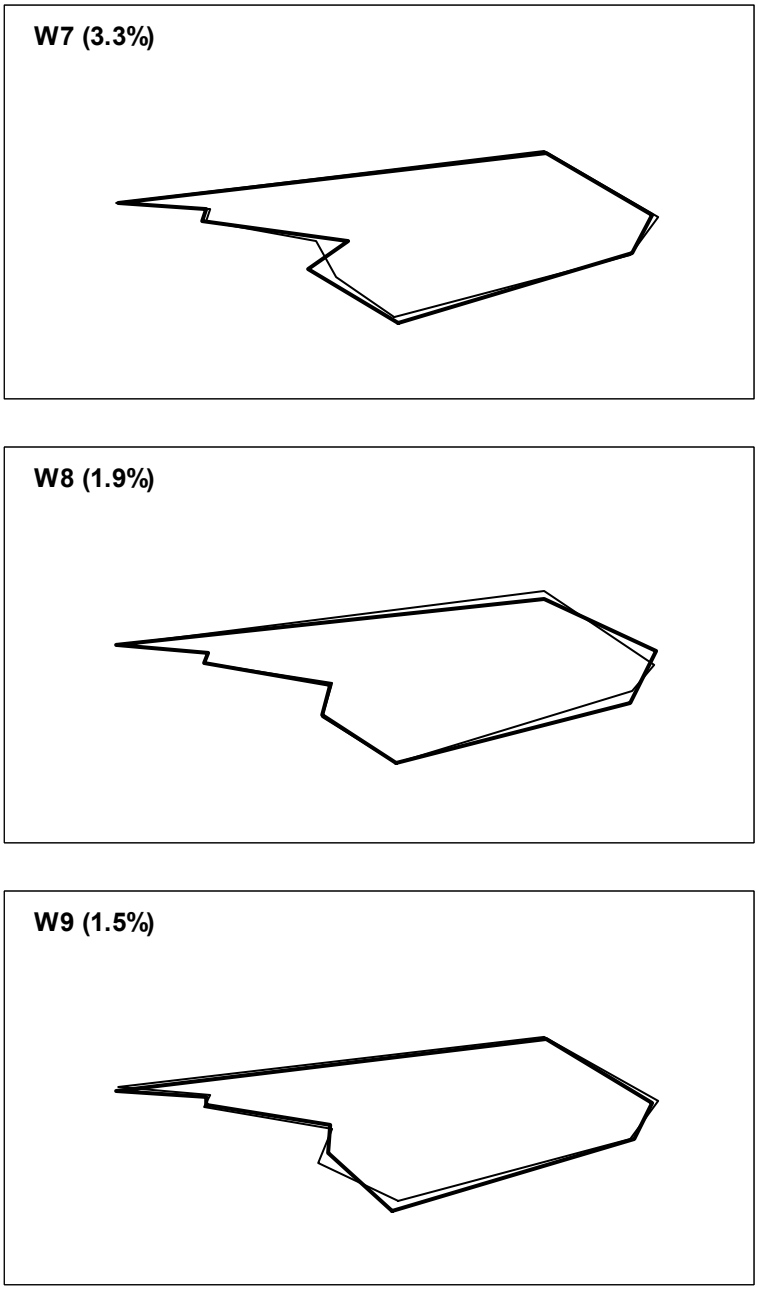


Figure 4.5. The consensus configurations for depicting the extreme 5% at both ends of the shape spectrum for whole wing warps (W1-W9). Slender lines correspond to positive values and bold to negative. The proportion of total variance in landmark data explained by each warp is reported in brackets.

Genotyping

Polymorphisms in *EGFR* were identified by sequencing ~11 kb corresponding to the 6 exons and flanking non-coding regions in the inbred lines as described in Chapter 3. In the inbred lines, WE and UCD, a total of 246 SNP's (5 replacements) and 21 indels at 0.05 frequency or higher constitute the genotypic matrix. SNP's and indels are numbered according to their location in a Genebank record 17571116 (Flybase number: FBgn0003731), which describes the 48 kb genomic region including *EGFR*. Polymorphisms will be labeled by ancestral base, location and derived base, for instance C30200T.

Statistical analysis of phenotypes

Descriptors of phenotypic distributions by populations and sex were calculated with Proc MEANS and Proc UNIVARIATE in SAS. Partitioning of the phenotypic variance was conducted with the SAS procedure Proc MIXED. I fit the terms population (POP) and sex (SEX) as fixed and lines (LINE) and replicates (REP) as random nested factors in the model.

$$Y = \mu + POP + SEX + POP \times SEX + LINE(POP) + SEX \times LINE(POP) + REP(LINE\ POP) + SEX \times REP(LINE\ POP) + \varepsilon$$

Reduced models were also run. Least square means warp scores for the terms POP, LINE(POP), SEX X LINE(POP) were calculated with the lsmeans option in Proc MIXED, and the difference between classes also assessed and tested with the estimate function.

Heritability and genetic correlations

Variance component estimation for the genetic and phenotypic factors were estimated with the Proc VARCOMP function in SAS. The inbred lines are considered clonal, since only 5-10% of lines differed at molecular markers (*EGFR* and *5HT-1A* data not shown) and trait heritabilities can be estimated. Assuming strict additivity then the component of additive genetic variance inferred from a set of isogenic lines is double that of an outbred population. Following this adjustment V_G is divided by V_P for the sexes separately. Heritability for sexes combined adds V_{GS} with the V_L to give the V_G. Genetic correlations between traits were estimated from the ratio of covariances of the traits divided by the square root of the product of genetic variances of individual traits (Falconer and MacKay 1996). Phenotypic correlations and covariances were estimated with Proc CORR for the 18 shape parameters and the 5 size measures divided by sexes. Calculation of genetic correlations for sexes combined gave some values that were out of bonds for the size measures, possibly due to the distribution of size traits being very different between the sexes. The genetic and phenotypic correlations were also estimated for datasets

divided by populations and sex. Confidence intervals were estimated for genetic correlations by using the z-function (Sokal and Rohlf 1995, Ungerer *et al.* 2001).

Tests of association

Tests of association between sequence variants in *EGFR* and individual parameters of wing shape were implemented with SAS and in Tassel, a java-based software package developed by Ed Buckler available at www.maizegenetics.org. The data matrix includes least square estimates of line means by sexes and the 246 SNP and 21 insertion deletion polymorphisms. In addition we tested three summary variables similar to transposon insertion vs. no-insertion classes in *achaete-scute* (MacKay and Langley 1990). Here the indicators lumped low frequency variants; major insertions around exon 1; and replacements or deletions in coding and non-coding portions of the transcript. A total of 18 shape parameters were analyzed along with 5 phenotypes of wing size and vein length.

Tassel enables assessment of significance of associations by permutation but does not accommodate sex or interaction terms, and therefore was only used for exploratory purposes. It can use estimates of population structure as a covariate but here populations were simply coded by an indicator. The analyses were conducted on data separated by sex either on individual datasets or joint set. The joint set was analyzed with and without population as a term. The significance of the ANOVA test-statistic of association of each site was assessed by contrasting it to distribution of test-statistics derived from analysis of 10000 random permutations of the data. Permutations involved randomizing the phenotypes in respect to the genotype matrix, thus retaining patterns of linkage disequilibrium. Tassel enables two kinds of comparisons to be made with the test statistic of each particular site in reference to the permuted data. Each site can be contrasted to distribution of test statistics derived from that exact site only assessing local significance. It can also be compared to the pool of all sites along the gene, estimating experiment wide significance.

In order to assess other sources of variation and interactions we tested a type 3 ANOVA model with the Proc Mixed function in SAS. The effects of polymorphism (SNP), population (POP) and Sex are considered fixed. SNP (Single Nucleotide Polymorphism) is defined broadly to include insertion-deletion polymorphisms. Line was included as a random term nested within polymorphism and population.

$$Y = \mu + SNP + SEX + POP + SNP \times SEX + SNP \times POP + SEX \times POP + SNP \times SEX \times POP + LINE(SNP\ POP) + SEX \times LINE(SNP\ POP) + \varepsilon$$

The model was reduced by omitting interaction terms between fixed factors but retaining the nested line term to avoid inflated test-statistics due to pseudo-replication. The effects and

significance of difference in polymorphism states are found with the lsmeans / diffs options in Proc Mixed. Our interest is in the genetic contribution to phenotype, i.e. terms including SNP, asking about contribution of the site to phenotype (SNP), and also about dependence on sex (SNP*SEX), population (SNP*POP) or both (SNP*SEX*POP). Analysis of reduced models showed that tests of associations to traits (B3, C2 and C3) required the population by sex interaction (SEX*POP) term to avoid excessive significance in the three-way interaction term. This is caused by the distribution of the traits differing significantly between sexes and populations. Results from the full model are reported.

The multiplicity of tests is an issue with 18 parameters of shape and 267 SNP's and Indels in EGFR. 17 sites can be excluded because they are in perfect LD, (r^2=1) to another sites, and there is also a 150 bp cluster of 7 highly linked sites in intron 3 that can not be regarded as independent tests. They comprise a heavy cluster of LD as represented on Figure 3.6 in Chapter 3. They may represent a deep haplotype structure in the region (data not shown). Cutoffs were determined by Bonferroni correction (Sokal and Rohlf 1995) accounting for 250 sites and 20 trait variables (18 shape and 2 size). Considering correlations between traits and linkage disequilibrium in the genotype matrix (see Chapter 3) that reduce the independence of individual tests, these corrections are conservative. Significance of *p*-value matrixes was also calculated with Proc MULTTEST in SAS.

Repeatability experiments

Two separate repeatability experiments were conducted, one by recrossing WE lines (Round robin or Experiment 2) and the other on sample of Kenyan chromosomes in a common background of Samarkand (Test cross or Experiment 3). This allows retesting of sites significant in the inbred lines. Experiment 2 involved round-robin crosses between 71 randomly chosen West End lines. The crosses were set up by randomizing the 71 lines with respect to each other, in 3 discrete blocks. As a consequence, each line was crossed three times as male and three times as female. We scored 8-10 females of each F1 generation cross in two replicates, set up 4 months apart. The second repeat involved an independent set of Kenyan alleles. 36 second chromosomes were substituted into the Samarkand (Sam) background utilizing stocks kindly provided by Trudy MacKay. In parallel an *EGFR* allele, *Ellipse* (E1) and a *blistered* allele were substituted similarly into Sam. The wild-type chromosomes were tested over these two mutations and Samarkand second chromosomes. Ten individuals of each sex were scored from three replicate crosses arranged in random blocks. All protocols of rearing, handling, and phenotyping were identical to previous experiments.

Genotypes for experiment 2 where deduced from allelic matrices of the WE lines. This allowed the construction of two homozygous classes and the heterozygotes. F1's missing a

genotype of one parent were treated as missing data. For experiment 3, the Kenyan chromosomes were sequenced as described in Chapter 3. They differed by 90 private sites and lacked 50 that are common to the North American sample. This only affected one site being retested. Also similar to the inbred dataset, there was heterogeneity in the depth of sampling between sites, due to incomplete genotyping. This prevented tests of two more sites in the Kenyan lines.

The estimation of line effects and extraction of line means was conducted by Proc GLM, and the LSMEANS options in SAS. The model for experiment 2, the round-robin, was:

$$Y = \mu + LINE + REP(LINE) + \varepsilon$$

where LINE represents each of the F1 generated by the round-robin crosses and REP the replicate vial. For the Kenyan chromosomes (experiment 3) the Proc MIXED model was more complicated accounting for the effects of CROSS (to Sam, *Ellipse* and *blistered*), SEX or LINE.

$$Y = \mu + CROSS + SEX + CROSS \times SEX + LINE + SEX \times LINE + CROSS \times LINE + CROSS \times SEX \times LINE + REP(CROSS \times LINE) + SEX \times REP(CROSS \times LINE) + \varepsilon$$

The line means obtained from this analysis were used in the tests of association, with models built in a similar way as before. For the Round-Robin experiment, the model in Proc GLM was:

$$Y = \mu + SNP + REP(SNP) + \varepsilon$$

With SNP being the genetic term and REP the replicate vials tested. The additional term of block was dropped from model as it was never significant (and did not alter the results, data not shown). The Mixed model ANOVA testing for the SNP effects in the Test cross experiment was again more elaborate:

$$Y = \mu + SNP + SEX + CROSS + SNP*SEX + SNP*CROSS + SNP*SEX*CROSS + LINE(SNP\ CROSS) + SEX*LINE(SNP\ CROSS) + \varepsilon$$

Here LINE is again treated as a random factor nested within the fixed terms SNP and CROSS to account for pseudoreplication. The mean effect of polymorphisms were estimated by the lsmeans option in the SAS function GLM. Reduced models were also conducted, by crosses.

Results

Shape analysis by relative warps

The relative warp procedure extracts major components of shape for a particular dataset. Because experiments differ by datasets then warp parameters can identify distinct axes of variation which can differ substantially, particularly between small samples. To address this I

calculated warp score for the total dataset (12,531 wings) and a subset including 70% of the WE wings. If the axes of variation are distinct between the populations then one expects a breakdown in correspondence. This can be formally tested by correlation and comparison of absolute values. However good congruence was found (Pearson correlation > 0.95, $p < 0.0001$) between the trait values derived from the partial and larger datasets, suggesting that the axes of wing shape variation are shared between the populations. Similarly inclusion of 3 other datasets, representing wings from crosses among WE lines, Test crosses with Kenyan chromosomes and tests of 15 heterozygous wing mutants (total of 40,626 wings), did not result in major shift in the axes of variation. The correlations between older parameters for the inbred lines and those derived for the combined set are high, ranging from 0.95 to 0.999 (Table 4.1).

Effects of EGFR on shape

The comparison of phenotypic values of three *EGFR* alleles after crossing to six wild type lines was used to evaluate the effect of the locus on wing shape. Contrary to the parameters from the inbred lines, the parameters for the dataset including major genes were more sensitive to recalculation of shape parameters. This means that mutants may cause novel deformations in shape, not commonly found in wild specimens. The *EGFR* alleles were tested over 6 wild type chromosomes, and inspected for consistent effect over backgrounds. The results for shape parameter W1 suggest that loss of *EGFR* function leads to shape changes corresponding to lower relative warp scores (Figure 4.6). The loss of function alleles had lower trait value than the E1 class in all crosses, sexes and lines. This translates into loss of *EGFR* leading to increased length of the wing, which changes the shape of both anterior and posterior regions, and extends the distance between the crossveins. The results across other traits are less convincing and confounded by differences between alleles (for C1) or background factors attributable to balancer stock (as in the case of D1: Figure 4.6, middle and lower panels).

Partitioning phenotypic variance

The distribution of shapes is summarized by least square means, standard deviation, skewness and kurtosis by sex and population in Appendix B. Negative kurtosis (14/16) was more prevalent in the UC population indicating flatter curves. Skewness showed no population dependence but some traits are consistently skewed, for instance B1 distribution is positively skewed in both populations and sexes.

The contributions of sex, population and line were tested with a mixed model analysis of variance for each of the traits (Appendix C). Line was highly significant in all cases, signifying a genetic component. Variance components for line and total variance were used to estimate trait heritability by sex and population (Falconer and MacKay 1996) (Appendix D). The values ranged from 0.28 to 0.68 with a definite skew towards high values (Table 4.2). Heritabilities

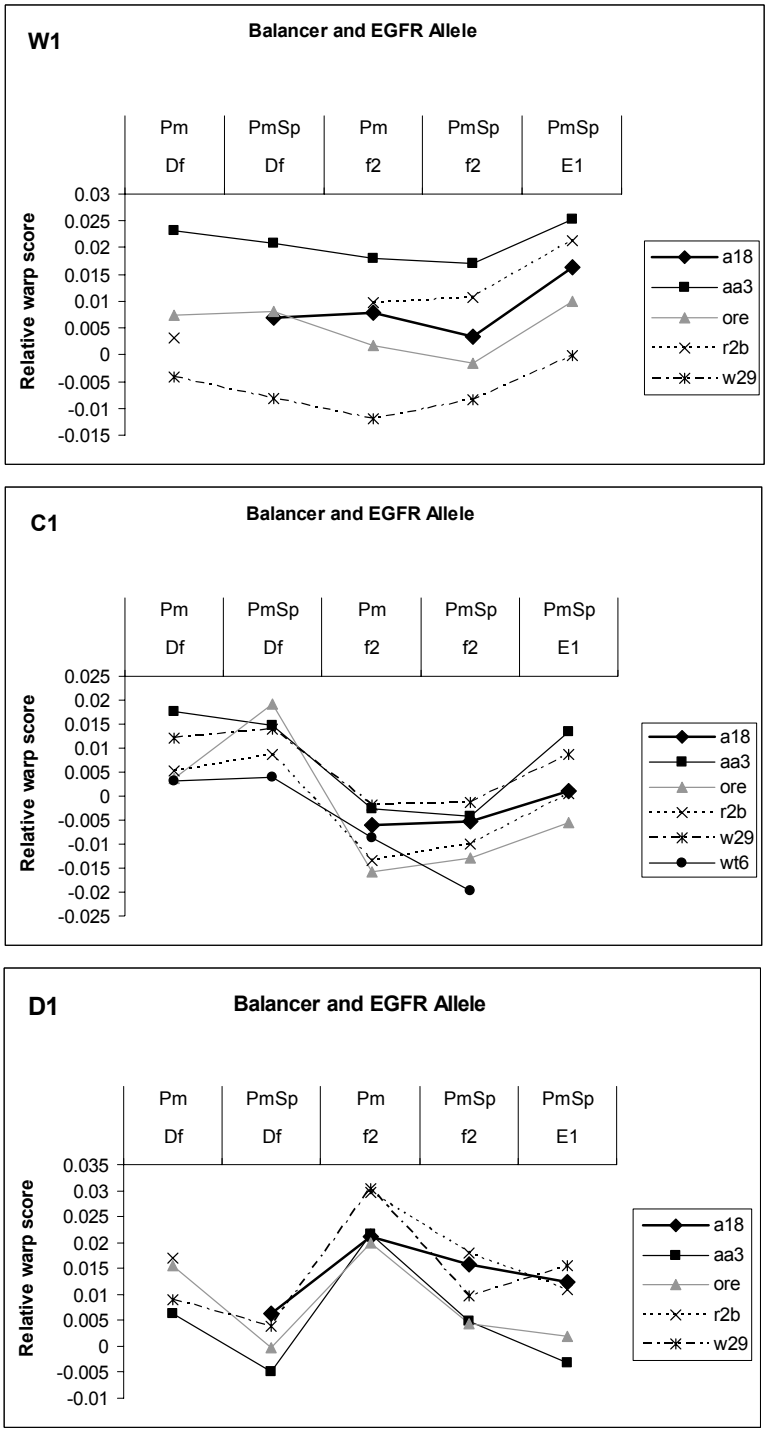


Figure 4.6. The allelic effects of EGFR mutations *Df(2R)Pu-D17* (Df) and *torpedo* (f2), both loss of function and the gain of function *Ellipse* (E1) on wing shape W1, C1 and D1. Assayed in heterozygous condition in 5 or 6 wild type backgrounds (see legend). The backgrounds are indicated by lines and explained in caption (see also chapter 2 for details). The loss of function alleles were tested twice, differing by balancer chromosomes (*Pm, PmSp*) that carried the mutation into the testcross.

Table 4.2. Heritability of 18 measures of shape (B1-W9) and 5 size related measures (T Area – L1)

	UC		WE		UC-WE[1]
	h^2_F	h^2_M	h^2_F	h^2_M	h^2_{FM}
B1	0.61	0.56	0.48	0.50	0.51
B2	0.55	0.51	0.43	0.42	0.46
B3	0.55	0.53	0.42	0.44	0.46
C1	0.58	0.59	0.43	0.47	0.50
C2	0.47	0.48	0.48	0.45	0.45
C3	0.38	0.40	0.35	0.32	0.34
D1	0.54	0.55	0.47	0.46	0.49
D2	0.49	0.48	0.41	0.42	0.43
D3	0.36	0.34	0.34	0.36	0.34
W1	0.67	0.68	0.56	0.59	0.61
W2	0.61	0.61	0.51	0.49	0.53
W3	0.50	0.50	0.45	0.47	0.46
W4	0.54	0.52	0.48	0.48	0.49
W5	0.54	0.53	0.41	0.38	0.46
W6	0.47	0.42	0.40	0.40	0.40
W7	0.30	0.30	0.26	0.30	0.28
W8	0.42	0.42	0.39	0.37	0.39
W9	0.41	0.38	0.37	0.34	0.37
T Area	0.23	0.25	0.17	0.17	0.19
B Area	0.27	0.28	0.22	0.20	0.23
C Area	0.27	0.30	0.19	0.20	0.22
D Area	0.22	0.22	0.20	0.19	0.19
L1	0.22	0.24	0.15	0.16	0.17

1. Heritability estimated for sexes and populations combined (WE: West End, UC: U. of California Davis).

where very similar by sex but exhibited distinct population effects, with WE having lower values (21/23). The estimates from the joint population are comparable. Finally there is a slight negative relationship between the heritability and factor number, as warps explaining large amounts of variation have highest heritability, for instance the estimate for W1 is 0.61 when calculated for males from both populations but warp W9 is 0.37. The high heritabilities of W9 and W8 are still intriguing as they explain only 1.9% and 1.5% of the variance in landmark placement. Assessment of sex and interaction terms including sex and replication suggests considerable environmental and sex dependence. However the population term was significant for only 4 of the 18 traits, B2 (p=0.004614), C1 (p=0.049), W2 (p=0.011024) and W5 (p=0.009416) (Table 4.3). The absolute magnitude of trait deviations between populations and standard errors are also tabulated. The interaction between sex and population proved significant for W5, and was marginal with p-values between 0.05 and 0.01 for B3, D2 and W8. All the size variables had large population by sex interaction terms, though only Area-C was significantly affected by population (p = 0.040). These results are graphed for B1, W5 and area of the whole wing (Figure 4.7).

Correlations of traits

Since intervein regions B, C and D share certain landmarks some correlation among trait measures was expected. Similarly, intervein regions and whole wing analyses obviously share landmarks. Phenotypic and genetic correlations reveal these relationships and estimate their strength (Table 4.4). Phenotypic correlations between the 18 shape parameters are generally weak while the 5 size measures are all highly correlated to one another. Despite low correlation coefficients, 170 of 253 correlations were significant at the 0.0001 level. Significance of weak correlations is not surprising given the size of the dataset (~12,000 wings) but must not be over interpreted. A better estimator of independence is the squared correlation coefficient, and it shows only 3 pairs of traits with r^2 higher than 0.5, for the phenotypic correlations. The interpretation in this case is that variation in one trait will explain 50% or more of the variation in the second, as for instance W1 and B1, where r^2 = 0.61 (r = -0.78 with p < 0.0001). The other tightly correlated parameter pairs are B2/W5 (r = 0.52) and D3/W7 (r = 0.68). Inspection of the shape outlines in Figs 2-5 highlights the shape changes common to the paired traits. Four more pairs of traits show dependence of r^2 in the 0.25-0.50 range and 20 pairs have values between 0.05 and 0.25. This means that 126 pairs of traits do not show marked correlation. Examples of these relationships are seen in Figure 4.8. The estimated genetic correlations are higher as expected (Table 4.4 below diagonal) but the pattern is the same. It is noteworthy that the genetic correlations detect more logical relations to the size variables than the phenotypic correlations. For instance B1 and W1 are highly correlated to L1 which captures the length of the wing, but much less so to other size measures. This reflects the shape changes in B1/W1

Table 4.3. F statistic and significance of population effects analyzed with Mixed Model ANOVA.

	Population		Population by sex[1]		Population[2]	
Trait	F	*P*	F	*P*	Est	SE
B1	0.16	.	0.26	.	-0.0008	0.0019
B2	8.93	***	1.38	.	-0.0036	0.0012
B3	0.05	.	6.69	*	0.0002	0.0009
C1	4.07	*	3.02	.	-0.0036	0.0159
C2	0.34	.	0.63	.	-0.0005	0.0080
C3	0.03	.	0	.	0.0000	0.0061
D1	0.19	.	1.65	.	-0.0011	0.0026
D2	0.90	.	3.89	*	-0.0015	0.0015
D3	0.05	.	1.35	.	-0.0003	0.0013
W1	1.20	.	0.12	.	-0.0023	0.0021
W2	7.03	***	2.04	.	0.0047	0.0018
W3	0.13	.	0.92	.	0.0005	0.0013
W4	0.16	.	2.17	.	-0.0005	0.0011
W5	7.17	***	9.43	**	0.0026	0.0010
W6	1.86	.	1.70	.	-0.0012	0.0008
W7	0.00	.	0.00	..	0.0000	0.0006
W8	0.00	.	5.51	*	0.0000	0.0005
W9	0.03	.	1.76	.	-0.0001	0.0004
T Area	1.89	.	34.02	****	0.2192	0.1595
B Area	0.50	.	30.04	****	0.0478	0.0676
C Area	0.76	.	27.19	****	0.0515	0.0589
D Area	5.63	*	29.47	****	0.1189	0.0501
L1	0.13	.	19.69	****	0.0143	0.0394

1. Population by sex interaction analyzed with the same model as above.
2. Est: Estimated difference between the CA and NC populations and with standard errors (SE).
Significance of F-statistics "." >0.05, * 0.05-0.01, ** 0.01-0.001, *** 0.001-0.0001, **** <0.0001.

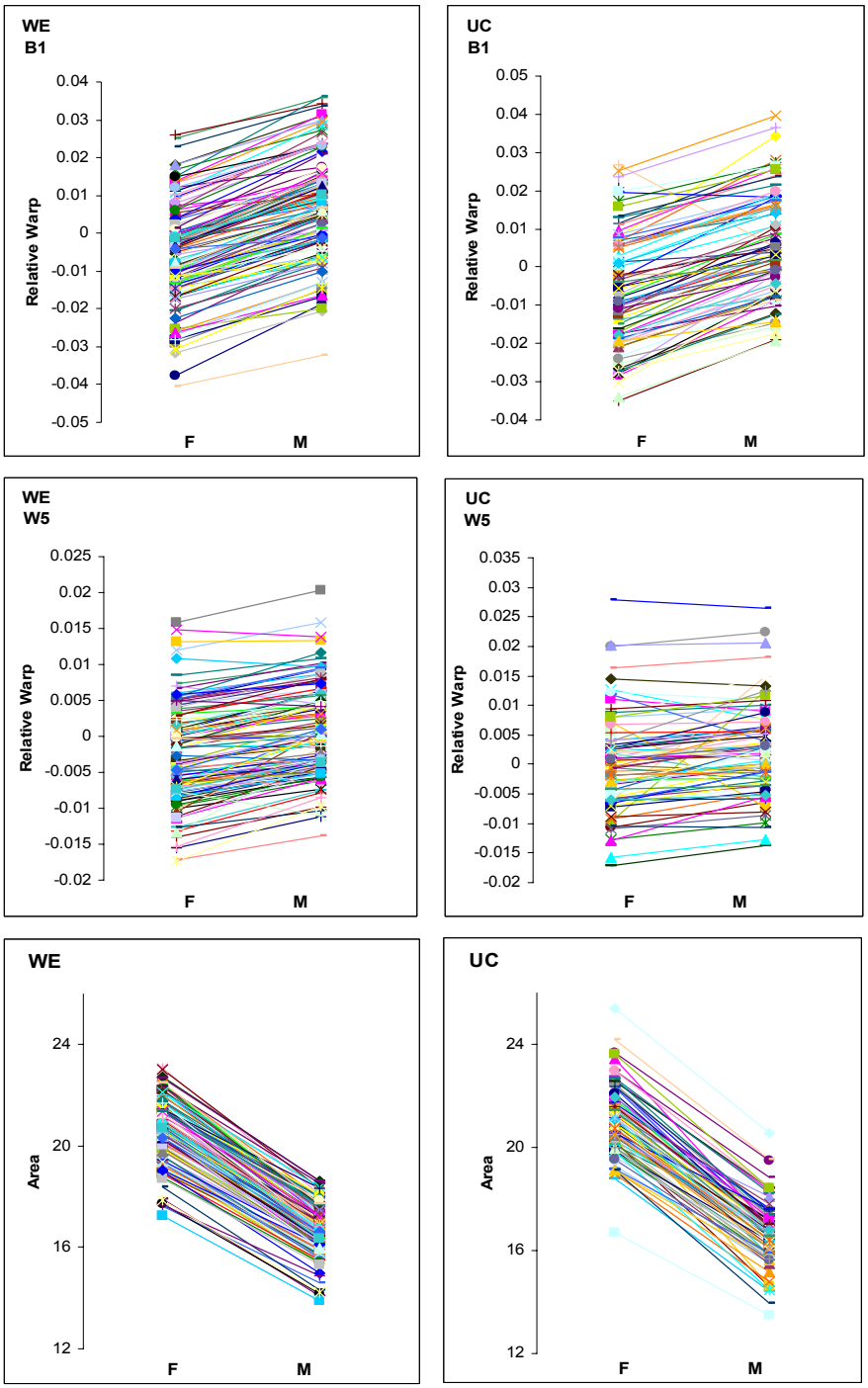


Figure 4.7. Estimated least square means for individual lines by sex and population. Three traits are represented, B1 (top panel), W5 (population by sex interaction significant at 0.01 level) and total area of intervein regions B, C and D (bottom). These are plotted individually for populations with WE (left) and UC (right).

Table 4.4. Phenotypic (above) and genetic (below) correlations between shape and size measures presented by r^2.

	B1	B2	B3	C1	C2	C3	D1	D2	D3	W1	W2	W3	W4	W5	W6	W7	W8	W9	T Area	B Area	C Area	D Area	L1
B1	*****	.	.	.	*	.	.	.	.	***	*	*	.	.	.	.	.	.	*	*	*	*	*
B2	.	*****	.	.	.	*	*	*	.	.	*	**	.	***	.	.	.	.	*	*	.	.	*
B3	*	.	*****	.	.	*	.	*	.	.	.	.	**	.	**	.	*	.	*	.	*	*	*
C1	*	.	.	*****	.	.	*	*	.	**	**	.	.	*	.	*	.	.	.	.	.	.	.
C2	*	*	.	.	*****	.	.	*	.	*	*	.	*	*	*	.	**	.	.	*	.	*	*
C3	.	*	*	.	.	*****	*	.	.	.	*	*	.	.	**	.	.	.	*	.	*	*	.
D1	.	*	.	*	*	*	*****	.	.	.	*	**	**	*	.	.	.	.	*	*	*	.	*
D2	.	*	*	**	*	*	.	*****	.	.	**	.	**	.	.	*	.	.	.	.	.	.	.
D3	.	.	.	.	.	.	.	.	*****	*	*	.	*	.	.	***	.	.	.	.	.	.	.
W1	****	*	.	****	*	*	.	.	*	*****	.	.	.	.	.	.	.	.	*	*	*	*	*
W2	**	*	.	****	**	*	**	***	*	.	*****	.	.	.	.	.	.	.	.	.	.	.	.
W3	**	****	*	*	**	**	***	*	.	.	.	*****	.	.	.	.	.	.	*	*	*	*	*
W4	*	.	***	*	*	*	****	***	*	.	.	.	*****	.	.	.	.	.	.	.	.	.	*	*
W5	*	****	.	*	*	.	**	.	*	.	.	.	.	*****	.	.	.	.	.	.	.	.	.	.
W6	.	.	****	*	*	****	.	.	.	.	.	.	.	.	*****	.	.	.	*	.	*	*	*
W7	.	.	.	**	.	.	.	***	****	.	.	.	.	.	.	*****	.	.	.	.	.	.	.	.
W8	.	.	***	.	****	.	.	.	.	.	.	.	.	.	.	.	*****	.	.	.	.	.	.	.
W9	.	.	.	.	*	.	.	*	*	.	.	.	.	.	.	.	.	*****	.	.	.	.	.	.
T Area	*	.	.	.	*	.	.	.	*	.	.	**	*	*	.	.	.	.	*****	*****	*****	*****	*****	*****
B Area	*	.	**	.	.	*	.	.	*	*	.	*	.	.	**	.	*	.	*****	*****	*****	*****	*****	*****
C Area	*	.	.	*	***	*	*	.	*	.	*	**	.	**	.	.	***	.	*****	*****	*****	*****	*****	*****
D Area	**	.	.	.	.	*	.	*	.	.	*	**	*	.	*	.	.	.	*****	*****	*****	*****	*****	*****
L1	***	.	.	*	.	*	.	.	.	***	.	*	.	.	*	.	*	.	*****	*****	*****	*****	*****	*****

Above diagonal are phenotypic correlations and below are genetic correlations between shape and size measures, calculated for both populations combined. Phenotypic correlations are estimated for sexes combined but genetic correlations are presented for males only.
The degree of independence is designated by the number of stars. r^2: ***** 1, **** 0.99-0.75, *** 0.75-0.50, ** 0.50-0.25, * 0.25-0.05, "." 0.05-0.0.

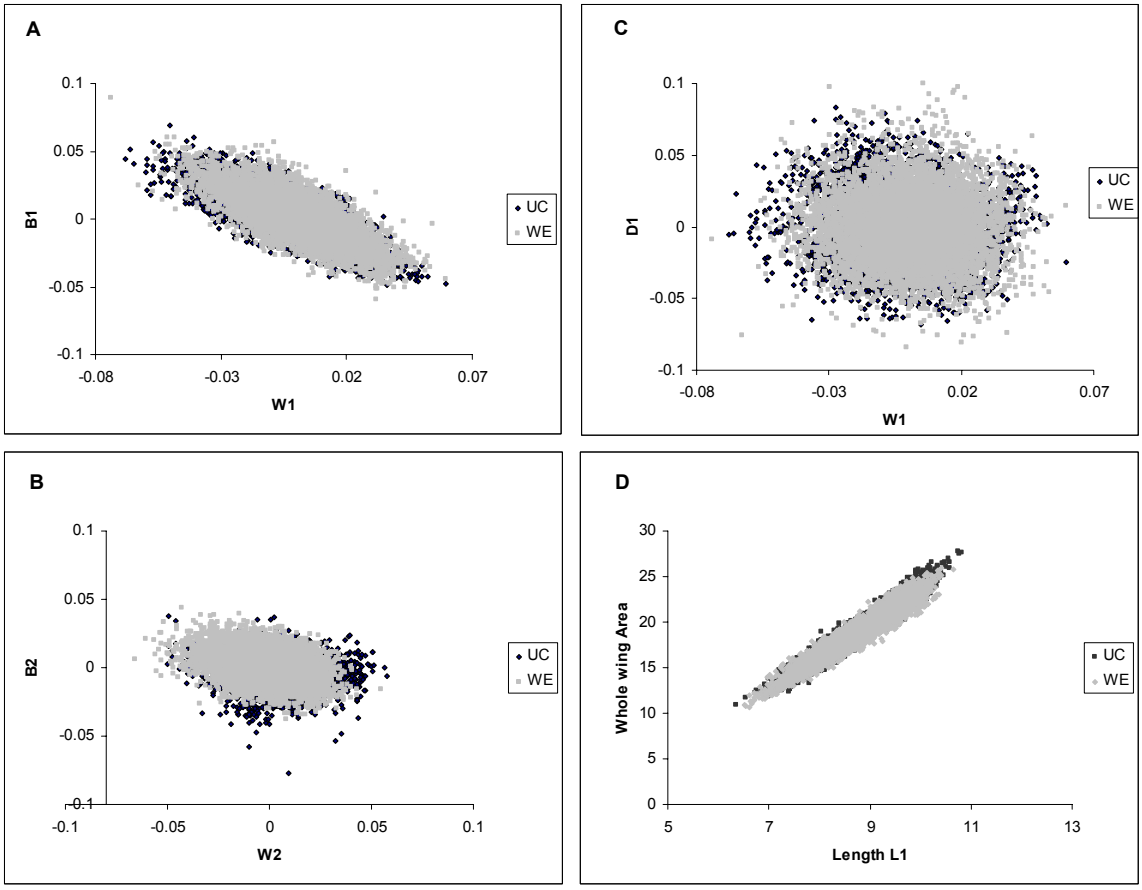


Figure 4.8. Examples of the relationships between pairs of relative warps and size measures graphed by populations (UC black, WE gray). Pairs of relative warps with strong and no phenotypic correlation (Pearson coefficient). A) B1 and W1 (r = - 0.7797, p < 0.0001), B) W5 and B2 (r = - 0.71942, p < 0.0001). C) D1 and W1 (r = -0.06589, p < 0.0001) D) The size measures, wing length (measured as length of L1) and the area of the wing (r = 0.945713, p < 0.0001).

corresponding to elongation of the wing. Also C2 and W8 both capture the width of intervein region C, and are strongly correlated at the genetic level with the size of that region. Another notable feature is the rapid decay of relations that can be seen in pairs of traits. For instance C1 has high genetic correlation to W1, and W1 has similar correlation to B1. But there is only weak genetic correlation between C1 and B1. Finally the results show that size and shape measures show only minor dependence of shape on size. An exception is W1 and B1, which are related to length of vein L1 and presumably the whole wing, in accord with the interpretation of the shape change in these parameters capturing wing elongation. This uncoupling of size and shape agrees with the inferences of Birdsall *et al.* (1999).

Genetic correlations are sensitive to allele frequencies (Falconer & MacKay 1996) and could thus differ between populations. The genetic and phenotypic correlations between traits were estimated for sexes and populations separately. Do these genetic correlations differ? Confidence intervals (CI) on the coefficients of genetic correlation can be used to test for significant deviation from zero (Ungerer *et al.* 2001) by using the z-function (Sokal and Rohlf 1995). I constructed 95% CI's for the genetic correlations calculated by sexes (Appendix E). CI's for correlations with absolute value 0.13 or higher did not surround 0 and are significant. This criterion is met by 165 out of the total 253 pairs of traits. The CI's for correlations calculated for the sexes individually overlapped in all cases. Populations were compared by estimating 95% CI's for genetic correlations within each population. CI's for seven pairs of traits do not overlap (data not shown). Construction of 95% CI on genetic correlations is a liberal test of significance as multiple comparisons are not considered. Genetic correlations will however impact the association tests, as correlated traits will give similar results. For instance in males the genetic correlation between B1 and W1 is 0.998, yielding comparable significance of associations (Figure 4.9 C).

Exploratory analysis of associations by sexes and population

Reduced model testing for the association between SNP's and wing shape were analyzed in Tassel on datasets divided by sex and population, to explore general tendencies of the data. Analysis by sex and population would otherwise result in 4X the tests, resulting in Bonferroni cutoff at $p = 0.0000025$, which no tests survived. The top sites by sex and population for each trait are reported in Table 4.5, and those gave robust signals with more elaborate models (see following section). Four main results emerge from these explorations. First peaks of association drop sharply as can be seen in association profiles for traits D1 and C1 (Figure 4.9), suggesting considerable independence of sites. This is important as we surveyed continuous genotypes, so detection of association in a particular region does not depend on linkage but is a direct assessment of the effect of each polymorphism. Sites close to the boundaries of the sequenced

Table 4.5. Significance of strongest associations for individual traits as tested on data divided by sex and population.

	WE Females			WE Males			UC Females			UC Males		
Trait	Site	F	*P*	Site	F	*P*	Site	F	*P*	Site	F	*P*
B1	38581	4.66	0.73	38039	6.14	0.27	6085	5.81	0.39	42043	5.56	0.36
B2	41601	5.65	0.44	40620	6.48	0.27	30733	3.98	0.88	42367	4.83	0.67
B3	31164	3.44	0.98	42336	5.41	0.49	**31624**	6.00	0.28	**31624**	4.92	0.64
C1	**30200**	10.24	0.00001	**30200**	9.99	0.02	**40149**	5.11	0.56	**40149**	4.52	0.71
C2	36644	5.03	0.61	39196	5.70	0.38	31245	4.72	0.73	42367	4.59	0.8
C3	**30676**	6.32	0.28	**30676**	4.60	0.74	40110	7.03	0.08	5895	5.04	0.6
D1	**39389**	7.81	0.08	40110	5.79	0.44	**39389**	7.75	0.06	**39389**	6.83	0.06
D2	6412	6.26	0.30	39199	6.92	0.16	**40149**	7.38	0.07	**40149**	6.97	0.1
D3	31443	4.91	0.66	40044	4.40	0.80	41154	4.66	0.72	41079	4.13	0.89
W1	**36644**	6.70	0.21	**36644**	7.05	0.05	36214	4.01	0.91	42367	4.41	0.81
W2	30200	6.49	0.24	5510	4.70	0.67	**40149**	6.62	0.22	**40149**	6.65	0.2
W3	39160	6.50	0.24	39300	8.47	0.02	6065	4.36	0.79	42367	4.22	0.77
W4	39389	4.44	0.82	38056	5.14	0.59	**6063**	3.35	0.98	**6063**	4.19	0.84
W5	35955	4.44	0.83	40044	3.74	0.93	42140	3.68	0.96	30565	4.24	0.92
W6	36248	5.70	0.42	41256	5.28	0.55	30403	6.79	0.18	37805	6.86	0.21
W7	**30505**	7.61	0.13	**30505**	9.01	0.04	**39389**	5.35	0.54	**39389**	5.98	0.26
W8	42043	6.85	0.19	39196	6.02	0.35	**6326.3**	5.48	0.48	**6326.3**	5.63	0.34
W9	**38025**	8.77	0.03	**38025**	7.84	0.09	**39912**	6.52	0.24	**39912**	7.28	0.11
T Area	**40722**	7.76	0.12	**37973**	6.37	0.24	30403	5.39	0.46	**41658**	5.06	0.64
B Area	**37973**	6.29	0.28	**37973**	5.14	0.58	**41658**	5.65	0.44	**41658**	4.99	0.63
C Area	**40722**	7.67	0.05	**40722**	6.10	0.32	41670	5.97	0.5	**41658**	6.27	0.48
D Area	41247	6.05	0.37	**37973**	5.86	0.41	**41658**	5.94	0.39	**41658**	6.10	0.34
L1	**40722**	**10.57**	0.01	**40722**	8.51	0.07	41712	6.12	0.29	5510	5.48	0.51

Implemented in Tassel, and significance asserted by permutation of phenotypic data on the allelic matrix for each combination of trait, sex and population separately, and does not correct for tests across traits.
Sites most significant in both sexes within a population are boldfaced

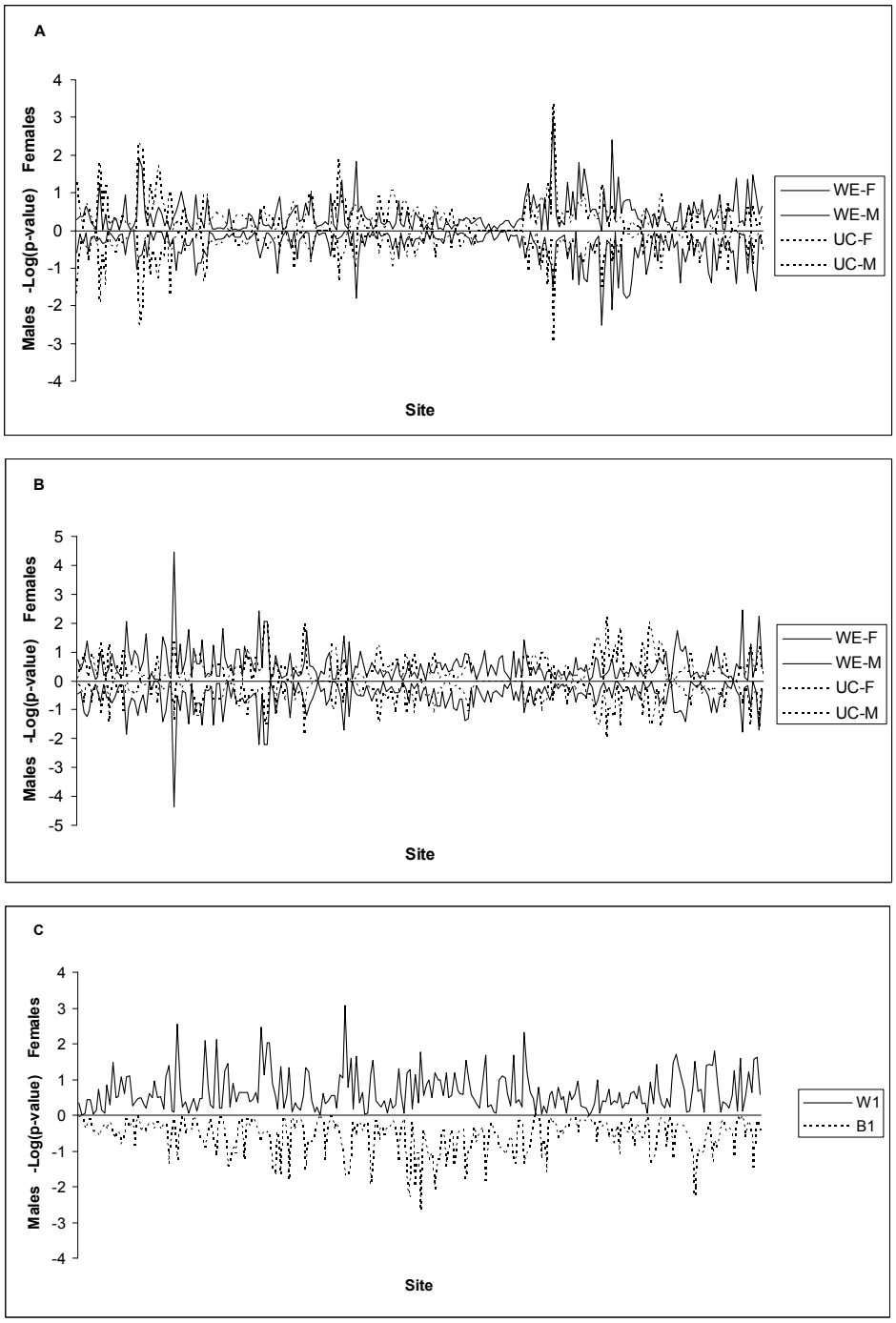


Figure 4.9. Associations between *EGFR* polymorphisms and four wing shape parameters. A) parameter D1 and B) parameter C1. Tested for sexes and populations separately (Solid line: WE, dotted: UC). C) Association profiles for two correlated traits, B1 (below) and W1 (above) in WE lines only. Site on the x-axis represents the sampled portions of the gene and significance of associations (as log transformation of *p*-value) is plotted on the y-axis. Males are below and females on top. Bonferroni cutoffs for 20 traits, 250 polymorphisms, 2 populations and 2 sexes are off the chart ($p = 0.0000025$, negative log transformation 5.6).

region are the obviously exempt. Second, apart from the three correlated pairs of traits, each of the shape traits had distinct profiles of association. Examples of dependent traits are W1 and B1, which are highly correlated (Figure 4.8), and have not surprisingly highly correlated association profiles (Figure 4.9. C). Third, shape parameters yielded stronger associations than any of the 5 size parameters with direct tests of nucleotide polymorphism effects. Finally, tests of association in Tassel with ANOVA and permutations gave nearly identical *p*-values (not shown), indicating that the phenotypes are in good agreement with the assumption of normality.

These reduced exploratory analyses yielded two more findings. The significance of associations was more similar between sexes of the same population, then between the same sex in distinct populations. For instance, for trait W1 the correlation of association test significance between sexes for each population is high (r = 0.91) while the correlation between profiles for female from each population is insignificant (r = 0.04) (Figure 4.10). The relation between populations is strongest for females in D1 (r = 0.48), mainly driven by a few sites that are significant in both populations (Figure 4.11). Secondly, there is an intriguing sex dimorphism, as males from North Carolina gave consistently weaker associations than WE females. See for instance the poor signal of WE males to trait D1 (part A Figure 4.9). These results suggest manifestation of variation in *EGFR* depends upon sex and factors segregating in the two populations, leading us to test more complex models.

Tests of association by Analysis of Variance

Analysis of variance implemented in SAS allows delineation of higher order interactions between sex, populations and polymorphisms on phenotypes. The ANOVA procedures in SAS do not accommodate permutation based evaluation of significance for complex models. We (N. Nikoh, I. Dworkin, A. Palsson and G. Gibson, data not shown) found that mixed model ANOVA with line nested within SNP and population gave very comparable results to permutations respectful of the sex and population identity. That is, values are randomized within each class, for instance West End females, to retain structure of the original data. The overall patterns of association are consistent with the reduced models implemented in Tassel. Uncoupling of tests by sites and traits is apparent, and as are differences in SNP effect dependence on gender and population. The association profiles for these terms as tested against D1 illustrate this clearly (Figure 4.12 and Table 4.6), for instance site G42377T shows most significant interaction with sex but no significance for other terms. Again consistent with the Tassel results, significance profiles differed substantially by traits. In Figure 4.13 the first profile (A) shows the significance of the SNP term when tested against C1, with site C30200T being the single highest (see below). Similarly, trait B3 returns just noise, with no genetic contribution of the *EGFR* polymorphisms. Finally, the LD in and around intron three generates a unique plateau-like

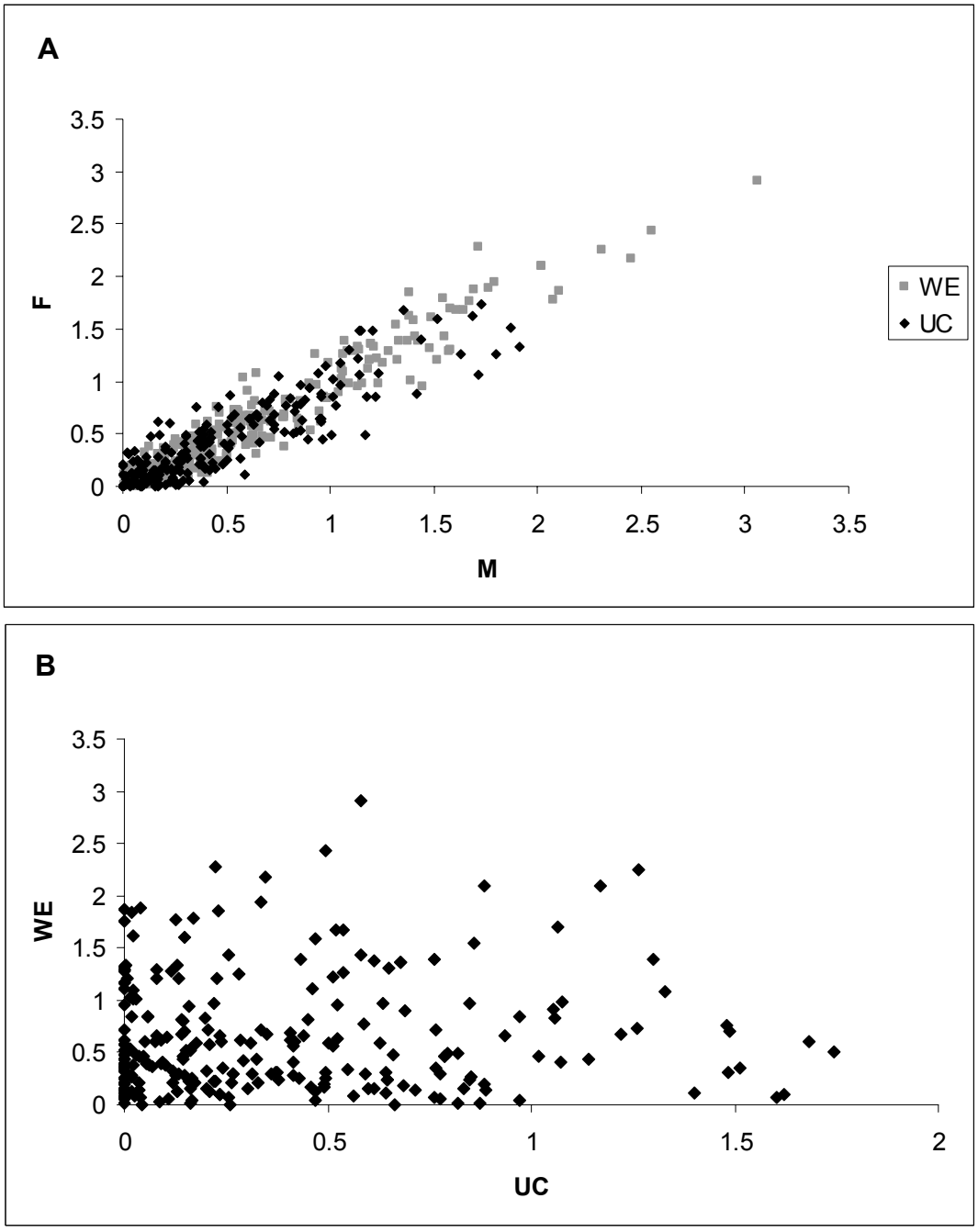


Figure 4.10. Relationship of significance of associations (Negative log of p-values) by sex and population for shape parameter W1. A) Estimated for sexes separately in each population, UC (black) Pearson correlation between sexes $r = 0.91$ $p < 0.0001$ and WE (gray) $r = 0.96$ $p < 0.0001$. B) Correspondance between populations for females ($r = 0.04$, not-significant).

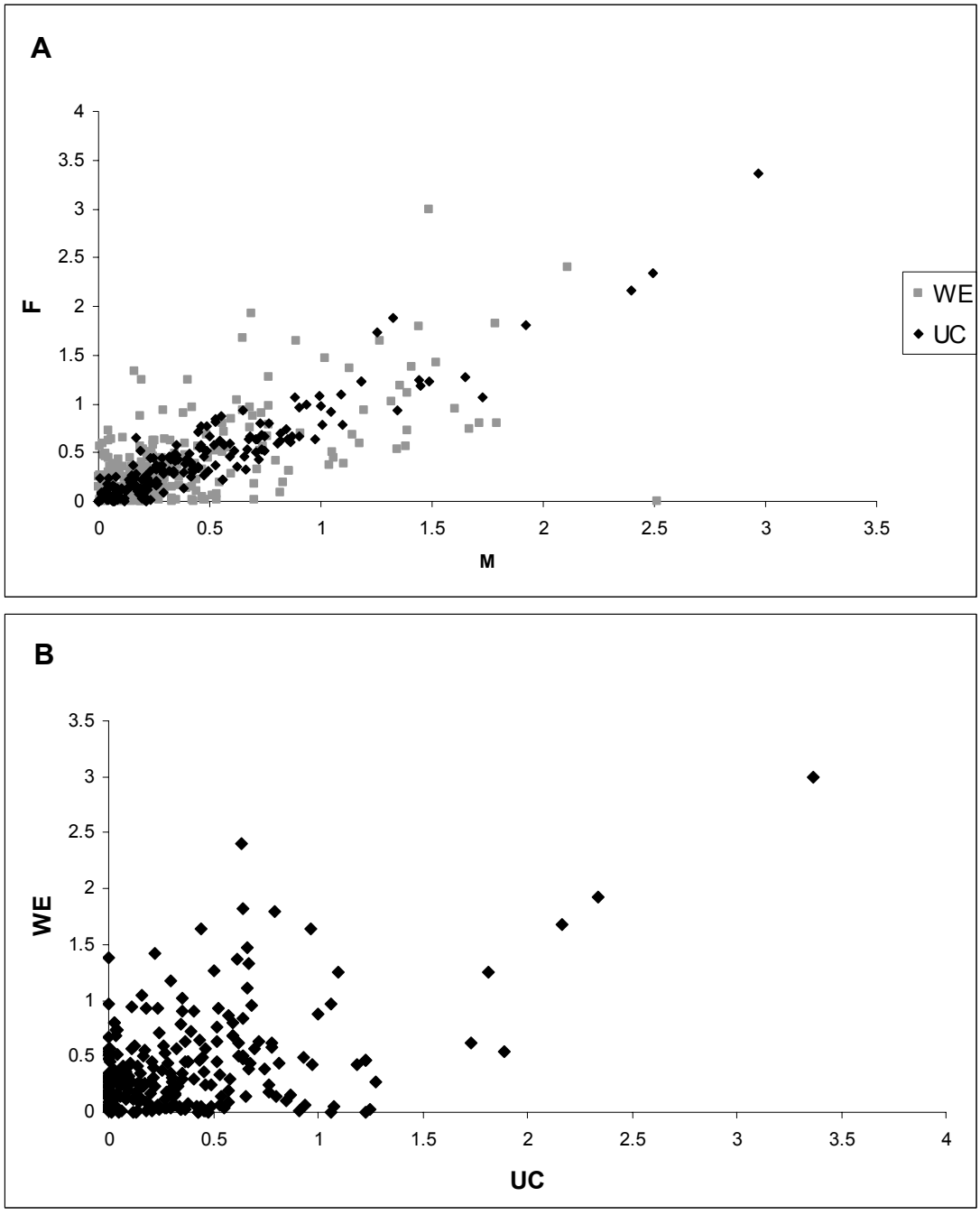


Figure 4.11. Relationship of significance of associations (Negative log of p-values) by sex and population for shape parameter D1. A) Estimated for sexes separately in each population, UC (black) Pearson correlation between sexes $r = 0.94$ $p < 0.0001$ and WE (gray) $r = 0.65$ $p < 0.0001$. B) Correspondance between populations for females ($r = 0.48$, $p < 0.0001$).

Table 4.6. ANOVA tables for trait D1 accompanying Figures 4.12 and 4.17.

Source	Site	F	*P*	Site	F	*P*
Pop	T39389C	0.12	0.72465	T40110C	0.79	0.37435
SNP		16.93	**0.00006**		11.67	**0.00078**
Sex		343.56	0.00000		429.45	0.00000
Pop*SNP		3.31	0.07050		0.00	0.94450
SNP*Sex		0.06	0.80087		0.96	0.32931
Pop*Sex		2.00	0.15944		0.64	0.42318
Pop*SNP*Sex		0.00	0.96598		0.33	0.56562
Pop	G6065T	1.33	0.25012	6218del13	6.96	0.00903
SNP		9.60	**0.00224**		1.23	0.26955
Sex		193.84	0.00000		335.13	0.00000
Pop*SNP		3.99	**0.04707**		10.77	**0.00123**
SNP*Sex		1.01	0.31644		3.90	**0.04969**
Pop*Sex		0.17	0.68464		1.27	0.26076
Pop*SNP*Sex		0.02	0.89299		0.41	0.52077
Pop	G42377A	0.09	0.76786			
SNP		0.15	0.70352			
Sex		170.99	0.00000			
Pop*SNP		0.01	0.92680			
SNP*Sex		9.36	**0.00258**			
Pop*Sex		2.35	0.12698			
Pop*SNP*Sex		5.00	**0.02663**			

The significance cutoff for a single trait is $p = 0.0002$, so only T39389C is significant.
Genetic terms are highlighted if $p < 0.05$.
The nominator degrees of freedom where higher than 180 for all tests.

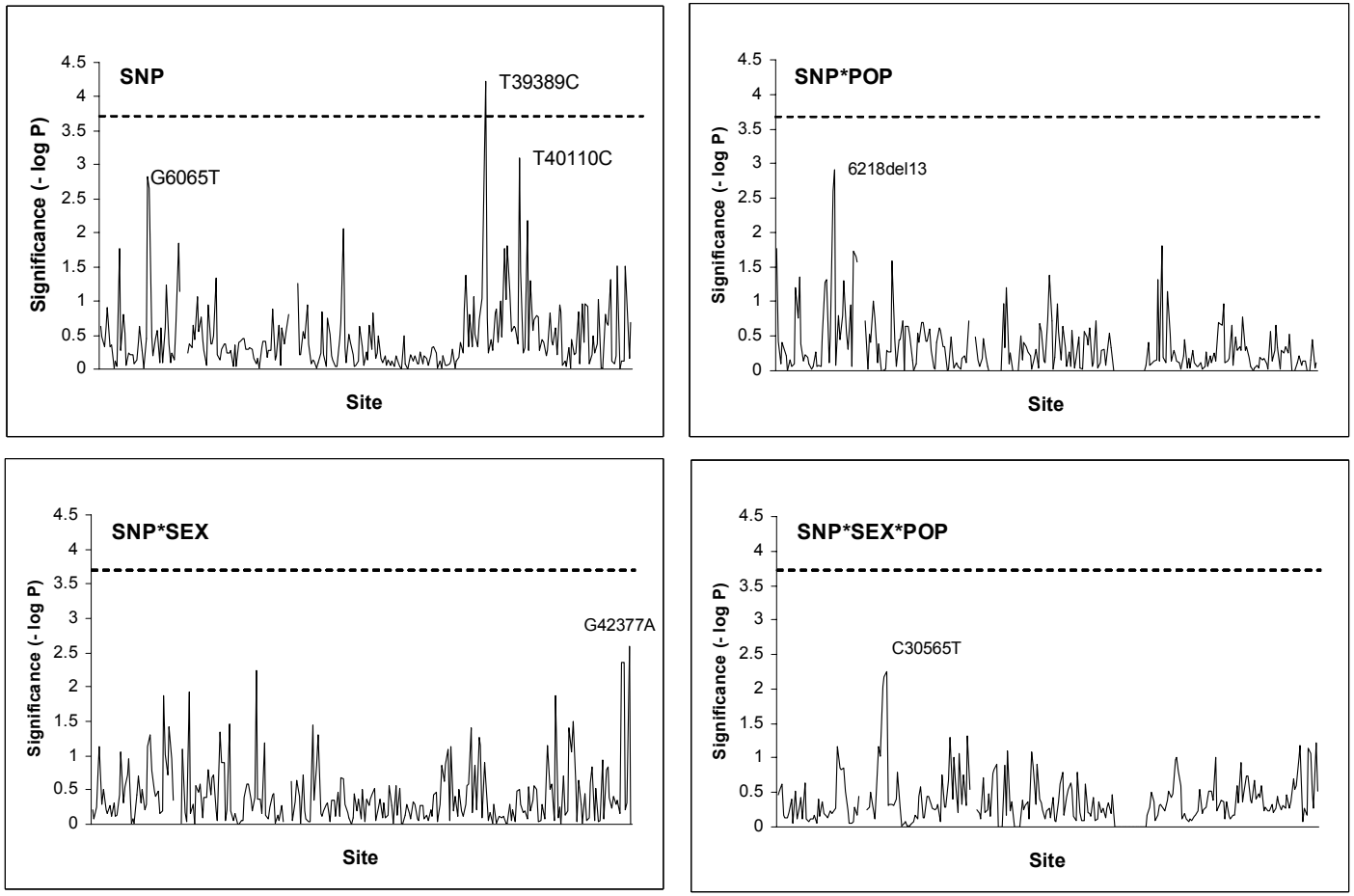


Figure 4.12. Significance of tests of association between polymorphisms in *EGFR* and wing shape parameter D1. The panel represent the 4 terms testing for the genotypic effect (SNP). The solid line represents Bonferroni cutoff for tests on a single trait. The experiment wide cutoff is at $p = 0.00001$. Significance reported as negative log transformation of *p*-values.

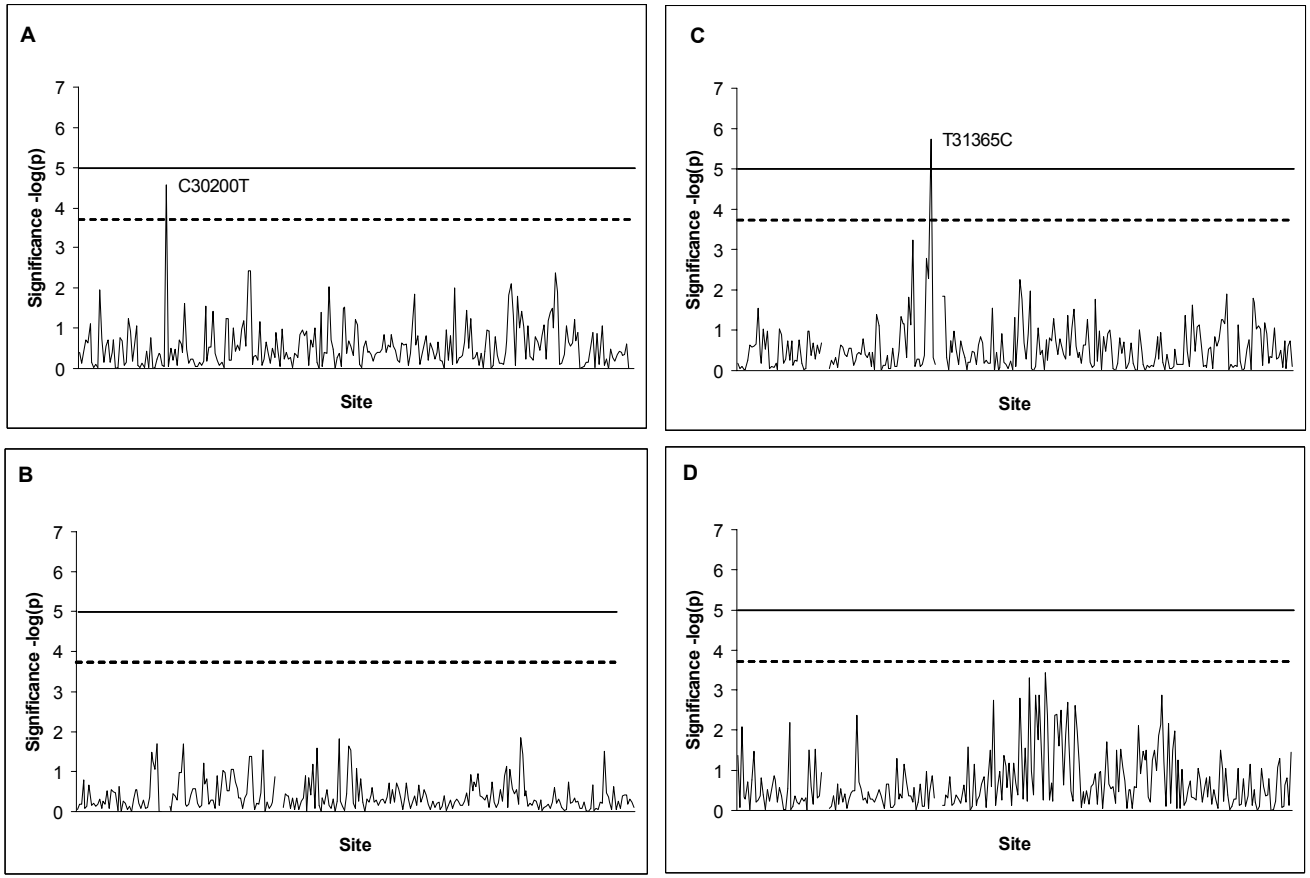


Figure 4.13. Examples of associations along *EGFR*. The lines represent the three multiple correction thresholds, experiment wide Bonferroni (cutoff p = 0.00001, top solid line) and single trait Bonferroni (cutoff p = 0.0002, lower dashed line). A). The significance of polymorphisms on trait C1, B) on trait B3, C) a Sex by SNP interaction for the total area (T Area) and D) associations of SNP's to W9. Significance reported as negative log transformation of *p*-values.

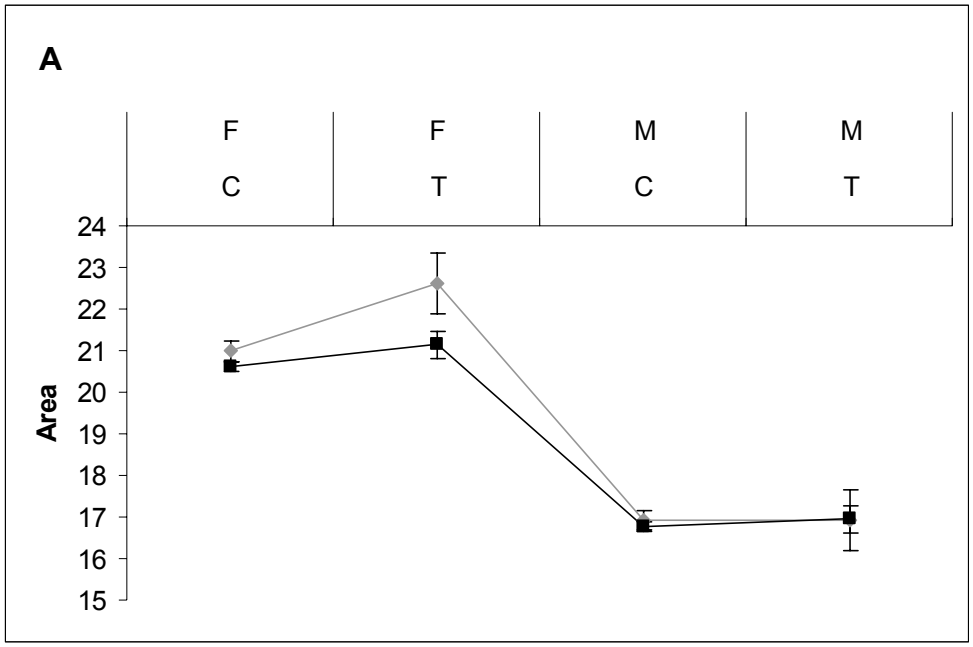
pattern of associations to W9 where no single site leaps out as most significant. Again the general result is the relative independence of sites, as high associations rarely extend beyond a handful of sites.

Correcting for multiple tests

Significance of associations can be evaluated at 3 levels; the significance of individual tests, the significance of tests for a given trait, and finally the whole experiment wide significance over all sites and traits. Multiple genetic terms in the complex ANOVA models are considered model fitting. Individual tests of association are judged against the probability of rejecting the null hypothesis when true 5% of the time (α). We must adjust α for the 250 sites tested against each trait. The most stringent correction is to divide by the number of tests, yielding the Bonferroni cutoff at $p = 0.0002$ per trait (negative log = 3.7). Experiment wide cutoffs, corresponding to 250 sites and 20 traits, give Bonferroni $p = 0.00001$ (negative log = 5). With the high correlation between size measures I only tested *EGFR* polymorphisms against area of the whole wing (T-Area) and length of L1. Application of Dunn-Sidak or stepwise corrections did not alter the results. Only one site gave a *p*-value lower then the experiment wide cutoff and is formally significant in the whole experiment. Seven *p*-values lower than the cutoff for individual traits are only suggestive and not formally significant.

EGFR and wing size

Size and shape of the *D. melanogaster* wing are largely unrelated and size was assumed to be a control phenotype for tests of association with *EGFR*. This rests on the assumptions that the major role of *EGFR* in wing development is patterning not growth and that polymorphisms in *EGFR* will affect shape more strongly then size. The results challenge these assumptions. While no polymorphisms are significantly associated with size measures when tested individually, one site shows dependence on sex in relation to wing size represented by total area (T Area). The association to site T31365C shows experiment wide significant sex dependence ($p = 1.85 \times 10^{-6}$) and a three-way interaction ($p = 0.001$) (Figures 4.13 C and 4.14). Inspection of graphs suggests the effects are unique to females and more pronounced in the Californian population. Site T31365C is located 500 bp downstream of exon 2, at a boundary between highly variable and less polymorphic regions. Two other sites are suggestive, significant if considered size in isolation, but not on an experiment wide basis. Those are T40722C and G30401A which also show population and population by sex dependence (Table 4.7 and Figure 4.15). Site T40722C is a silent site in exon 6, the RTK domain and similarly G30401A is in the promoter of exon 2 and shows three way interaction with the strongest effect observed in UC females.



B

Source	F	*P*
Pop	1.45	0.230558
SNP	2.06	0.154439
Sex	2044.02	9.47E-70
Pop*SNP	0.31	0.579016
SNP*Sex	25.59	**1.85E-06**
Pop*Sex	18.72	3.52E-05
Pop*SNP*Sex	11.28	**0.001101**

Figure 4.14. Site T31656C affecting area of the whole wing (parameter T-area), shows sex and population dependence A), as detailed in the ANOVA table B). Degrees of freedom are 1,104 except for interaction terms where the denominator *df* are reduced by 1. Least square line means (on Y-axis) with standard errors are graphed for each genotype by sex (on X-axis) and population configuration. Gray line indicates UC and black WE.

Table 4.7. ANOVA tables for the sites most strongly associated with wing parameters.

Source	Trait	Site	F	*P*	Trait	Site	F	*P*
Pop	T Area	T31656C	1.45	0.230558	C1	C30200T	5.58	0.01983
SNP			2.06	0.154439			19.06	**0.00003**
Sex			2044.02	0.000000			17.38	0.00006
Pop*SNP			0.31	0.579016			0.28	0.59553
SNP*Sex			25.59	**0.000002**			0.22	0.63793
Pop*Sex			18.72	0.000035			0.98	0.32434
Pop*SNP*Sex			11.284	**0.001101**			0.16	0.69437
Pop	C2	C31634T	0.24	0.62325	T Area	T40722C	14.65	0.00018
SNP			0.03	0.86024			0.15	0.69755
Sex			317.09	0.00000			6287.36	0.00000
Pop*SNP			0.39	0.53172			16.97	**0.00006**
SNP*Sex			8.94	**0.00370**			0.12	0.73318
Pop*Sex			10.20	0.00201			33.23	0.00000
Pop*SNP*Sex			18.35	**0.00005**			4.49	**0.03545**
Pop	D1	T39389C	0.12	0.72465	W9	5683del1	0.05	0.81544
SNP			16.93	**0.00006**			1.93	0.16665
Sex			343.56	0.00000			135.39	0.00000
Pop*SNP			3.31	0.07050			0.26	0.60745
SNP*Sex			0.06	0.80087			16.58	**0.00007**
Pop*Sex			2.00	0.15944			1.30	0.25519
Pop*SNP*Sex			0.00	0.96598			3.81	0.05244
Pop	T Area	G30401A	0.19	0.66415	W7	C30505A	0.08	0.7808
SNP			2.44	0.12155			15.12	**0.0002**
Sex			4940.99	0.00000			0.00	0.9869
Pop*SNP			4.48	**0.03690**				
SNP*Sex			4.29	**0.04093**			1.05	0.3083
Pop*Sex			15.50	0.00015			0.01	0.9377
Pop*SNP*Sex			16.48	**0.00010**				

Bonferroni cut-off ($p = 0.0002$) on trait basis, but experiment wide basis ($p = 0.00001$).
Terms including SNP are highlighted if p-value is lower than 0.05.
The nominator degrees of freedom where higher than 170 for all sites except C30505A (*df* 1,95), G30401A (*df* 1,98), C31656T (*df* 1,104) and C30200T (*df* 1,117).

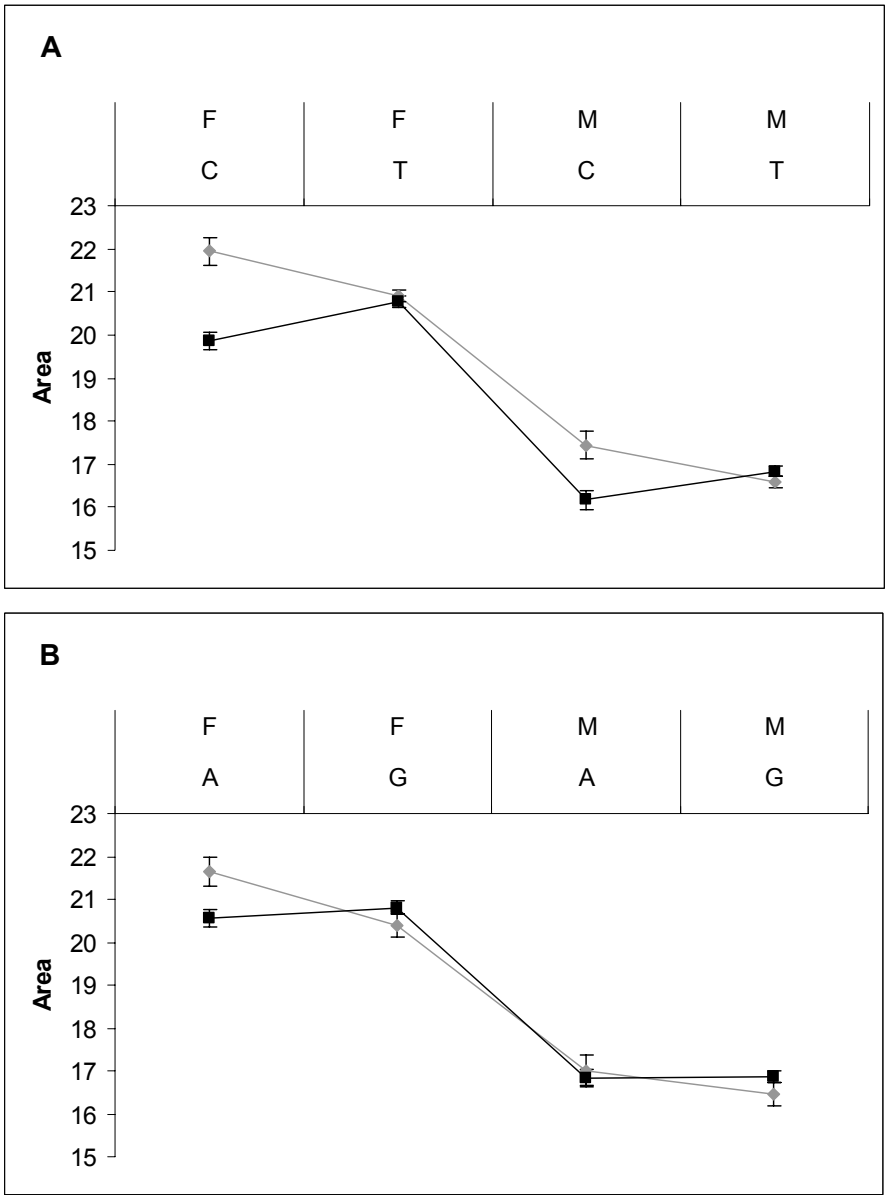


Figure 4.15. Two polymorphisms showing significant interaction with size. A) Site T40722C in exon 6 yields a SNP by population term of $p = 0.00006$. B) G30401A in the promoter of exon 2 has three way interaction with population and sex at $p = 0.0001$. See Table 4.7 for full ANOVA's. Least square line means (on Y-axis) with standard errors are graphed for each genotype by sex (on X-axis) and population configuration. Gray line indicates UC and black WE.

Suggestive associations for shape

No tests of associations to shape survive Bonferroni correction accounting for 250 polymorphisms and 20 traits. Mining for borderline signals risks identifying false positives and must be interpreted suggestive at best. For instance when correcting for multiple tests by individual traits, instead of all 20 traits, five sites with presumed effects on shape subsist. The strongest site is C30200T, related to intervein warp C1 (p = 2.71 x 10^{-5}) with no significant interaction terms (seen clearly on Figure 4.16 A). C30200T is also associated to W1 and W2, at 0.005 level and D2 at 0.05. Recall that W1 and W2 are orthogonal by principles of procedure, yet they are both strongly correlated to C1, at the phenotypic and genetic level (Table 4.4). The common shape change is the relative placement of crossveins (Figures 4.2-5) with the derived T causing elongation of the distance between the crossveins. C30200T is the only polymorphism at high frequency in a presumed GAGA factor binding motif. A derived T disrupting a conserved CN pattern has risen to 0.75 in frequency in the two North American populations while still being at 50% in a Kenyan sample (Figure 3.4 in Chapter 3). The site is in linkage equilibrium with sites downstream, and sites in the first exon (data not shown) but as C30200T is on the edge of a genotyped region then the possibility of the effect being caused by a linked upstream site cannot be excluded.

The four other borderline sites are C31634T which affects C2 in a sex and population dependent manner (p = 0.00005), T39389C with p = 0.00006 in tests against trait D1 (Figures 4.13. and 4.16), single base deletion 5683del1 whose affects on W9 are conditioned on sex and site (p = 0.00007) and finally C30505A impacts W7 (p = 0.0002) (Figures 4.16, and 4.17). ANOVAs for those sites are summarized in Table 4.7. Site C31634T is in the vicinity of exon 2, placed 800 bp into intron 2. It affects parameter C2 in a sex and population dependent manner, with UC females apparently giving the signal (Figure 4.16 B). It is located on the upstream boundary of a conserved region which starts close to site T31365C (that affects size, see above) and affects the width of the central region of the wing. T39389C affects D1 with a clear SNP term but also shows dependence on population (Figure 4.17). The third most significant site within *EGFR* affecting shape is T39389C, and is located in exon 5 but does not alter the protein sequence. This site is being driven by association in both populations as it showed up in the individual analysis in Tassel (Table 4.5), and the population by SNP term is not significant. The weakest of these marginal signals, C30505A, was only found segregating in WE hence the lack of population term in the model. But since only 10 alleles were sampled in UC we can not confirm the site's absence from the UC population, particularly as it also segregates in the Kenyan sample. C30505A is located within the 5'-untranslated region of the RB transcript, 13 bp away from the start codon. All of those sites show distinct effects of character states as can be

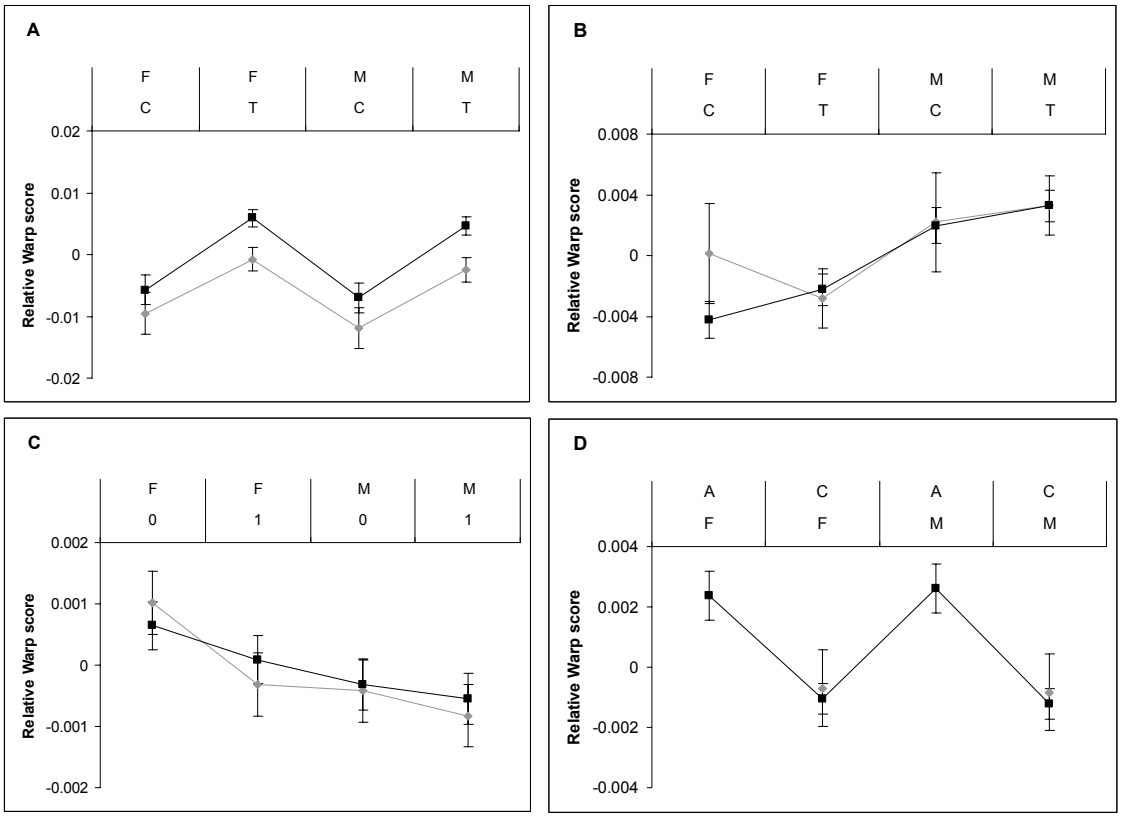


Figure 4.16. Four sites giving significant polymorphism effect on aspects of shape, when correcting for tests on each trait alone (cutoff $p = 0.0002$) but not on experiment wide basis (cutoff $p = 0.00001$). A) Site C30200T impacts C1 with $p = 0.000027$. B) Site C31634T affects C2 in sex and population dependent manner $p = 0.00005$. C) Indel 5683del1 has SNP by sex effects with $p = 0.00007$ on trait W9. D) Trait W7 is affected by site C30505A ($p = 0.0002$). See ANOVAs in Table 4.7. Least square line means (on Y-axis) with standard errors are graphed for each genotype by sex (on X-axis) and population configuration. Gray line indicates UC and black WE.

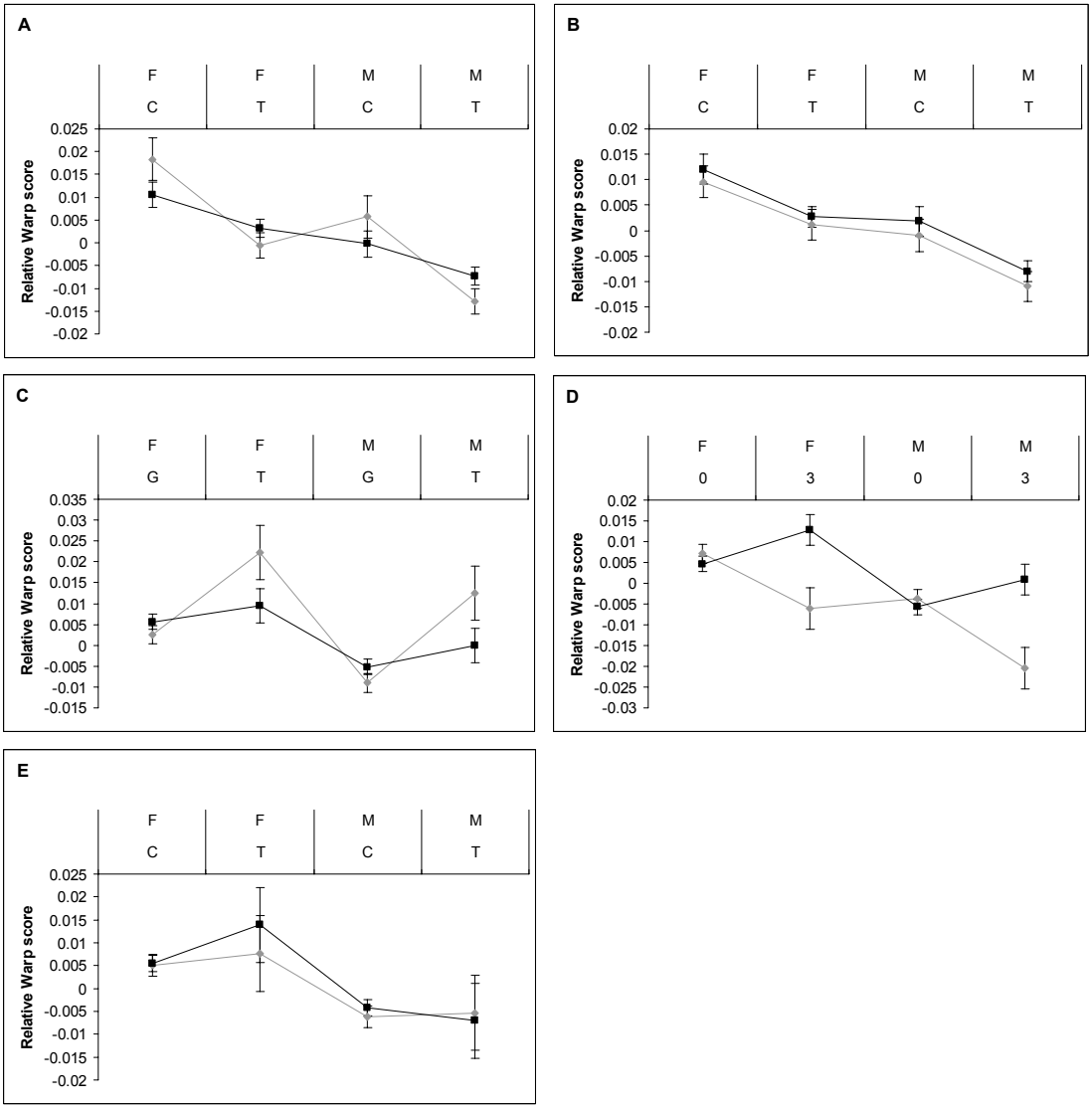


Figure 4.17. Example of associations of *EGFR* polymorphisms and shape. The five smallest *p*-values for trait D1, also graphed in Figure 4.13 and ANOVAs tabulated in Table 4.7. A) Only one site T39389C (SNP term, p = 0.00006) is significant after corrections for multiple tests on this trait alone (Bonferroni cutoff = 0.0002). It does not pass experiment wide correction (p = 0.00001). Other sites are not significant when correcting for multiple tests to this trait. Two have low SNP term , B) site T40110C (p=0.00078), C) G6065T p=0.00224. D) Indel 6218del13 has a interaction by population (*p* = 0.00123) E) site G42378A has interaction with sex (*p* = 0.0025) and a three-way interaction (SNP*Pop*Sex) at *p* = 0.02. Least square line means (on Y-axis) with standard errors are graphed for each genotype by sex (on X-axis) and population configuration. Gray line indicates UC and black WE.

seen in Figure 4.16 and 4.17. The distribution of line-means graphed by genotypic classes also represents this see for example T39389C and D1 in Figure 4.18.

All sites detailed so far are either synonymous or non-coding. The strongest replacement involved a G6065T polymorphism and D1 (p = 0.00224) (Figures 4.12 and 4.17, Table 4.6). It changes serine 17 to isoleucine in the signal peptide for isoform RA and also affects 3 other parameters more weakly (Appendices H and I). The effect are more pronounced in the UC's, documented in the population by SNP term (p = 0.047). Interestingly this was also the only replacement found significantly associated with variation in eye-development (I. Dworkin, K. Birdsall, A. Palsson and G. Gibson submitted, data not shown). Other sites with associations to eye-roughness did not yield a signal to wing shape. It must be stressed that those sites reported are not significant when considering number of tests experiment wide.

Interactions between population and SNP

Two questions are addressed here. Can we detect dependence of SNP effects on genetic backgrounds, sex or population? Do sites known to differ in frequency between populations create sporadic associations? Contrary to the sex term, population was not significant for the majority of sites and traits tested. The largest difference between the populations is 10% of difference seen between sexes, and in the same range as the effects detectable by the overall association tests. Inclusion of a population term in higher order models should account for these effects. Significant interaction of polymorphism to population suggests a genetic conditionality of allelic effects. The significance can either be caused by stochastic sampling of alleles, as the depth of sampling from the two populations differs between sites, or it could be a true signal. The dataset offers insight into our ability to test for population dependence. Four of the 16 traits had significantly different phenotypic distributions (B2, C1, W2 and W5). The question is does this population term translate into artificial SNP associations? 75 ANOVA models gave significant population and SNP terms at the 0.05 level and 45 of those involve the four traits known to differ between populations. It is worth asking if this discussion would take place if the interaction term discussed was sex, and interaction term would be interpreted as reflecting true genetic background dependence of the polymorphisms.

The second issue involves the six sites with highest F_{ST} values between the North American populations (Chapter 3). Do they give excess of associations for the 4 shape parameters differing between populations? These sites give only 3 associations at 0.05 level with only one to a parameter with population difference (B2), making generalizations impossible. These results suggest differences in allele frequency and phenotypes between populations in this study are not large enough to generate artificial associations.

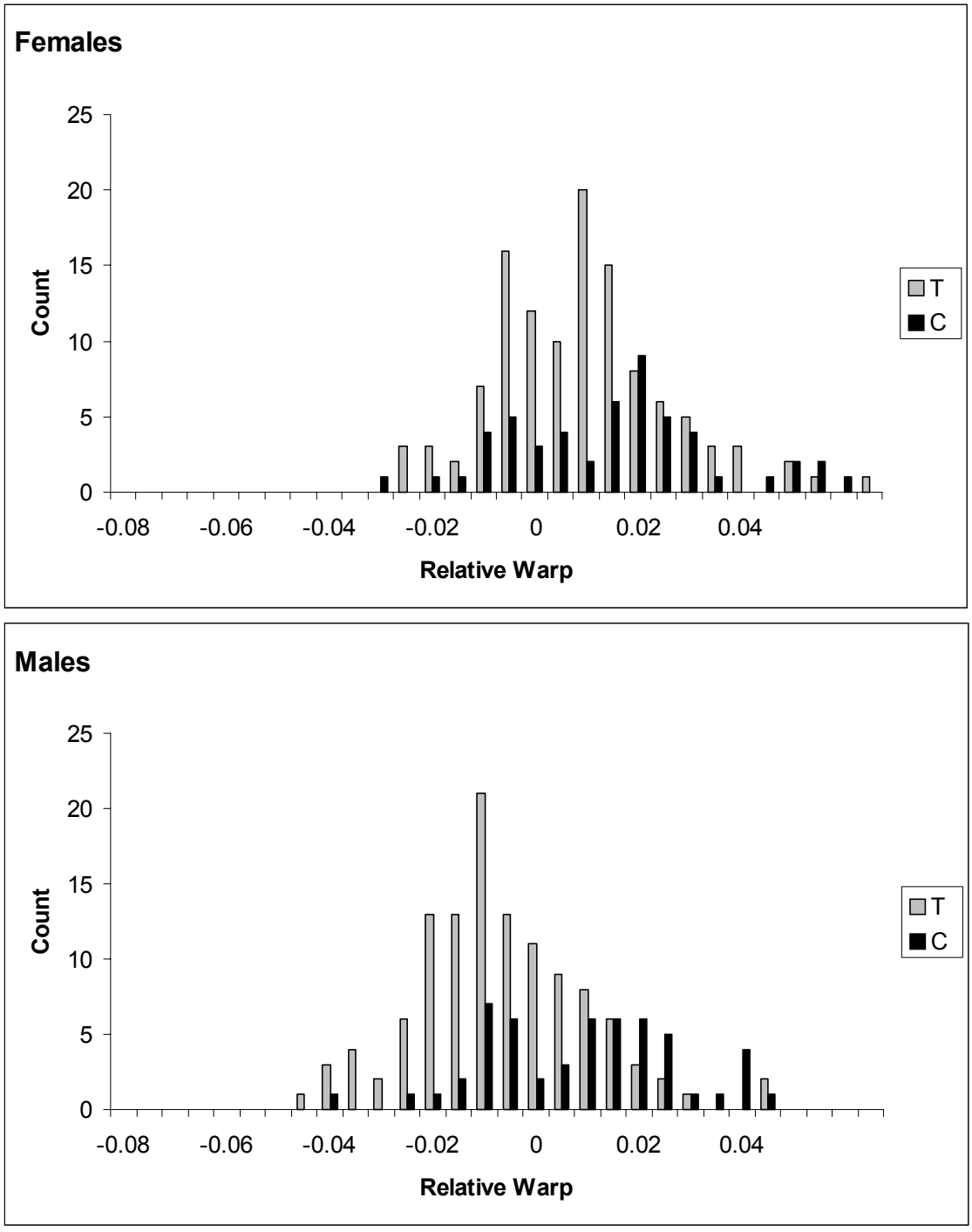


Figure 4.18. The distribution of line means for D1 as categorized by site T39389C. The distributions are plotted separately by sexes, females above and males below. Lines with T are graphed in gray and C in black.

Rare potentially deleterious variants

According to evolutionary theory deleterious variants are expected to segregate at low frequency. Those are out of reach for classical tests of association as implemented above, that are restricted to sites at a frequency higher than 0.05. Several indicator variables were constructed on the basis of the molecular nature of mutations, pooling 11 rare replacements, 4 deletions in coding and non-translated parts of transcript, and large inserts around exon 1. The results of test of these 3 indicator variables, and a compound variable incorporating all types of rare variants are portrayed in Figure 4.19. The rare replacements are associated at the 0.05 level with parameters B2, W3 and W9, while the deletions give a nominal signal to W2. Similarly presence of a pogo transposon and other large insertions around exon 1 affects W7 at the same level. The compound variable also shows a signal to B2, W3 and W9. Naturally these associations would not survive corrections of multiple tests over the whole experiment and must be considered indicative. Pritchard (2001) modeled the effects of rare variants on disease susceptibility and predicted the unavoidable allelic heterogeneity would complicate classical mapping procedures. The test conducted here does not depend on linkage between marker and contributing sites but has little statistical power, particularly if implemented as part of comparable association study.

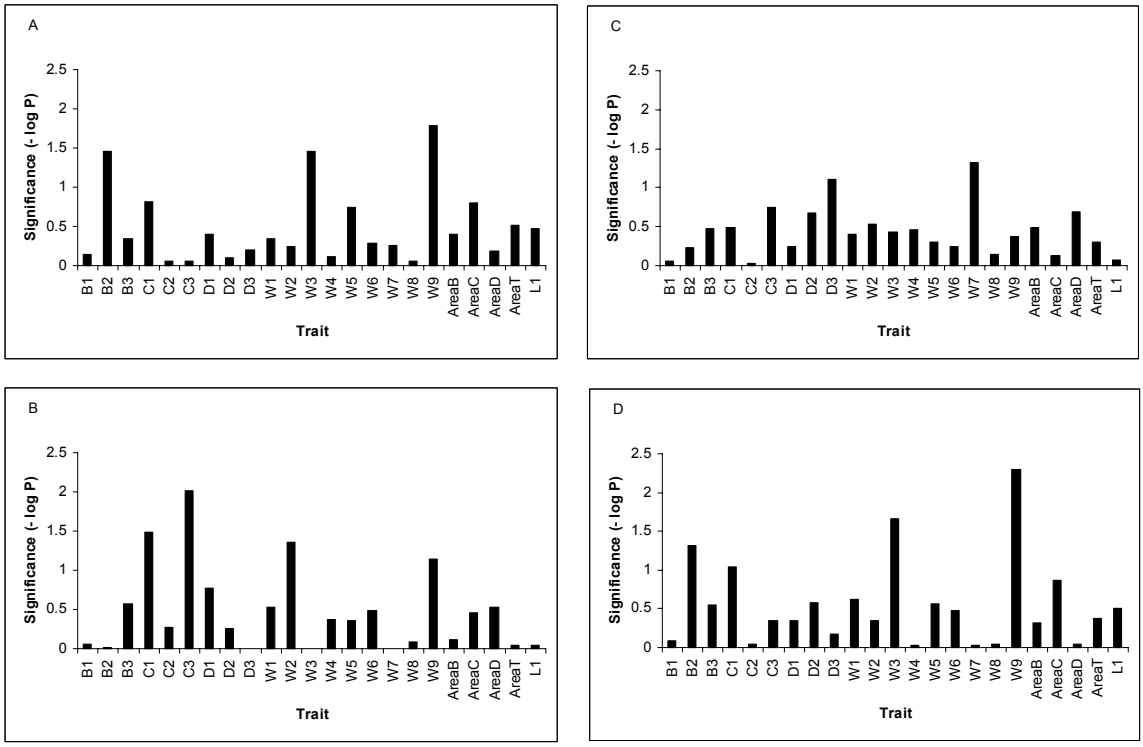


Figure 4.19. Significance of the effects of compound indicators of rare alleles in *EGFR* on shape and size the wing. A) Testing all 11 rare replacements, B) deletions in transcript and C) large introns around exon 1. D) All compound indicators combined in a single test. Significance is reported as the negative log of the *p*-value of the SNP term in ANOVA.

Repeatability of associations

Currently the best way to confirm the effects of quantitative trait nucleotides implicated by association tests is to repeat a study. Two separate experiments were preformed to address repeatability of associations between polymorphisms in *EGFR* and wing parameters. I retested site 31365, that passed experiment wide Bonferroni correction and seven sites with suggestive signals discussed above. Experiment 2 involved randomized recrosses among the WE line to generate F1 lines reconstituting the homozygous state and generating heterozygotes for the sites of interest. In experiment 3, a completely independent sample of alleles was substituted into a common background and tested in heterozygous condition over wild-type and mutant chromosomes. First I investigated the phenotypic dimension of the two datasets.

Phenotypes of Round robin and Kenyan test cross

We have already established that shape metrics are comparable between the inbred experiment and other datasets (Last column of Table 4.1). However it is possible that lingering differences will interfere with tests of association. This is not the case as exemplified by association profiles for trait W1 (Figure 4.20). Axes of shape variation of the original study and the replicates are therefore very similar, justifying comparison. Also the phenotypic range of the two follow up experiments overlaps the range of the inbred lines, as can be seen in Appendices J and K. The test cross design, with the Kenyan chromosomes substituted into Samarkand background, does not enable a test for allelism at *EGFR*, as a gain of function alleles was used. All traits have significant Line effect confirming heritable variation in the sample and 20 out of 23 traits show significant Cross x Line interaction (Appendix L). Of the traits retested only D1 does not show an interaction effect. Again, significant Cross x Line effect does in this case not explicitly mean allelism at *EGFR* (MacKay and Fry 1996, Chapter 2) as it was not tested over a deletion of the locus. It could be caused by differential suppression of the E1 gain of function allele, epistatic interactions, allelism of other loci on the tester chromosome (for instance *blistered*), or genetic backgrounds effects (detailed discussion in Chapter 2).

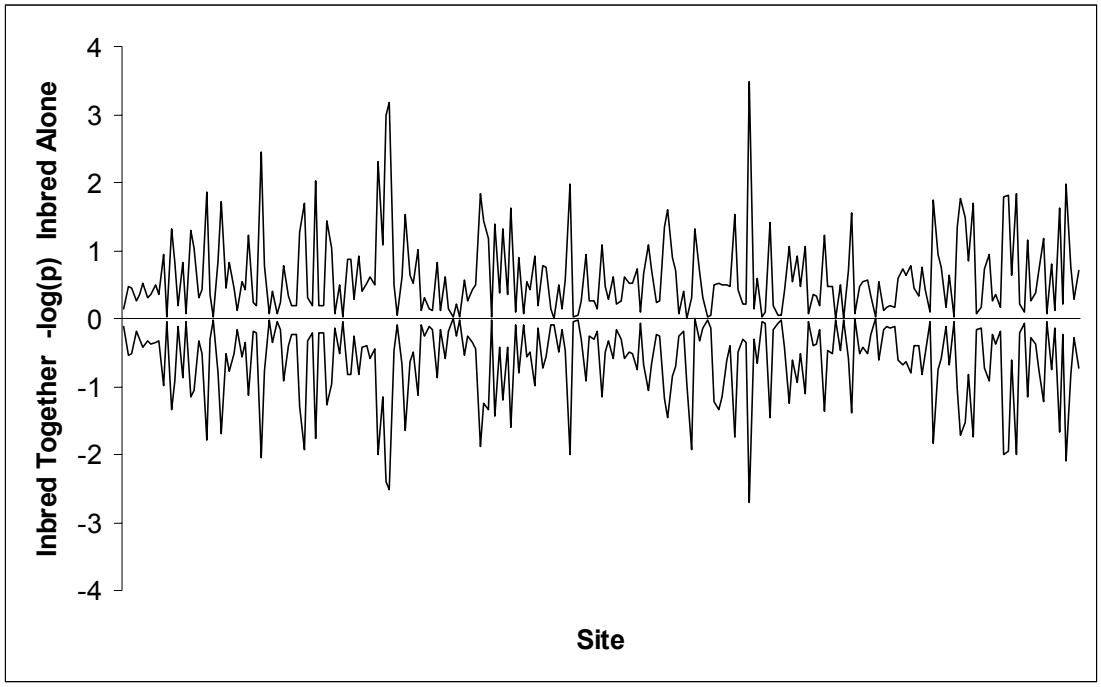


Figure 4.20. The effects of calculating warps for inbred specimens separately (above) or as part of the whole set of wings (below). Significance of associations to trait W1 along *EGFR* depicted as before.

Associations in Round robin and Kenyan test cross

One of the eight sites survives the retesting procedure in both experiments (Table 4.8, Figure 4.21, Appendix M). Interestingly, site T31365C, which showedthe most significant original association does not have replicable effects on size. Site C30200T impacts C1 significantly in all experiments in the same direction (Table 4.8). Site C30505A and trait W7 produce a more complicated signal. It has the same significant effects in the recrossing among the WE lines, but only approaches significance and has opposite effects in the Kenyan test cross (Figure 4.21, see Appendix M for full ANOVA tables for Kenyan test cross). None of the other top sites are significant. Three of those could not be tested in the Kenyan test cross, because markers were not scored, not segregating or rare (<5%).

These results show C30200T has replicable effects on wing morphology, the relative distance between the crossveins. The effect is strongest in the inbred experiment, with the estimated difference between allelic classes being 0.01 relative warp units. The crosses in the Round robin experiment generated all allelic classes at C30200T, but due to the low frequency of the C only four lines are C/C homozygotes. This prevented accurate assessment of dominance. The difference between T/T homozygotes and the heterozygotes is 0.0059 (Round robin). Intriguingly this is close to the same difference as estimated between the two homozygotes in the Kenyan test cross (0.0060) (Table 4.9). The effects proportional to standard deviations of C1 show the same pattern, 0.82 in the inbreds and 0.53 and 0.50 in the follow up. Pseudoreplication was a potential issue for both follow up experiments, especiallyin the Round Robin experiment, as each line provided a total of 3 genomes (mated three times as females and three times as males). This was addressed by testing for associations by the 3 experimental blocks, where each line is only represented by a single female and a single male. The results are reported in Table 4.9 and the effects of C30200T are consistent in all cases, approaching and achieving significance in one case. These analyses are described as reduced model in Table 4.9 while they truly are not. The experimental block term was not included in the full model because it was never significant. Similarly, reduced models were run for the test-cross, analyzing the associations for each test cross individually. The results demonstrate the clear effects of polymorphism C30200T as the direction of effects is consistent and all *p*-values are below 0.1 and one below 0.05 (Table 4.9).

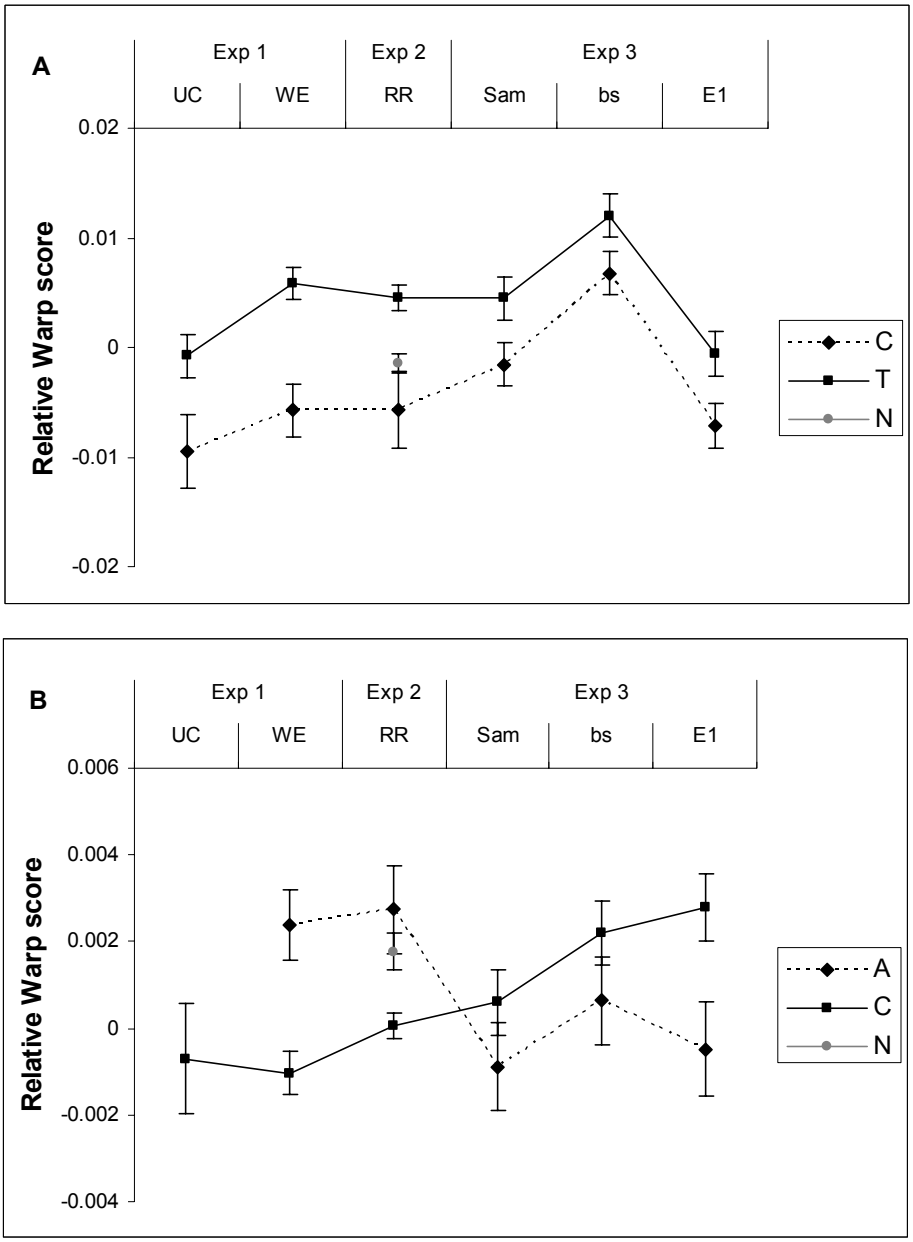


Figure 4.21. The effects of polymorphism in *EGFR* on wing shape the three experiments. A) Site C30200T affecting C1 and B) site C30505A affecting W7. See Table 4. 8 for *p*-values. Least square line means of relative warp units (on Y-axis) with standard errors are graphed by genotype (see legends) for females. The X-axis designates the experiment, 1 the inbred populations UC and WE. Experiment 2, round robin crosses of 71 WE lines and experiment 3, test crosses of Kenyan alleles to three chromosomes, Samarkand (Sam), *blistered* (*bs*) and *EGFR* (E1).

Table 4.8. Summary of the top eight associations between *EGFR* and wing parameters.

Loc.	SNP	Trait	Term	*p-value*	Type	*f*	π	Div.	Repeat 1 *p-value*	Repeat 2 *p-value*	Effect
31656	T/C	Area	SNP x Sex	1.85E-06	N	0.86	0.014	0.0634	0.88587	0.61263	-
30200	C/T	C1	SNP	0.00003	N	0.71	0.0027	0.1003	0.00008	0.018	Correct
31634	C/T	C2	SNP x Sex x Pop	0.00005	N	0.53	0.0115	0.0504	0.35925	0.53630	-
40722	T/C	Area	SNP x Pop	0.00006	S	0.21	0.0089	0.00065	0.46681	NA	-
39389	T/C	D1	SNP	0.00006	S	0.3	0.0112	0.0059	0.44171	0.91484	-
5683	0/T*	W9	SNP x Sex	0.00007	N	0.48	0.0159	0.0153	0.84510	ND	-
30401	G/A	Area	SNP x Pop x Sex	0.0001	N	0.36	0.0026	0.0130	0.71134	NT	-
30505	C/A	W7	SNP	0.0002	N	0.26	0.0042	0.0693	0.00072	0.0635**	Reverse

Loc: Refers to the location of polymorphisms in Genebank record 17571116.
SNP: The ancestral state of the polymorphism and the derived condition as inferred by a comparison to *D. simulans.* * site is an insertion/deletion polymorphism.
Term: The genetic term and its significance (p-value) describe the association.
Type: Describes the nature of the base change, S: Synonymous, N: Non-coding,
f: The frequency of the derived allele.
Molecular evolution parameters π, the average pair-wise number of differences between alleles π and divergence (Div.) in 50 bp windows surrounding the site (25 bp on either side).
Repeat 1, p-value of SNP term as tested with Proc GLM in SAS.
Repeat 2, p-value of SNP terms, except ** which is a Cross by SNP term. In all cases were other terms including SNP non-significant.
NA: Site 40722 is not segregating in the Kenyan sample, ND: Area surrounding site 5683 was not sampled in the Kenyan population, NT: Site 30401 represented by a single allele in the Kenyan population, not tested.
Direction of significant (or almost significant) SNP effects, correct or reversed.

Table 4.9. Effects of site C30200T in *EGFR* on crossvein placement (relative warp C1).

	LSM of genotype					
Experiment	T/T	T/C	C/C	Est.	SDU	*P*
Inbred	0.0018		-0.0086	0.0104	0.82	2.7E-05
Round Robin						
Full	0.0045	-0.0014	-0.0057	0.0059	0.5306	7.6E-05
Reduced						
Block-A	0.0032	-0.0022	-0.0019	0.0054	0.5166	0.0542
Block-B	0.0054	0.0011	-0.0095	0.0043	0.3503	0.0609
Block-C	0.0049	-0.0023		0.0072	0.6577	0.0036
Test Cross						
Full	0.0044		-0.0016	0.0060	0.5036	0.0180
Reduced						
Sam	0.0015		-0.0052	0.0068	0.6187	0.0142
bs	0.0127		0.0069	0.0057	0.6137	0.0835
E1	-0.0010		-0.0069	0.0059	0.6082	0.0777

Experiment, refers to the initial association study (inbred) and the two follow up studies (Round Robin with WE lines and Test crosses with Kenyan lines). Data are presented for full analysis and reduced models.
Round Robin: Each WE line was crossed total of 3 times as female and 3 times as male. They were paired in 3 independent randomizations, represented by blocks A-C. The reduced analysis were done on each of those blocks separately and the full model on the united dataset.
Test cross: In addition to the full model, the effects are estimated for backgrounds only.
Sex was omitted in all the above models as they were very similar, and in case of the Round Robin because it was just included females.
Est: Estimated difference of genotypic classes of C30200T in relative warp units. In Inbreds and Test cross panels, the difference between T and C. Except in Round Robin where the difference was estimated between T and the heterozygous condition (T/C) as C homozygotes are rare. For instance C/C homozygotes are not even present in Block C.
SDU: Standard deviation units, the estimated difference between genotypes as a fraction of the standard deviation of line means. For Inbred and the Test cross those are averaged by sexes.

Discussion

Shape of the wing

Establishing the trait is the initial step when studying continuous variation. Description of shape by relative warp analysis on landmarks is a purely algorithmic process unconstrained by the scientist's preconception of shape, except for his/her choice of landmarks. Here I describeddissection of the shape variation in an insect wing as captured by two alternate approaches to the landmark data. One approach is to extract variation simultaneously for all 9 landmarks, representing the whole wing (Figure 4.1). The other is based on knowledge of wing development, and is to subdivide the landmarks into 3 groups defining the boundaries of intervein regions B, C and D. I chose to investigate 9 components from each methodology. There are three major findings from our analysis of shape. The axes of shape variation in the wing are restricted to few dimensions, shared by multiple datasets. The current enterprise proceeded by incorporating every emerging dataset and recalculating the warps. The absolute values of warps did not change significantly, and the correlation between old and new warps was high, particularly after the dataset entered tens of thousands of specimens. This is a major finding, suggesting that the evolutionary dimensionality of shape is restricted. The second major result is the high heritability of individual warps, suggesting a substantial genetic component even for subtle aspects of shape, like relative warps W8 and W9 that account for 1-2% of shape variation. This suggests seemingly minor aspects of form can be selected, consistent with the selection experiments of Weber (1990a 1990b, 1992). Those two results together document wing shape as occupying restricted domain of shape space and patterns of genetic variation as one of limiting factors. Testing for relative contribution of developmental, environmental and genetic causes for shape variation is however a complicated matter (Klingenberg 2002). So are analysis of quantitative variation in gene-action and developmental mechanics in the wing and other structures. Those could potentially be bypassed by searching for "oblique" solutions of principal components (Hatcher 1994). Such adjustments have not been implemented in standard morphometrics packages. The third observation is practical, documenting that the two approaches, whole wing or by intervein regions, capture similar but not entirely overlapping aspects of shape. Further studies of these relations in partial and larger datasets are needed.

The meaning of warp values in terms of shape attributes is graphically clear (Figures 4.2-4.5) but less so verbally. Morphometricians emphasize that relative warps are not capturing the displacement of individual landmarks but changes in the underlying geometry of shape (Bookstein 1991, Rohlf 1996). Biologically, changes in shape are brought about by differential growth in fields of cells or focal points that may or not correspond to individual landmarks. It remains to be seen if landmark based methods have the power to identify developmental

centers of organization and be used to construct hypothesis about the processes. Recent advances, enabling description of curved forms with quasi-landmarks offer a promising tool (Adams *et al.* 2003). Regardless, the relative warps have distinct meanings in terms of deviation from consensus shape. Some appear to capture common features, as for instance D3 and W7, both of which apparently relate to rotation of crossvein 2. Also W1, W2 and W4 all appear to reflect a width to length ratio but are still by mode of derivation independent parameters. Those first 6 parameters for the whole wing reflect integrated variation across the wing, while the latter 3 parameters are clearly identifying local changes in shape. This is a function of the procedure and must not be over interpreted. Landmark data enable identification of the contribution of individual landmarks to the overall variation in shape. Application of the new version of the TPS package (Rohlf 2002) shows that landmarks defining the crossveins seem to contribute bulk of the variation in shape (data not shown). Analysis of the developmental stability of wing phenotypes also showed that the cross-veins are less constrained then overall parameters (Woods *et al.* 1999). Waddington also noted this variation and utilized it to develop his ideas on canalization (1957). Those results should be interpreted with caution as the nature of the parameter extraction introduces a bias in this estimation, as the landmarks in closest proximity in each dataset always contribute the largest portion of the variance (data not shown). Intuitively the overall shape of the wing could be retained by fixing peripheral landmarks leaving the exact locations of the internal junctions of veins and crossveins to drift. It is not clear how to test this experimentally. In summary, TPS based relative warp analysis proved a powerful approach to quantifying variation in shape of *D. melanogaster* wings.

Effects of *EGFR* on shape

Alleles of several loci, including these key developmental genes, are known to have phenotypes disrupting particular veins or intervein regions. Birdsall *et al.* (2000) argued that positioning of veins is a major determinant of shape, and put forth a hypothesis that loci controlling vein development might contribute to standing variation in shape. Two lines of evidence support this, a QTL mapping experiment suggests a non-random aggregation of vein loci under QTL peaks. Secondly, quantitative complementation tests are consistent with main vein loci; *dpp, tkv, hh* and *EGFR*, harboring segregating variation for shape. Here I tested this hypothesis directly by asking if polymorphisms in *EGFR* affect wing shape.

To set the stage for the association tests, I aimed to establish a direct relation between *EGFR* function and parameters of shape. This was done by reanalysis of the data described in Chapter 1. First, I demonstrated that the axes of variation are sufficiently similar in joined and individual datasets and then explored the effects of major alleles of *EGFR* on wing shape. The logic is that a deletion and gain of function allele should have distinguishable effects as summed

up over multiple backgrounds. This pattern was found for a subset of parameters, for example W1 (Figure 4.6). Focusing on W1 we can, courteously disregarding the vocabulary of morphometrics, say loss of *EGFR* results in shortening of the wing, posterior displacement of crossvein 2 (landmarks 2 and 3) and the junction of L2 and the wing margin (landmark 6). The effects of *EGFR* loss on C1 would be particularly interesting given the repeatable associations but the results are inconclusive. The results for other parameters suggest that if *EGFR* plays a role, then it does so through complex dependence on genetic backgrounds and sex. The current experimental design is not ideal to address this question, as the alleles of interest differed by not by the lesions in *EGFR* but also by other polymorphisms on the same chromosome. That could be the case for C1 as the two deletions have opposite effect, with the gain of function allele in the middle (Figure 4.6). Better strategies would be stable P-element insertions in a common strain or introgressed major alleles into one or more background. The first strategy has been successfully implemented for several traits, (Lyman *et al.* 1996, Clark and Wang 1997, Lai *et al.* 1998, Fedorowicz *et al.* 1998) while the latter and more laborious procedure has only been used in a handful of studies (Gibson and van Helden 1998, MacKay and Lyman 1998, Lyman *et al.* 1999, Polaczyk *et al.* 1999, Long *et al.* 2000, Robin *et al.* 2002) to investigate variation among natural alleles.

Heritability and population differentiation

The two main questions regarding the distribution of wing shape metrics concern the degree of heritability and population subdivision. Partitioning of the variance by ANOVA (Appendix C) established significant sex and line effects, the latter indicating a genetic component. The proportion of additive genetic variance for a trait in a given population is quantified by the heritability and was estimated by the variance component for line. I estimated heritability for sex individually on distinct and joined populations (Table 4.2), and yielded values ranging between 0.28 and 0.68. Size related measures have on the other hand distinctly lower heritability (0.15-0.27), consistent with Roff (1997) and Birdsall *et al.* (2000). The heritability estimates by sex are congruent but the North Carolina population had lower heritability estimates across traits and sexes. This could reflect different distribution of alleles in the two populations, or be a function of the inbreeding intensity. Consistently the Californian sample underwent twice the generations of sib-mating, which can increase the additive genetic component. The heritabilities for shape are at the high end of estimates for *Drosophila* traits (Falconer and MacKay 1996, Roff 1997) and higher then other metrics of wing shape reported by Cowley and Atchley (1990). It is possible that their estimates where confounded by size, as the tools of new morphometrics were not available.

A test of population differentiation at the phenotypic level was preformed at several levels. The most convincing evidence of population differences were provided by ANOVA of phenotypes, as four traits had significant population effect. In addition there where significant population by sex interactions for four more shape traits and all the size measures. Finally, 95% confidence intervals constructed for genetic correlations don't overlap for 7 pairs of traits. This observation is conditioned on the fact that these confidence intervals do not account for multiple tests. Interestingly these signals were observed for the smallest relative warps, and thus do not account for largest portion of the variance. However these results together imply significant geographic differentiation in allelic variation contributing to the traits of interest. It could be caused by population subdivision or differential stratification between the populations that can not be distinguished with analysis of phenotypes. In addition these results demonstrate the evolutionary uncoupling of shape components, which can differentiate at the genetic level, either by drift or selection. Significance of the population term also places the tests of association into a context were the phenotypic differences can generate false positives. This will be discussed further below.

Relations of shape parameters

The current study aims to embrace the multidimensional shape space of the *D. melanogaster* wing. The practical question is concerned with the level of independence between variables, as we would like to contrast the two strategies for shape analysis. The biological one asks how three dimensional form is realized by development. Both questions were addressed by studying the phenotypic and genetic correlations between shape and size parameters as partitioned by sex and population.

The phenotypic correlations are low and only highlight exceptionally strong relations (Table 4.4) like the size measures. The genetic correlations were more pronounced and disentangled more intuitive relations between variables, like the relation between W1 and wing length (L1). Comparisons of genetic correlations between parameters derived from the whole wing vs. inter-vein regions shows a largely one-to-one correspondence. There are exceptions and in summary the two methods capture common but not entirely overlapping aspects of shape. For the purpose of successfully describing the phenotypic space then I believe this dual approach is justifiable. The outcome is an abundance of parameters which affects the significance cutoff for the association tests. Presence of significant genetic correlations does not pose problems for identification of natural genetic variants affecting phenotypes. For instance analysis on bristle number focused on two regions of the adult fly, the abdominal segments and the sternum (MacKay 1995, MacKay 2001). While there is significant genetic correlation

between the phenotypes, both QTL analysis and association test routinely attribute phenotypic variation in each to distinct genetic factors (reviewed by MacKay 2001).

Cowley and Atchley (1990) surveyed wing shape along with additional adult structures and found strength of genetic correlations reflecting common developmental origins. The wing develops from the same imaginal disc, but is then compartmentalized by the action of multiple loci (Chapter 1). Klingenberg and Zaklan (2000) provided evidence for pervasive integration in wing development that did not adhere to compartment boundaries. Here I note the primary whole wing parameters, comparable to Klingeberg's metrics, are correlated to multiple IVR-parameters consistent with developmental integration being important. However again, it is not clear if relative warp analyses are the optimum tool to elucidate these relationships. The two main observations may seem at odds, one concluding independence and the other assimilation, but constrained, uncoupled and evolvable traits are at the essence of evolutionary developmental biology (Raff 1996) and appreciated by evolutionary geneticists (Falconer and MacKay 1996, Lynch and Walsh 1998). In short my findings differ barely from those of paleontology documenting patterns of constraint and divergence in fossilized bones over large timescales (Gould 1977). Except here we have a living population amenable to dissection at the genetic, developmental and fitness level and a set of techniques to unravel the relationship between phenotypic attributes. The utility of the wing system for dissection of evolutionary genetics of shape is probably only matched by the mouse mandible (Atchley and Hall 1991, Leamy *et al.* 1998). Either structure should be used in studies of the molecular basis of genetic correlations to link those with genetic variation, either polymorphisms or genomic transcriptional attributes (Jin *et al.* 2001, Rifkin *et al.* 2002).

Association tests

The hypothesis tested here is: do polymorphisms in *EGFR* contribute to natural variation in wing shape in fruit flies? This was done by conducting three association experiments, an initial study and two smaller studies for verification. Together the results lead to rejection of the null hypothesis of no association, and are consistent with QTN's for wing shape segregating in *EGFR*.

Test of association for wing size

Size measures were *a priori* considered control phenotypes for the association tests, as *EGFR* was considered a major candidate for shape variation, on the basis of our previous results and developmental genetics. Its best characterized function in the wing is differentiation and proliferation in vein tissue. The second role occurs after the specification of vein tissue as the receptor is required in the intervein regions, potentially by providing a survival signal (Martin-Blanco *et al.* 1999). Decapentaplegic has been shown to act precisely in that manner during

wing development (Moreno *et al.* 2002). However EGFR stands for Epidermal growth factor receptor and has been shown to contribute to cell growth in *Drosophila* tissues (Held 2002). Still it was a considerable surprise that the most significant association in the inbred lines was to attribute of size. It was significant after Bonferroni adjustment of the false discovery rate to account for the staggering 5000 ANOVA's. Variant T31365C in intron 2 affected the size of the wing contingent on sex and population by sex. The derived C variant is at high frequency (0.85) and may cause larger wings in females (Figure 4.16). The effects are more pronounced in the Californian sample, suggesting why the association was not replicable in experiment 2 (which included only WE lines). Lack of repetition in the Kenyan cross could be attributed to its small scale. The fact that QTN T31365C with complex effects on size does not replicate suggests, regardless of the initial *p*-value, that it might be a false positive. On the other hand, if its dependence in the Californian lines is caused either by novel mutations or alleles differing substantially in frequency between the populations, then it could be real. Two other sites which had complex suggestive signals to size did not replicate either. One is a silent site in the exon 6 while the other is in the promoter of exon 2. The third possibility is that the additive effects of the size SNP's are drowned by dominance of other factors in the genetic backgrounds. Consistently QTLs for wing size show extensive dominance while shape measures do not (Zimmerman *et al.* 2000). Finally, the reason for lack of replication could be that variants with quantitative effects are sensitive to stochastic variation which reduces repeatability (Long and Langley 1999). This remains a significant problem for studies fewer than 500 subjects when modeled for human disorder settings. The accurate measures enabled by utilizing clonal stocks of fruit flies allow us to get away with 200 "subjects" but the problems still persists, particularly for my stratified sample.

Test of association for wing shape

The association study in inbred lines did not yield variants with experiment wide significant effects on wing shape. There were five sites which were described as suggestive, and retested in the follow up experiments. Of those C30200T, the variant with strongest initial signal had replicable effects on distance between crossveins. More importantly, the effects of the site are repeatable, both in recrosses of WE lines and in a new pool of alleles (Table 4.8 and Figure 4.20). The magnitude of the effect ranges from 0.8 standard deviation units to ~0.5 units in the follow up experiments. Diminishing effects and significance in repeated studies have been documented for multiple human disease variants (Ioannidis *et al* 2001), suggesting that identification of causative variants depends greatly on the favorable effects of chance in the original sample. At the molecular level, C30200T disrupts a putative GAGA binding factor in alternate promoter RB and can be regarded first regulatory mutation in *EGFR*. Several pieces of evidence argue for the role of the element and potentially the C30200T polymorphism for

function. i) The element is located in the more conserved promoter, with the regions surrounding the element highly conserved. ii) The whole element is similar to characterized GAGA elements in *Ubx* (Hodgson *et al.* 2001). iii) The pattern of polymorphism and divergence conserve a di-cytidine periodicity that is important for GAGA factor binding. iv) Site C30200T is the only polymorphism at high frequency disrupting the C periodicity of the element. v) The association to shape parameter C1 is replicable and consistent in terms of effect in two populations and two follow up studies. The hypothesis of the effects of site 30200 on *EGFR* function could be addressed with transgenic experiments, or controlled genetic tests of the interaction between natural and major *EGFR* variants and the GAGA factors in the Drosophila genome. Of the remaining sites, only C30505A affecting W7 showed a degree of replication. It is also located in the 5' untranslated region of the RB-transcript, 13 bases up from the start codon. However the fact that this site does not replicate in the Kenyan test cross, and has opposite effect casts doubt on its true significance. Clearly a more extensive sampling is required to verify its effect. It is interesting that significant and suggestive associations cluster in the vicinity of exon 2 (Table 4.8 and Figure 4.22). Temporal and spatial expression of the transcripts is largely overlapping during development and no sex based differences have been reported (Lev *et al.* 1985, Scheiter *et al.* 1986, Kammenmayer and Wadsworth 1987). The molecular effects for the non-coding polymorphisms are most probably at the level of RNA transcript, before or after splicing. The polymorphisms could affect transcriptional regulation, and given the dynamics of *EGFR* expression in the vein primordia, where down-regulation of expression appears as the major transition then disruption of binding sites for negative regulators is a plausible mechanism. However the effects could also be on splicing control or mRNA stability. As neither the regulation of transcription nor RNA stability of *EGFR* have been investigated those will both remain viable hypotheses.

The results are unique as all previously described alleles of *EGFR* affect the protein proper. The phylogenetic shadowing, polymorphism levels and significant associations could be an indication of the specific regulatory regions of *EGFR* in Drosophila and suggestive of transcription factors important for *EGFR* regulation in the wing. This is achieved by testing for the association between naturally occurring polymorphisms and variation in shape, and does therefore not run into the same problems as studies of major alleles of pleiotropic genes. According to Clifford and Schüpbach (1989, 1992 and 1994) major *EGFR* mutations fall into three distinct complement groups with a complicated pattern of developmental consequences. There was also a dust-bin category of alleles that could not be assigned into these classes due to specificity of effects and peculiar patterns of complementation. This argues for complex effects of *EGFR* in development, where the locus acts in multiple tissues, in a combinatorial fashion with local pool of proteins to elicit a range of effects. This is corroborated by another

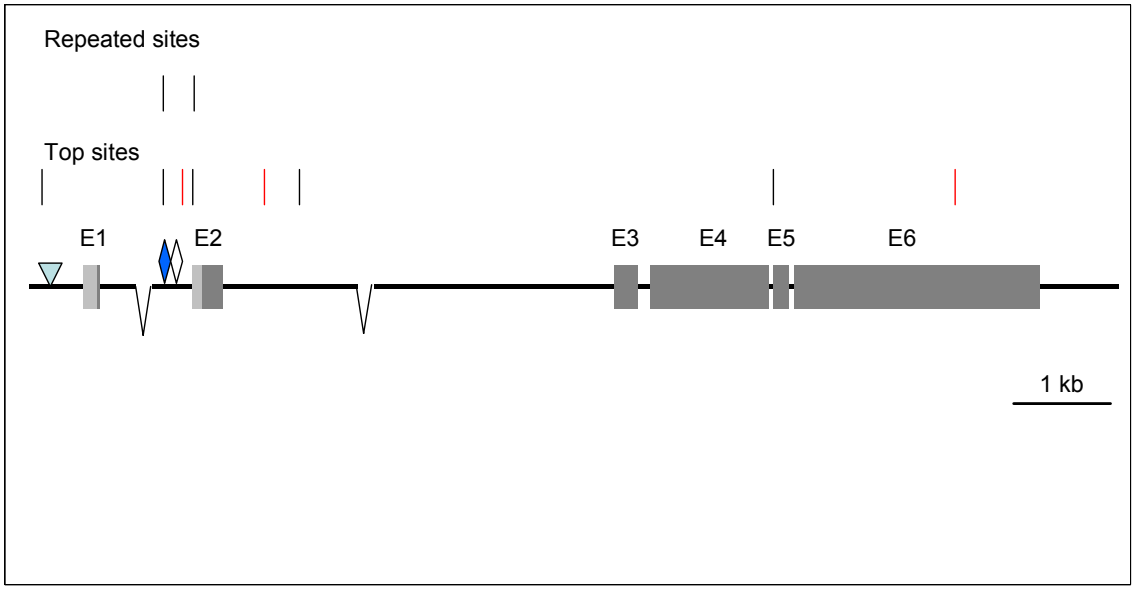


Figure 4.22. Location along *EGFR* of the 8 sites reported in Table 4.8, affecting size (red) and shape (black) of the *D. melanogaster* wing in natural populations. Two of those gave significant associations in follow up experiment and are represented above (Repeated sites). Gene structure is the same as in Figure 3.1.

study in our laboratory, where the same populations and genotypes were used to test for the contribution of *EGFR* polymorphisms to cryptic variation for photoreceptor determination (I Dworkin, K. Birdsall, A. Palsson and G. Gibson submitted, data not shown). Eye development was sensitized by crosses to *Ellipse*, a gain of function allele of *EGFR.* A set of three silent sites with in high pair-wise LD in the receptor tyrosine kinase (RTK) domain were significantly associated with eye-roughness. The effects were replicated in a fresh sample of wild chromosomes as judged by case control and TDT tests. There is also evidence for epistatic effects between the sites, suggesting a complicated mode of action. Distinct *EGFR* polymorphisms associating with aspects of wing and eye development indicate general pleiotropic effects of segregating variants in D. melanogaster *EGFR*.

I can not firmly state C30200T is the causative factor as it could be linked with another truly contributing polymorphism. This is particularly important as we sequenced only ~11 kb of the ~ 40 kb locus, and none of the flanking regions. Such comprehension should be goal in future association studies. The most parsimonious interpretation is that the effects are caused by C30200T or a closely linked site, which is most probably within the locus.

Population and sex dependence of effects

What does a polymorphism by sex or population interaction really mean? Both sex and population can be viewed as indicators of genetic backgrounds. The capacity of segregating variation to suppress or enhance the effects of specific mutations has been widely documented. For example the penetrance of the homeotic *Ultrabithorax* mutations in *Drosophila* are modified by backgrounds after being introgressed into several strains (Gibson and van Helden 1998). Similar results have been found for Antenna to leg transformations by *Antennapedia* (Gibson *et al.* 1999), sensory bristles on the wing induced by *hairy* (Fletcher and Thompson 2001), and eye-roughness in *sevenless* and *Ellipse* mutants (Polaczyk *et al* 1998). In humans the relative risk of breast cancer in carriers of the major BRCA-2 is increased by a factor of four, if they also have 135C variant of RAD51 (Levy-Lahad *et al.* 2000). The RAD51 variant did not confer significant risk independently and can therefore be regarded as a modifier. The specificity of the effects is marked by BRCA-1 carriers not being affected by the allelic state of RAD51. Precisely, in terms of population effects, the first association study in *Drosophila* to sample multiple populations surveyed only 36 chromosomes. No differences between populations or population dependence of allelic effects were identified (MacKay and Langley 1990). These results might be attributable to small sample size so the current experiment can be considered the first that has sufficient power to detect population specific effects of individual SNP's.

The current experiment, as implemented, did not enable distinction of the exact nature of the population differences. Population stratification or subdivision can be estimated by survey

of genomic markers, with 40-100 being sufficient to reject hypothesis of major stratification (Pritchard and Rosenberg 1999, Remington *et al.* 2001, Ardlie *et al.* 2002. The experiment did however allow estimation of sex and population dependence of SNP effects. Half of the significant and suggestive SNP's reported in table 4.8 show conditional effects. The data also showed traits with distinct phenotypic distributions in the two populations also had excess of population by SNP terms significant at the 0.05 level (by individual tests). This could either be a true pattern or a statistical artifact. The fact that none of those associations are significant when correcting for multiple tests, neither experiment wide nor less stringently by traits, supports the artifact hypothesis. But the issue will remain unresolved until we can compare large number of purely additive QTN's and polymorphisms with true sex or genetic background dependent effects. The second issue that the two populations allowed us to tackle concerns the rate of false positives from sites with significant F_{ST}. The data are nor persuasive as only three associations at 0.05 level were detected with those sites, and only one to a trait with different distributions in the two populations. But the fact only few associations were detected with the F_{ST} sites is consistent with these differences in allele frequency not being large enough to create artificial associations. Coupling of quantitative and ecological genetics in a large scale systematic survey of nucleotide variation in several clines is needed to provide basis for tests of naturally occurring variants in phenotypes.

Conclusions

By utilizing the amenability of fruit flies for genetic studies and the robust statistics of quantitative genetics and morphometrics I have tackled a composite morphological structure and provided evidence for the contribution of segregating variation in *EGFR* on wing shape parameters. This study extends on previous studies of association of natural genotypic and phenotypic variation in *Drosophila* along three lines, the multidimensionality of trait space, the depth of genotyping and by testing alleles from more than one population. It is easy to envision an extension of this approach where numerous traits and genes are scored simultaneously and then analyzed in a unified manner. While analysis of variance provided a solid assessment of effects then permutation based methods could have provided better assessment of significance then standard corrections for multiple tests. Development of software modules within standard statistical packages to perform randomization with structured datasets would be a great benefit for similar future studies.

I detected significant genetic effects of a segregating variant in the *EGFR* locus on wing shape. The signal was only suggestive in the initial experiment, due to the multiple dimensions of the experiment, but the individual p-value (2.71×10^{-5}) is among the largest reported in *D. melanogaster* (MacKay and Langley 1990, Lai *et al.* 1994, Long *et al.* 1998, Lyman *et al.* 1999,

Long *et al.* 2000, Robin *et al.* 2002, Geiger-Thornsberry and MacKay 2002, De Luca *et al.* 2003). The association was replicated in two separate follow up experiments adding support for the inference that allelic variation in EGFR affects wing shape. The non-coding variant resides in a region of presumed regulatory function, a GAGA factor binding element in promoter of alternate 5'-exon. This is particularly noteworthy because no mutants have been found that impact transcriptional or translation regulation of *EGFR* in *D. melanogaster*. It remains possible that the true causative site was not scored in the current sample as sampling focused on coding and promoter regions. Molecular transgenic experiments could provide formal proof of the functional effect of this natural polymorphism (Choudhary and Laurie, Dunn and Laurie 1995, Stam and Laurie 1996, Ludwig *et al.* 1998). Association to warp C1, which captures the relative distance between crossveins, is particularly exciting given the results of the complementation tests, which implicated allelic variation in *EGFR* affecting the anterior (IVR-B) and central (IVR-C) part of the wing (Chapter 2). These results are also interesting considering that Ennos (1988) postulated, using biomechanical analysis of wings, that number, strength, and location of crossveins were primary determinants of wing rigidness. The flexibility of wings is of key importance for flight performance, when the wing undergoes complex rotations and flaps (Dickinson *et al.* 1993, Dickinson *et al.*1999, Dudley 2000, Fry *et al.* 2003). It is therefore tempting to monitor the frequency of the C30200T variant along established clines of *D. melanogaster*, and study both wing shape and flight attributes.

Chapter 5

Thesis conclusions

Project aims and success

My interest was to identify segregating polymorphisms contributing to the genetic and biological basis of naturally occurring variation in phenotypes. I chose to investigate the contribution of genetic variation in fifteen regulatory genes and molecular variation in the *EGFR* locus on aspects of wing shape in fruitflies. Have the specific aims of the study have been met?

A. The first goal was to define morphometric procedures useful for investigating the biology of wing shape. Utilizing the established tools of morphometrics and building on work by Klingenberg (2002), my advisor and lab members (Birdsall *et al.* 2000) we have now a clear protocol for analysis of wing shape. A wealth of data accumulated for this thesis and other experiments in the lab suggests that variation in wing shape follows common trajectories. Our results are consistent with those of Klingenberg and Zaklan (2002) which demonstrated pervasive integration in the wing, but also imply that a substantial fraction of the genetic variation does have localized effects on particular regions of the wing (Zimmerman *et al.* 2000, Chapters 2 and 4).

B. As resolution of QTL mapping is low (Zeng 1994, Zeng *et al.* 2000, MacKay 2001) additional tests of allelic effects are required to identify a subset of loci harboring segregating variation for a trait. I applied quantitative complementation tests to directly test the hypothesis put forth by Zimmerman *et al.* (2000) concerning the role of venation loci for shape variation. Fifteen major wing mutations including *wg, en, dpp, hh*, and *EGFR* representing the predominant patterning pathways were tested. The results are consistent with the venation pathways and the canonical members, *hh, dpp* and *EGFR* harboring segregating variation affecting shape. Along with MacKay and Fry (1996), Lyman and MacKay (1998), Long *et al.* (1996), Lyman and Mackay (1998), Ashton *et al.* (2001) and Pasyukova *et al.* (2000) this demonstrates the utility of the approach for fine mapping of QTL's and testing of candidate loci.

C. Molecular evolution and population genetics offer a distinct way to address questions about the long and short term evolutionary forces affecting candidate loci. The selection of *EGFR* for these analyses was channeled by our quantitative genetic evidence and the prominent role of *EGFR* in vein formation. We sequenced 10.9 of the total 40 kb locus, focusing on exons and adjacent region, and the results are consistent with the protein and certain non-coding regions experiencing purifying selection. Peculiarly, one of two alternate N-termini was

evolving at a higher rate then the rest of the protein. The different levels of analysis also highlight an interesting pattern of conservation in a putative GAGA factor binding site. Linkage disequilibrium is low within the locus, with high values of r^2 only spanning 500 bp at the most, and no differences noted between the three populations. Analysis of F_{ST} along the locus demonstrated a degree of differentiation, particularly in reference to the African population. The two North American populations also differed by just a few sites, but those differences in allele frequency did not translate into distortion in LD. If selection was affecting distinct alleles in each population then the effects must be small, and nothing that can be regarded as a selective sweep or adaptive evolution (Barrier *et al.* 2001) was detected.

D. The last goal was to try to resolve a QTL down to individual nucleotide differences. Tests of association between polymorphisms in *EGFR* and variation in wing size and shape are significant. Site T31656C has sex-dependent experiment wide significant association with size, but does not replicate in follow up experiments. The lack of replication may have been caused by unique sex by population dependence of the SNP. On the other hand site C30200T did not pass the experiment wide cutoff but is repeated consistently, showing no dependence on sex or population. It disrupts a proposed GAGA binding domain in the promoter for exon 2 and affects crossvein placement in the central portion of the wing. Genotyping was extensive but not comprehensive, leaving open possibility for an unknown second site being the true cause of effects. With the low LD in the region then the most plausible scenario is that C30200T is the causative site within *EGFR,* substantiating the hypothesis that natural variation in *EGFR* affects wing shape.

Practical lessons

The applicable results from the project are mainly twofold. First the sequence analyses of *EGFR* utilized different evolutionary timescales allowing additional insights. In particular, coupled surveys of polymorphism and divergence suggest that purifying selection is operating on particular regions while not causing significant deviations from neutral expectations. The increased magnitude of the population sample did not greatly elevate the power of standard molecular evolution statistics, but did yield insights into indel polymorphism distribution. Second, the depth of sampling, by sequencing of 25% of the locus, was of primary importance for resolution in the association mapping. Within a particular region, each site became an individual test, thus reducing dependence on linkage to mediate effects of QTN to marker. The main justification for selection of the approach was the low LD in Drosophila. The obvious caveat is that 75% of *EGFR* and the flanking genomic regions were not sequenced. The effects detected could therefore be caused by other polymorphisms in LD with the scored marker, particularly markers bordering on non-sampled regions. In case of the sampled region, less detailed

genotyping would not have been sufficient to identify the two putative QTN's. Also, the high number of sites and phenotypes tested did affect the conclusions as we proceeded to correct for multiple tests with classical methods. This has ramifications for attempts to scale up association studies along both of those axes. Permutation based methods should be explored for this or comparable datasets.

With more extensive genotyping of candidate loci then background effects and the stochastic Beavis effects will contribute greatly to formal detection of experimentwide significance. The Beavis effect states that some portion of a detected QTL effect can be attributed to chance distribution of variance, resulting in inflated estimates of QTL effects. In an inbred design, other segregating alleles may also confound or enhance the contribution of the polymorphisms tested. Those are less important in more controlled designs, (MacKay 2001; Chapter 1), that conversely risk detecting conditional polymorphisms. Coupling of association studies with inbred lines and with chromosomes extracted from those inbred lines into a common background could illuminate this issue.

Intriguingly, even with only 25% of the *EGFR* region sampled we still detected significant associations with wing shape. Certainly the sampling favored exons and promoter regions expected to affect function, while another 50% of the remaining region encodes three other genes nested in an *EGFR* intron. The simplest interpretation is that the scored polymorphism causes the effects but it is a distinct possibility that an unscored variant in LD is really responsible. These associations found here and the rapid decay in LD are similar to the conclusions of MacKay and coworkers (MacKay *et al.* 1990, Lai *et al.* 1994, Long *et al.* 1998, Lyman *et al.* 1999, Long *et al.*2000, Geiger-Thornsberry *et al.* 2002 and Robin *et al.* 2002) on allelic variation in candidate loci for bristle number. They sampled in each case on the order of 30 polymorphisms in loci of length comparable to *EGFR* and found significant associations. They did not claim to identify exact QTN's but concluded that allelic polymorphisms in those loci affect bristle number Together the evidence underscores the utility of Drosophila as a model for refining fine mapping methods. The relative ease by which the protocol was extended to a multidimensional phenotype is also encouraging for broader use of the fly as a model for complex disease.

The evolutionary fate of wing QTN's

The evidence from complementation tests and the associations between *EGFR* and wing shape suggest that major developmental genes can be functional agents in the evolution of morphologies. These inferences are similar to those from work on the quantitative genetics of *Drosophila* bristles (Mackay 2001). Comparison of gene expression in butterflies suggests key loci also contribute to variation in other insects (Carroll *et al.* 2001, McMillan *et al.* 2002).

Mapping in butterfly lines selected for eye-spot size also implicated a key regulatory gene *distal-less* as a contributor, with molecular evidence for corresponding changes in mRNA level in the eyespot primordia (Beldede *et al.* 2002).

Are effects observed in controlled laboratory experiments going to be relevant in nature? Are they substantial enough to be seen by natural selection, and what are their fitness effects? Weber (1992) proved quite elegantly that shape does respond to selection and suggested that it does so in a gradual fashion, suggesting abundance of available additive variation. Selection may act directly on wing shape in nature as repeated clines in wing length across continents are observed (Imasheva *et al.* 1995, James *et al.* 1996, Huey *et al.*2000). Clines of a trait can be a byproduct of population structure or pleiotropic relations to another trait under selection (Eanes 1999). Gockel *et al.* (2001) showed that the Australian cline in wing length is not generated by serious population stratification. Similarly Coyne and Beecham (1987) found different clines for size, wing length and bristles down the North American heartland, arguing that clines in traits can be quite uncoupled. Wing length and presumably shape may therefore be a target of selection. Selection can act quite rapidly as in the case of *D. subobscura,* which 30 years after populating the new world had reproduced the old world cline (Huey *et al.* 2000). The exact mechanism of selection is not known but the biological functions of the wing include flight, courtship, and perception through its bristles. Likewise no ecological attributes that may be responsible for clinal differences have been identified, though wing development is sensitive to temperature, humidity, and nutrition (Calboli *et al.* 2003).

Two broader arguments can be made that the effects of natural polymorphisms on phenotypes are indeed subject to selection. First, evolutionary forces act on the allelic variation in the entire population over time. This is a notably Fisherian view (Fisher 1930, Lessard 1997), but defendable as the utility of Wright's shifting balance theorem was cast in doubt (Coyne *et al.* 1997). Therefore, an allelic variant explaining 0.5% of the phenotypic variance in current experimental settings can be assumed to be subject to selection if the trait influences fitness. Second, analysis of codon usage across taxa demonstrate persistent and lineage specific patterns. Selection on codon usage has a very low selection differential but is still able to mold synonymous allelic variation (Powell 1996, Comeron and Kreitman 1998), arguing that the small effects observed here may indeed be subject to selection. Analysis of wing shape, flight and the frequency of C30200T along an established wing shape cline could address this question.

Quantitative developmental biology

Loyal to the idea of large scale dissection of multiplicity of composite traits, my efforts are part of larger experiment in which seven other loci are being genotyped and four more phenotypes assessed in the same two sets of inbred lines. As the traits fall into two classes of pharmacological (heartbeat and drug survival) and morphological (wings, eye roughness and

dorsal appendage spacing), with candidate genes corresponding to each class, this dataset will allow a unique comparison of associations across traits and genes. We can therefore use the biologically real data as a baseline for our statistical analysis and evolutionary models. But experiments of this design prompt a philosophical question: how exhaustive a description of natural variation should we aim for? Let us consider the genotype level because the problem is commonly acknowledged. The lowest level of resolution is DNA. This notably disregards altered bases such as methylated Cytidines that can be sidelined as a product of epigenetic machinery or mitotic mutations, and are generally considered of minor consequence, except in cancers. It can be safely said that the overwhelming portion of the genetic variation impacting phenotypes will be observable at the A, T, C and G level. Imagine that our interest is in the genetic diversity of a particular population. It is common to extract a sample for analysis and proceed to genotype. Generally researchers choose a subset of markers to score, but recent advances in sequencing have led to proposals to sequence complete genomes in a study population. Such ventures can be considered the “complete” experiment of heritable variation and clearly out of reach for the average scientist. Instead we design the experiments to survey a fraction of the genotypes and hope that we can detect signals that warrant further study, as for instance in a QTL mapping experiments or survey of a candidate locus. In the case of Drosophila I have now demonstrated that genotypes achieved by sequencing of large fragments gives us additional understanding when testing for phenotypic associations.

Now, would the same hold for the phenotypic dimension? While no empirical data are available we can still contemplate the issue. Biologists normally describe or quantify certain features of the organism under study in a targeted fashion. For instance in most investigations in fruit flies bristles on two specific organs, the ventral abdominal plate and the sternum, are counted. This is by no means the complete representation of Drosophila bristles, as bristles are found on virtually every body part and appendage of the creature. Counting them all is clearly impossible, especially for quantitative genetic purposes. Still if absolute counts of all the bristles were available then they would most certainly be useful for questions on developmental and natural genetic variation. The third axis we consider in this thought experiment is perpendicular to the other axes, as it focuses on the patterns of variation during development. This is naturally based on our interest in opening up the black box of quantitative genetics and to determine how a genetic polymorphism transforms into phenotypic variation. One can imagine tracking variation at different levels during development, on morphological entities like limb buds or cellular populations, on concentrations or distribution of macromolecules including proteins, the dynamics of mRNA stability, chromosome arrangements and so forth. The metrics could assess variation in abundance, stability, shape, size, adhesion, enzymatic speed and other biochemical properties. These levels of complexity make this hypothetical enterprise considerably more complicated then for instance sequencing to get full genotypes.

This thought experiment must be considered in its philosophical surroundings, as it rests on deterministic ideas. Evolution has long been recognized as a capricious process but the degree of unpredictability in development is less appreciated. Even though developmental biologists talk of deterministic and regulated processes as clearly distinct, the underlying assumption is one of an unfolding program. Certainly development, by definition, has to be deterministic but it is the extent of variation in this process that is being explored here and utilized by evolution. From the practical standpoint differences in development can be perceived by selection and some fraction of those should be detectable by our analysis. It is reviving LaPlace's Demon to think that if we would acquire a full description of variation in morphological and molecular attributes during development then we would understand how genetic differences manifest in phenotypes. Life is a product of a stochastic physical world and it is possible that our investigations run into uncertainty principles that may prevent us from answering elementary evolutionary questions. Phenomological descriptions of variation in development trace back to the naturalists, but manifest more recently in construction of developmental and expression atlases (Meir 1997, White *et al.* 1999). Recent work on growth curves in mice (Atchley *et al.* 1997) provide a better quantifiable level of analysis and found substantial variation in late vs. early growth. A particularly striking example is provided by the nematode C*aenorhabditis elegans,* which proceeds through a stereotypical, commonly considered invariant, pattern of cell divisions to produce the adult. de Lattre and Felix (2001) report on variation in developmental fate of individual cells in the vulval complement group. They observed within-species differences in the cells of the vulva that have diverged compared to related species. Both of these studies address differences at the level of tissues or cells, but one can also focus on biochemical attributes like enzyme function or transcriptional variation. The most applicable technique is microarray analysis both because of streamlined protocols and its comprehensive nature (Jin *et al.* 2001, Rifkin *et al.* 2003). For instance one could study transcriptional variation at several developmental points among distinct natural isolates. Those kinds of experiments will give us ideas about the tracts of variation during development, aiding our building and testing of hypotheses about the developmental manifestation of quantitative trait nucleotides.

Classical and quantitative genetics started on the level of differentiated phenotypes, but have now faithfully embraced the molecular nature of the gene. It is standard technique to isolate genetic factors and test for correlations between the phenotypic and genotypic state, but the exact mechanism of an effect normally requires molecular analysis, which is by default more difficult for loci of small quantitative effect. It is my hope that later genetic analysis of natural variation will manage to address the process of genotypic representation by coupled investigations of phenotypic, genotypic and developmental variation.

References cited

Abi-Dargham, A., Rodenhiser, J., Printz, D., Zea-Ponce, Y., Gil, R., Kegeles, L. S., Weiss, R., Cooper, T. B., Mann, J. J., Van Heertum, R. L., Gorman, J. M., and Laruelle, M. (2000). "Increased baseline occupancy of D2 receptors by dopamine in schizophrenia." *Proc Natl Acad Sci U S A*, 97(14), 8104-9.

Abouheif, E., and Wray, G. (2002). "Evolution of the gene network underlying wing polyphenism in ants." *Science*, 297, 249-252.

Adams, D., Rohlf, F. J., and Slice, D. E. (2003). "Geometric morphometrics: ten years of progress following the "revolution"." *Italian Journal of Zoology*. In press.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., and et al. (2000). "The genome sequence of *Drosophila melanogaster*." *Science*, 287, 2185-2195.

Aquadro, C. F., Bauer DuMont, V., and Reed, F. A. (2001). "Genome-wide variation in the human and fruitfly: a comparison." *Current Opinion in Genetics and Development*, 11, 627-634.

Ardlie, K. G., Lunetta, K. L., and Seielstad, M. (2002). "Testing for population subdivision and association in four case-control studies." *American Journal of Human Genetics*, 71(2), 304-311.

Ashton, K., Wagoner, A. P., Carrillo, R., and Gibson, G. (2001). "Quantitative trait loci for the monoamine-related traits heart rate and headless behavior in *Drosophila melanogaster*." *Genetics*, 157(1), 283-294.

Atchley, W. R., and Hall, B. K. (1991). "A model for development and evolution of complex morphological structures." *Biological Reviews*, 66, 101-157.

Atchley, W. R., Xu, S. Z., and Cowley, D. E. (1997). "Altering developmental trajectories in mice by restricted index selection." *Genetics*, 146(2), 629-640.

Bainbridge, S. P., and Bownes, M. (1988). "Ecdysteriod titer during *Drosophila* metamorphosis." *Insect Biochem Mol Biol*, 6618, 185-197.

Barnes, P. T., Sullivan, L., and Villella, A. (1998). "Wing-beat frequency mutants and courtship behavior in *Drosophila melanogaster* males." *Behavioral Genetics*, 28(2), 137-51.

Barrier, M., Robichaux, R. H., and Purugganan, M. D. (2001). "Accelerated regulatory gene evolution in an adaptive radiation." *Proc Natl Acad Sci U S A*, 98(18), 10208-10213.

Barton, N. H., and Turelli, M. (1989). "Evolutionary quantitative genetics: how little do we know?" *Annual Review of Genetics*, 23, 337-370.

Begun, D. J., and Aquadro, C. F. (1993). "African and North-American populations of *Drosophila melanogaster* are very different at the DNA level." *Nature*, 365, 548-550.

Beldade, P., Brakefield, P. M., and Long, A. D. (2002). "Contribution of Distal-less to quantitative variation in butterfly eyespots." *Nature*, 415, 315-318.

Bergman, C., and Kreitman, M. (2001). "Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences." *Genome Research*, 11, 1335-1345.

Bergman, C. M., Pfeiffer, B. D., Rincon-Limas, D. E., Hoskins, R. A., Gnirke, A., Mungall, C. J., Wang, A. M., et al. (2002). "Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome." *Genome Biology*, 3(12), 1-20.

Berry, A. J., Ajioka, J. W., and Kreitman, M. (1991). "Lack of polymorphism on the *Drosophila* 4th chromosome resulting from selection." *Genetics*, 129(4), 1111-1117.

Betancourt, A. J., and Presgraves, D. C. (2002). "Linkage limits the power of natural selection in *Drosophila*." *Proc Natl Acad Sci U S A*, 99, 21.

Biehs, B., Sturtevant, M. A., and Bier, E. (1998). "Boundaries in the *Drosophila* wing imaginal disc organize vein-specific genetic programs." *Development*, 125(21), 4245-57.

Bier, E. (2000). "Drawing lines in the *Drosophila* wing: initiation of wing vein development." *Current opinion in genetics and development*, 10(4), 393-8.

Birdsall, K., Zimmerman, E., Teeter, K., and Gibson, G. (2000). "Genetic variation for the positioning of wing veins in *Drosophila melanogaster*." *Evolution and development*, 2(1), 16-24.

Bitner-Mathe, B. C., and Klaczko, L. B. (1999). "Heritability, phenotypic and genetic correlations of size and shape of *Drosophila mediopunctata* wings." *Heredity*, 83(Pt 6), 688-96.

Bonic, R. A., Hajian, G. V., Cranford, E., and Krantz, S. (1971). *Freshman Calculus*, DC Heath and Company, Lexington, Massachusetts.

Bookstein, F. L. (1991). *Morphometric tools for landmark data: geometry and biology*, Cambridge University Press, Cambridge, Massachusetts.

Bookstein, F. L. (1996a). "Biometrics, biomathematics and the morphometric synthesis." *Bulletin Mathematical Biology*, 58, 313-365.

Bookstein, F. L. (1996b). "A hundred years of morphometrics." *Acta Zoologia Academiae Scientiarum Hungaricae*, 44(1-2), 7-59.

Botella, J., Kretzchamar, D., Kiremayer, C., Feldmann, D., Hughes, D., and Schneuwly, S. (2003). "Deregulation of the EGFR/Ras signaling pathway induces AGA-related brain degeneration in the *Drosophila* mutant *vap*." *Molecular biology of the cell*, 14(1), 241-250.

Brower, D. L., Piovant, M., and al., e. (1987). "Identification of a specialized extracellular matrix component in *Drosophila* imaginal discs." *Developmental Biology*, 119(2), 373-381.

Bryant, P. J. (1978). "Pattern formation in imaginal discs. In The genetics and biology of

Drosophila." The genetics and biology of *Drosophila*, M. Ashburner and T. R. F. Wright, Eds., Academic press, London, 229-335.

Buckler IV, E. S., and Thornsberry, J. M. (2002). "Plant molecular diversity and application to genomics." *Current opinion in Plant Biology*, 5, 107-111.

Butler, M. J., Jacobsen, T. L., Cain, D. M., Jarman, M. G., Hubank, M., Whittle, J. R. S., Phillips, R., and Simcox, A. (2003). "Discovery of genes with highly restricted expression patterns in the *Drosophila* wing disc using DNA oligonucleotide microarrays." *Development*, 130, 659-670.

Calboli, F. C. F., Gilchrist, G. W., and Partridge, L. (2003). "Different cell size and cell number contribution in two newly established and one ancient body size cline of *Drosophila subobscura*." *Evolution*, 57(3), 566-573.

Caracristi, G., and Schlotterer, C. (2003). "Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles." *Molecular Biology and Evolution*, 20(5), 792-799.

Carpenter, G. (2000). "The EGF receptor: a nexus for trafficking and signaling." *Bioessays*, 22, 697-707.

Carrillo, R., and Gibson, G. (2002). "Unusual genetic architecture of natural variation affecting drug resistance in *Drosophila melanogaster*." *Genetical Research*, 80(3), 205-213.

Carroll, S. B., Grenier, J. K., and Weatherbee, S. D. (2001). *From DNA to diversity, molecular genetics and the evolution of animal design*, Blackwell Science, Inc, Malden.

Celniker, S. E. (2000). "The *Drosophila* genome." *Current Opinion in Genetics and Development*, 10, 612-616.

Clark, A. G., and Wang, L. (1997). "Epistasis in measured genotypes: *Drosophila* P-element insertions." *Genetics*, 147, 157-163.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., and Sing, C. F. (1998). "Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase." *American Journal of Human Genetics*, 63, 595-612.

Clifford, R., and Schüpbach, T. (1989). "Coordinately and differentially mutable activities of *torpedo*, the *Drosophila-melanogaster* homolog of the vertebrate EGF receptor gene." *Genetics*, 123(4), 771-787.

Clifford, R., and Schüpbach, T. (1992). "The *torpedo* (DER) receptor tyrosine kinase is required at multiple times during *Drosophila* embryogenesis." *Development*, 115(3), 853-872.

Clifford, R., and Schüpbach, T. (1994). "Molecular analysis of the *Drosophila* homolog reveals that several genetically defined classes of alleles cluster in subdomains of the receptor protein." *Genetics*, 137(2), 531-550.

Comeron, J. M., and Kreitman, M. (1998). "The correlation between synonymous and

nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints?" *Genetics*, 150, 767-775.

Comstock, J. H., and Needham, J. G. (1898-99). "The wings of insects." *American Naturalist*, 32, 43-903.

Condic, M. L., Fristrom, D., and Fristrom, J. W. (1990). "Apical cell shape changes during *Drosophila* imaginal leg disc elongation: A novel morphogenetic mechanism." *Development*, 111, 23-33.

Cong, B., Liu, J., and Tanksley, S. D. (2002). "Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations." *Proc Natl Acad Sci U S A*, 99(21), 13606-13611.

Cortese, M., Norry, F., Piccinali, R., and Hasson, E. (2002). "Direct and correlated responses to artificial selection on developmental time and wing length in *Drosophila buzzatii*." *Evolution*, 56(12), 2541-2547.

Cowley, D. E., and Atchley, W. R. (1988). "Quantitative genetics of *Drosophila melanogaster*. II. Heritabilities and genetic correlations between sexes for head and thorax traits." *Genetics*, 119, 421-433.

Cowley, D. E., and Atchley, W. R. (1990). "Development and quantitative genetics of correlation structures among body parts of *Drosophila melanogaster*." *American Naturalist*, 135, 242-268.

Coyne, J. A., Barton, N. H., and Turelli, M. (1997). "Perspective: a critique of Sewall Wright's shifting balance theory of evolution." *Evolution*, 51(3), 643-671.

Coyne, J. A., and Beecham, E. (1987). "Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster*." *Genetics*, 117(4), 727-37.

Curtsinger, J. W. (1986). "Quantitative wing variation in inbred and outbred lines of *Drosophila melanogaster*." *Journal of Heredity*, 77(4), 267-71.

de Celis, J. F. (1997). "Expression and function of decapentaplegic and thick veins during the differentiation of the veins in the *Drosophila* wing." *Development*, 124(5), 1007-18.

de Celis, J. F. (1998). "Positioning and differentiation of veins in the *Drosophila* wing." *International Journal of Developmental Biology*, 42, 335-343.

de Lattre, M., and Félix, M.-A. (2001). "Microevolutionary studies in nematodes: a beginning." *Bioessays*, 23, 807-819.

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." *Nature Genetics*, In press.

de Moed, G. H., Jong, G. D., and Scharloo, W. (1997). "Environmental effects on body size in *Drosophila melanogaster* and its cellular basis." *Genetical Research*, 70, 35-43.

David, J. R., and Capy, P. (1988). "Genetic variation in Drosophila melanogaster natural

populations." *Trends in Genetics*, 4(106-111).

Dermitzakis, E. T., Bergman, C. M., and Clark, A. G. (2003). "Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites." *Molecular biology and evolution*, 20(5), 703-714.

Díaz-Benjumea, F. J., Gaitan, M., and Garcia-Bellido, A. (1989). "Developmental genetics of the wing vein pattern of *Drosophila*." *Genome*, 31(2), 612-619.

Díaz-Benjumea, F. J., and García-Bellido, A. (1990). "Genetics analysis of the wing vein pattern of *Drosophila*." *Roux's Archives in Developmental Biology*, 198, 336-354.

Dickinson, M. H., Lehmann, F. O., and Gotz, K. G. (1993). "The active control of wing rotation by *Drosophila*." *Journal of Experimental Biology*, 182, 173-89.

Dickinson, M. H., Lehmann, F. O., and Sane, S. P. (1999). "Wing rotation and the aerodynamic basis of insect flight" *Science*, 284(5422), 1954-60.

Dilda, C. (2002). "The genetic architecture of *Drosophila* sensory bristle number," Ph. D. Thesis, North Carolina State University, Raleigh.

Dryden, I. L., and Mardia, K. V. (1998). *Statistical shape analysis*, J. Wiley and sons, Chichester, New York.

Duchek, P., and Roth, P. (2001). "Guidance of cell migration by EGFR receptor signaling during *Drosophila* oogenesis." *Science*, 291, 131-133.

Dudley, R. (2000). *The biomechanics of insect flight*, Princeton University Press, Princeton, New Jersey.

Duncan, I. W. (2002). "Transvection effects in *Drosophila*." *Annual Review of Genetics*, 36, 521-556.

Dunn, R. C. and C. C. Laurie (1995). *"*Effects of a transposable element insertion on *Alcohol dehydrogenase* expression in *Drosophila melanogaster."* Genetics **140**: 667-677.

Eanes, W. F. (1999). "Analysis of selection on enzyme polymorphisms." *Annual Review of Genetics*, 30, 301-326.

Ennos, A. R. (1989). "Comparative functional morphology of the wings of Diptera." *Zoological journal of the Linnean society*, 96, 27-47.

Fain, M. J., and Stevens, B. (1982). "Alterations in the cell-cycle of *Drosophila* imaginal disk cells precede metamorphosis." *Developmental Biology*, 92(1), 247-258.

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics (4 etd.)*, Longman group Ltd., Essex, England.

Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." *Genetics* **155**(3), 1405-1413.

Fedorowicz, G. M., Fry, J. D., Anholt, R. R., and Mackay, T. F. (1998). "Epistatic interactions between *smell-impaired* loci in *Drosophila melanogaster*." *Genetics*, 148(4), 1885-1891.

Fink, W. L., and Zelditch, M. L. (1995). "Phylogenetic analysis of ontogenetic shape

transformations: a reassessment of the piranha genus Pygocentrus (Teleostei)." *Systematic Biology*, 44, 343-360.

Fisher, R. A. (1930). *The genetical theory of natural selection*, Clanderon, Oxford.

Fletcher, R., and Thompson, J. (2000). "Spatial differences in patterns of modification: selection on *hairy* in *Drosophila melanogaster* wings." *Genetica*, 109, 169-181.

Frary, A., and al., e. (2000). "A quantitative trait locus key to the evolution of tomato fruit size." *Science*, 289, 85-88.

Fraser, H. B., Hirsh, A. E., Steinmetz, L., Scharfe, C., and Feldman, M. W. (2002). "Evolutionary rate in the protein interaction network." *Science*, 296, 750-752.

Fristrom, D., Wilcox, M., and Fristrom, J. W. (1993). "The distribution of ps integrins, laminin-a and f-actin during key stages in *Drosophila* wing development." *Development*, 117(2), 509-523.

Fry, S. N., Sayaman, R., and Dickinson, M. H. (2003). "The aerodynamics of free-flight maneuvers in *Drosophila*." *Science*, 300, 495-498.

Fu, Y. X., and Li, W. H. (1993). "Statistical tests of neutrality of mutations." *Genetics*, 133, 693-709.

Funakoshi, Y., Minami, M., and Tabata, T. (2001). "*mtv* shapes the activity gradient of the Dpp morphogen through regulation of *thickveins*." *Development*, 128(1), 67-74.

Gabay, L., Seger, R., and Shilo, B. Z. (1997). "In situ activation pattern of *Drosophila* EGF receptor pathway during development." *Science*, 277, 1103-1106.

Gadau, J., Page, R. E., and Werren, J. H. (2002). "The genetic basis of the interspecific differences in wing size in *Nasonia* (Hymenoptera; Pteromalidae): Major quantitative trait loci and epistasis." *Genetics*, 161, 673-684.

Galpern, P. (2000). "The use of common principle component analysis in studies of phenotypic evolution: an example from the Drosophilidae," MS Thesis, University of Toronto, Toronto.

García-Bellido, A. (1975). "Genetic control of imaginal disc morphogenesis in *Drosophila*." ICN-UCLA Symposia on Molecular and Cellular Biology "Developmental Biology", D. McMahon and C. Frec Box, Eds., W.A. Benjamin Inc., 40-59.

García-Bellido, A., and de Celis, J. F. (1992). "Developmental Genetics of the venation pattern of *Drosophila*." *Annual Review of Genetics*, 26, 275-302.

Gasperini, R., and Gibson, G. (1999). "Absence of protein polymorphism in the *Ras* genes of *Drosophila melanogaster*." *Journal of Molecular Evolution*, 49(5), 583-90.

Geiger-Thornsberry, G. L., and Mackay, T. F. C. (2002). "Association of single-nucleotide polymorphisms at the *Delta* locus with genotype by environment interaction for sensory bristle number in *Drosophila melanogaster*." *Genetical Research*, 79(3), 211-218.

Gerhart, J., and Kirschner, M. (1997). *Cells, embryos and evolution*, Blackwell Science, Inc,

Malden.

Gibson, G., and van Helden, S. (1997). "Is function of the *Drosophila* homeotic gene Ultrabithorax canalized?" *Genetics*, 147, 1155-1168.

Gibson, G. C., and Palsson, A. (2001). "A complement for developmental genetics." *Current opinion in biology*, 11, R74-6.

Gibson, M. C., and Schubiger, G. (2001). "Peripodial membrane cells regulate imaginal disc development in *Drosophila*." *Developmental Biology*, 235(1), 52.

Gilchrist, S., and Partridge, L. (1999). "A comparison of the genetic basis of wing size divergence in three parallel body size clines of *Drosophila melanogaster*." *Genetics*.

Gockel, J., Kennington, J., Hoffmann, A., Goldstein, D. B., and Partridge, L. (2001). "Nonclinality of molecular variation implicates selection in maintaining a morphological cline of *Drosophila melanogaster*." *Genetics*, 158, 319-323.

Goldstein, D. B. (2001). "Islands of linkage disequilibrium." *Nature Genetics*, 29(109-137).

Gould, S. J. (1977). *Ontogeny and Phylogeny*, Belknap Press of Harvard University Press, Cambridge.

Grant, P., and Grant, R. (2002). "Unpredictable evolution in a 30-year study of Darwin's finches." *Science*, 296, 707-711.

Gurganus, M.C., Nuzhdin, S.V., Leips, J.W., and Mackay T.F.C. (1999) “High-resolution mapping of quantitative trait loci for bristle number in *Drosophila melanogaster”*. Genetics 152: 1585-1604

Harrison, P. J. (2000). "Dopamine and schizophrenia--proof at last?” *Lancet*, 356(9234), 958-9.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*, SAS institute, Cary, NC.

Held JR, L. I. (2002). *Imaginal discs, the genetic and cellular logic of pattern formation*, Cambridge University Press, Cambridge.

Hodgson, J. W., Argiropoulos, B., and Brock, H. W. (2001). "Site-specific recognition of a 70-base-pair element containing d(GA)n repeats mediates bithoraxoid polycomb group response element response element-dependent silencing." *Molecular and Cellular Biology*, 21(14), 4528-4543.

Hogeweg, P. (2000). "Evolving mechanisms of morphogenesis: on the interplay between differential adhesion and cell differentiation." *Journal of Theoretical Biology*, 203(4), 317-33.

Hudson, R. R., Kreitman, M., and Aguade, M. (1987). "A test of neutral molecular evolution based on nucleotide data." *Genetics*, 116, 153-159.

Huey, R. B., Gilchrist, G. W., Carlson, M. L., Berrigan, D., and Serra, L. (2000). "Rapid evolution of a geographic cline in size in an introduced fly." *Science*, 287(5451), 308-309.

Imasheva, A. G., Bubli, O. A., and Lazebny, O. E. (1994). "Variation in wing length in Eurasian

natural populations of *Drosophila melanogaster*." *Heredity*, 72, 508-14.

Imasheva, A. G., Bubli, O. A., Lazebny, O. E., and Zhivotovsky, L. A. (1995). "Geographic differentiation in wing shape in *Drosophila melanogaster*." *Genetica*, 96(3), 303-306.

Ioannidis, J.P.A., Ntzani, E.E., Trikalinos, T.A. and Contopoulos-Ioannidis, D.G. (2001). "Replication validity of genetic association studies." *Nature Genetics*, 29, 306-309.

James, A., Azevedo, R. B. R., and Partridge, L. (1997). "Genetic and environmental response to temperature of *Drosophila melanogaster* from a latitudinal cline." *Genetics*, 146, 881-890.

Jeong, H., Tombor, B., Albert, R., Oltval, Z., and Barabasi, A.-L. (2000). "The large-scale organization of metabolic networks." *Nature*, 407, 651-654.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001). "The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster." *Nature Genetics*, 29, 389 - 395.

Kaminker, J. S., Bergman, C., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Firse, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., and Celniker, S. E. (2002). "The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics persepective." *Genome Biology*, 3(12).

Kammerer, S., Burns-Hamuro, L. L., Ma, Y., Hamon, S. C., Canaves, J. M., Shi, M. M., R., N. M., Sing, C. F., Cantor, C. R., Taylor, S. S., and Braun, A. (2002). "Amino acid variant in the kinase binding domain of dual-specific A kinase-anchoring protein 2: A disease susceptibility polymorphism." *Proc Natl Acad Sci U S A*, 100(7), 4066-4071.

Kammermeyer, K. L., and Wadsworth, S. C. (1987). "Expression of *Drosophila* epidermal growth-factor receptor homolog in mitotic cell-populations." *Development*, 100(2), 201-210.

Kauffman, S. (1993). *On the origins of order*, Oxford University Press, Oxford.

Kerrigan, L. A., Croston, G. E., Lira, L. M., and Kadonaga, J. T. (1991). "Sequence-specific transcriptional anti-repression of the *Drosophila Kruppel* gene by the GAGA factor." *Journal of Biological Chemistry*, 266, 574-582.

Klingenberg, C. P. (2002). "Morphometrics and the role of the phenotype in studies of the evolution of developmental mechanisms." *Gene*, 287(1-2), 3-10.

Klingenberg, C. P., McIntyre, G. S., and Zaklan, S. D. (1998). "Left-right asymmetry of fly wings and the evolution of body axes " *Proc R Soc Lond B Biol Sci*, 265(1402), 1255-9., [erratum appears in Proc R Soc Lond B Biol Sci 1998 Dec 22;265(1413):2455].

Klingenberg, C. P., and Nijhout, H. F. (1999). "Genetics of fluctuating asymmetry: a developmental model of developmental instability." *Evolution*, 53(2), 358-375.

Klingenberg, C. P., and Zaklan, S. D. (2000). "Morphological integration between developmental compartments in the *Drosophila* wing." *Evolution*, 54(4).

Kopp, A., and True, J. (2002). "Evolution of male sexual characters in the Oriental *Drosophila melanogaster* species group." *Evolution and Development*, 4, 278-291.

Kreitman, M., and Hudson, R. R. (1991). "Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphim and divergence." *Genetics*, 127, 565-582.

Kwok, P.-Y. (2001). "Genetic association by whole-genome analysis?" *Science*, 294, 1669-1670.

Kyriacou, C. P., and Hall, J. C. (1980). "Circadian rhythm mutations in *Drosophila melanogaster* affect short-term fluctuations in the male's courtship song." *Proc Natl Acad Sci U S A*, 77(11), 6729-33.

Lage, z. (1997). "Requirement for *EGF receptor* signalling in neural recruitment during formation of *Drosophila* chordotonal sense organ clusters." *Current Biology*, 7(3), 166-175.

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 of *Drosophila melanogaster*." *Science*, 266(266).

Lai, C., McMahon, R., Young, C., Mackay, T. F. C., and Langley, C. H. (1998). "*quamao*, a *Drosophila* bristle locus, encodes a geranylgeranyl pyrophosphate synthase." *Genetics*, 149, 1051-1061.

Langley, C. H., and Crow, J. F. (1974). "The direction of Linkage disequilibrium." *Genetics*, 78, 937-941.

Laurie, C. C., Bridgeham, J. T., and Choudhary, M. (1991). "Association between DNA sequence variation and variation in expression of the *Adh* gene in natural populations of *Drosophila melanogaster*." *Genetics*, 129, 489-499.

Leamy, L. J., Routman, E. J., and Cheverud, J. M. (1998). "Quantitative trait loci for fluctuating asymmetry of discrete skeletal characters in mice." *Heredity*, 80(Pt 4), 509-18.

Lesokhin, A. M., Yu, S.-Y., Katz, J., and Baker, N. E. (1999). "Several levels of EGF receptor signaling during photoreceptor specification in wild-type, Ellipse and null mutant *Drosophila*." *Developmental Biology*, 205, 129-144.

Lessard, S. (1997). "Fisher's fundamental theorem of natural selection revisited." *Theoretical population biology*, 52, 119-136.

Lev, Z., Shilo, B. Z., and Kimchie, Z. (1985). "Developmental changes in expression of the *Drosophila melanogaster* epidermal growth factor receptor gene." *Developmental Biology*, 110, 499-502.

Levy-Lahad, E., Lahad, A., Eisenberg, S., Dagan, E., Paperna, T., Kasinetz, L., Catane, R., Kaufman, B., Beller, U., Renbaum, P., and Gershoni-Baruch, R. (2001). "A single nucleotide polymorphism in the RAD51 gene modifies cancer risk in BRCA2 but not BRCA1 carriers." *Proc Natl Acad Sci U S A.*, 98(6), 3232-3236.

Lewin, B. (1997). *Genes 6*, Oxford University Press, Oxford.

Lewontin, R. C. (1988). "On measures of gametic disequilibrium." *Genetics*, 120(3), 849-852.

Lewontin, R. C. (1995). "The detection of linkage disequilibrium in molecular sequence data." *Genetics*, 140(1), 377-388.

Liu, J., Mercer, J. M., Stam, L. F., Gibson, G. C., Zeng, Z. B., and Laurie, C. C. (1996). "Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*." *Genetics*, 142(4), 1129-45.

Long, A. D., and Langley, C. H. (1999). "Power of association studies to detect the contribution of candidate genetic loci to complexly inherited phenotypes." *Genome Research*, 9, 720-731.

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(2), 999-1017.

Long, A. D., Lyman, R. F., Morgan, A. H., Langley, C. H., and Mackay, T. F. C. (2000). "Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the *achaete-scute* complex are associated with variation in bristle number in *Drosophila melanogaster*." *Genetics*, 154(3), 1255-1269`.

Long, A. D., Mullaney, S. L., Mackay, T. F., and Langley, C. H. (1996). "Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*." *Genetics*, 144(4), 1497-510.

Ludwig, M. Z. (2002). "Functional evolution of noncoding DNA." *Current Opinion in Genetics and Development*, 12, 634-639.

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). "Evidence for stabilizing selection in a eukaryotic enhancer element." *Nature*, 403(6769), 564-7.

Ludwig, M. Z., Patel, N. H., and Kreitman, M. (1998). "Functional analysis of *eve* stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change." *Development*, 125(5), 949-58.

Lycett, G., Blass, C., and Louis, C. (2001). "Developmental variation in epidermal growth factor receptor size and localization in the malaria mosquito, *Anopheles gambiae*." *Insect molecular biology*, 10(6), 619-628.

Lyman, R. F., Lai, C. Q., 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*." *Genetical Research*, 74(3), 303-311.

Lyman, R. F., Lawrence, F., Nuzhdin, S. V., and Mackay, T. F. (1996). "Effects of single P-element insertions on bristle number and viability in *Drosophila melanogaster*."

Genetics, 143, 277-292.

Lyman, R. F., and Mackay, T. F. (1998). "Candidate quantitative trait loci and naturally occurring phenotypic variation for bristle number in *Drosophila melanogaster*: the *Delta-Hairless* gene region." *Genetics*, 149, 983-998.

Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits*, Sinauer associates, Inc publishers, Sunderland, Massachusetts.

Mackay, T. F. (1995). "The genetic basis of quantitative variation: numbers of sensory bristles of *Drosophila melanogaster* as a model system." *Trends in Genetics*, 11(12), 464-470.

Mackay, T. F. (1996). "The nature of quantitative genetic variation revisited: lessons from *Drosophila* setae." *Bioessays*, 18, 113-121.

Mackay, T. F. C. (2001). "The genetic architecture of quantitative traits." *Annual Review of Genetics*, 35, 303-339.

Mackay, T. F. C., and Fry, J. D. (1996). "Polygenic mutation in *Drosophila melanogaster*: Genetic interactions between selection lines and candidate quantitative trait loci." *Genetics*, 144(2), 671-688.

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.

Mahmoudi, T., Katsani, K. R., and Verrijzer, C. P. (2002). "GAGA can mediate enhancer function in trans by linking two separate DNA molecules." *EMBO journal*, 21(7), 1775-1781.

Marcus, J. M. (2001). "The development and evolution of crossveins in insect wings." *Journal of Anatomy*, 199(1-2), 211-216.

Martin-Blanco, E., Roch, F., Noll, E., Baonza, A., Duffy, J. B., and Perrimon, N. (1999). "A temporal switch in *DER* signaling controls the specification and differentiation of veins and interveins in the *Drosophila* wing." *Development*, 126, 5739-5747.

Mauricio, R. (2001). "Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology." *Nature reviews genetics*, 2, 370-381.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I. (2000). "Vista: visualizing global DNA sequence alignments of arbitrary length." *Bioinformatics*, 16(11), 1046-1047.

McAdams, H., and Arkin, A. (1997). "Stochastic mechanisms in gene expression." *Proc Natl Acad Sci U S A*, 94, 814-819.

McDonald, M., and Kreitman, M. (1991). "Adaptive protein evolution at the *Adh* locus in *Drosophila*." *Nature*, 351, 652-654.

McMillan, W. O., Monteiro, A., and Kapan, D. D. (2002). "Development and evolution on the wing." *Trends in Ecology and Evolution*, 17(3), 126-133.

Meir, E. (1997). "Building a 3-d gene expression atlas for early *Drosophila* embryos."

Developmental Biology, 186(2), B266-B266.

Meir, E., von Dassow, G., Munro, E., and Odell, G. M. (2002). "Robustness, flexibility, and the role of lateral inhibition in the neurogenic network." *Current Biology*, 12(10), 778-.

Mezey, J., Cheverud, J. M., and Wagner, G. P. (2000). "Is the genotype-phenotype map modular?: A statistical approach using mouse quantitative trait loci." *Genetics*, 156, 305-311.

Milner, M. J., Bleasby, A. J., and Pyott, A. (1983). "The role of the peripodial membrane in the morphogenesis of the eye-antennal disc of *Drosophila melanogaster*." *Wilhelm Roux's Archive of Developmental biology*, 192, 164-170.

Montange, J., Groppe, J., Guillemin, K., Krasnow, M. A., Gehring, W. J., and Affolter, M. (1996). "The *Drosophila* serum response factor gene is required for the formation of intervein tissue of the wing and is allelic to *blistered*." *Development*, 122, 2589-2597.

Moreno, E., Basler, K., and Morata, G. (2002). "Cells compete for decepentaplegic survival factor to prevent apoptosis in *Drosophila* wing development." *Nature*, 416, 755-759.

Moreteau, B., Capy, P., Alonso-Moraga, A., Munoz-Serrano, A., Stockel, J., and David, J. R. (1995). "Genetic characterization of geographic populations using morphometrical traits in *Drosophila melanogaster*: isogroups versus isofemale lines." *Genetica*, 96(3), 207-15.

Neufeld, T. P., de la Cruz, A. F., Johnston, L. A., and Edgar, B. A. (1998). "Coordination of growth and cell division in the *Drosophila* wing." *Cell*, 93, 1183-1193.

Nicholas, K. B., Nicholas, H. B. J., and Deerfield, D. W. I. (1997). "GeneDoc: Analysis and Visualization of Genetic Variation." *EMBO news*, 4, 14.

Nijhout, H. F. (2002). "The nature of robustness in development." *Bioessays*, 24, 553-563.

Nuzhdin, S. V., Pasyukova, E. G., Dilda, C., Zeng, Z. B., and Mackay, T. F. (1997). "Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster*." *Proc Natl Acad Sci U S A*, 94, 9734-9739.

O'Brien, T., Wilkins, R. C., Giardina, C., and Lis, J. T. (1995). "Distribution of GAGA protein on Drosophila genes in vivo." *Genes and Development*, 9, 1098-1110.

Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J.-H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., and Yokohama, S. (2002). "Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains." *Cell*, 110(775-787).

O'Keefe, D. D., and Thomas, J. B. (2001). "*Drosophila* wing development in the absence of dorsal identity." *Development*, 128(5), 703-710.

Olayioye, M. A., Neve, R. M., Lane, H. A., and Hynes, N. E. (2000). "The ErbB signaling network: receptor heterodimerization in development and cancer." *EMBO journal*, 19(13), 3159-3167.

Olsen, K., Womack, A., Garrett, A., et al. (2002). "Contrasting evolutionary forces in the

Arabidopsis thaliana floral developmental pathway." *Genetics*, 160, 1641-1650.

OMIM (2003). "Online Mendelian Inheritance in Man." http://www.ncbi.nlm.nih.gov.

Ondek, B., Gloss, L., and Herr, W. (1988). "The SV40 enhancer contains two distinct levels of organization." *Nature*, 333(6168), 40-45.

Palsson, A., and Gibson, G. (2000). "Quantitative developmental genetic analysis reveals that the ancestral dipteran wing vein prepattern is conserved in *Drosophila melanogaster*." *Development Genes and Evolution*, 210(12), 617-22.

Pasyukova, E. G., Vieira, C., and Mackay, T. F. (2000). "Deficiency mapping of quantitative trait loci affecting longevity in *Drosophila melanogaster*." *Genetics*, 156(3), 1129-46.

Polaczyk, P. J., Gasperini, R., and Gibson, G. (1998). "Naturally occurring genetic variation affects *Drosophila* photoreceptor determination." *Development, Genes and Evolution*, 207(7), 462-470.

Powell, J. R. (1996). *Progress and prospects in evolutionary biology: the Drosophila model*, Oxford University Press, New York.

Powell, J. R., and DeSalle, R. (1995). "*Drosophila* molecular phylogenies and their uses." Evolutionary Biology, M. K. Hecht, R. J. Macintyre, and M. T. Clegg, eds., Plenum Press, New York, 87-138.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex disease?" *American Journal of Human Genetics*, 69, 124-137.

Pritchard, J. K., and Przeworski, M. (2001). "Linkage disequilibrium in humans: models and data." *American Journal of Human Genetics*, 69, 1-14.

Pritchard, J. K., and Rosenberg, N. A. (1999). "Use of unlinked genetic markers to detect population stratification in association studies." *American Journal of Human Genetics*, 65(1), 220-228.

Prober, D. A., and Edgar, B. A. (2000). "Ras1 promotes cellular growth in the *Drosophila* wing." *Cell*, 100(4), 435-46.

Przeworski, M. (2002). "The signature of positive selection at randomly chosen loci." *Genetics*, 160, 1179-1189.

Raff, R. A. (1996). *The shape of life*, The university of Chicago press, Chicago.

Ramirez-Weber, F. A., and Kornberg, T. B. (1999). "Cytonemes: Cellular processes that project to the principal signaling center in *Drosophila* imaginal discs." *Cell*, 97(5), 599-607.

Rand, D. M., and Kann, L. M. (1996). "Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice and humans." *Molecular Biology and Evolution*, 13(6), 735-748.

Raz, E., Schejter, E. D., and Shilo, B. Z. (1991). "Interallelic complementation among *DER/flb* alleles: implications for the mechanism of signal transduction by receptor tyrosine kinases." *Genetics*, 129, 191-201.

Rebay, I. (2002). "Keeping the receptor tyrosine kinase signaling pathway in check: lessons from *Drosophila*." *Developmental Biology*, 251, 1-17.

Remington, D. L., Ungerer, M. C., and Purugganan, M. D. (2001a). "Map-based cloning of quantitative trait loci: progress and prospects." *Genetical Research*, 78(3), 213-218.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler IV, E. S. (2001b). "Structure of linkage disequilibrium and phenotypic associations in the maize genome." *Proc Natl Acad Sci U S A*, 98(20), 11479-11484.

Resino, J., Salama-Cohen, P., and Garcia-Bellido, A. (2002). "Determining the role of patterned cell proliferation in the shape and size of the *Drosophila* wing." *Proc Natl Acad Sci U S A*, 99(11), 7502-7507.

Richter, B., Long, M., Lewontin, R. C., and Nitasaka, E. (1997). "Nucleotide variation and conservation at the dpp locus, a gene controlling early development in Drosophila." *Genetics*, 145(2), 311-323.

Rifkin, S. A., Kim, J.-H., and White, K. P. (2003). "Evolution of gene expression in the *Drosophila melanogaster* subgroup." *Nature Genetics*, 33, 138-144.

Riley, R. M., Jin, W., and Gibson, G. (2003). "Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*." *Molecular Ecology*, 12(5), 1315-1323.

Risch, N., and Merikangas, K. (1996). "The future of genetic studies of complex disease." *Science*, 273, 1516-1517.

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

Rockman, M. V., and Wray, G. A. (2002). "Abundant raw material for cis-regulatory evolution in humans." *Molecular Biology and Evolution*, 19(11), 1991-2004.

Roff, D. A. (1997). *Evolutionary Quantitative Genetics*, Kluwer Academic Publishers.

Rohlf, F. J. (2002). "TPS relative warp analysis software."SUNY, Stony Brook.

Rohlf, J. F. (1996). "Morphometric spaces, shape components and the effects of linear transformations." NATO series: Advances in Morphometrics, L. F. Marcus Eds., Plenum Press, New York.

Rong, Y. S., and Golic, K. G. (2000). "Gene targeting by homologous recombination in *Drosophila*." *Science*, 288(5473), 2013-8.

Rozas, J., and Rozas, R. (1999). "DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis." *Bioinformatics*, 15, 174-175.

Salazar-Ciudad, I., Garcia-Fernandez, J., and Sole, R. V. (2000). "Gene networks capable of pattern formation: from inductio to reaction diffusion." *Journal of Theoretical Biology*, 205, 587-603.

Salazar-Ciudad, I., Newman, S. A., and Sole, R. V. (2001). "Phenotypic and dynamical transitions in model genetic networks I. Emergence of patterns and genotype-phenotype relationships." *Evolution and Development*, 3(2), 84-94.

Salazar-Ciudad, I., Sole, R. V., and Newman, S. A. (2001). "Phenotypic and dynamical transitions in model genetic networks II. Application to the evolution of segmentation mechanisms." *Evolution and Development*, 3(2), 95-103.

SAS version 8.02 (2002). "SAS." SAS institute, Cary NC.

Sawamura, K., Davis, A. W., and Wu, C.-I. (2000). "Genetic analysis of speciation by means of introgression into *Drosophila melanogaster*." *Proc Natl Acad Sci U S A*, 97(6), 2652-2655.

Schejter, E. D., Segal, D., Glazer, L., and Shilo, B. Z. (1986). "Alternative 5' exons and tissue-specific expression of the *Drosophila EGFR receptor* homolog transcripts." *Cell*, 46(7), 1091-1101.

Scherf, U., and al., e. (2000). "A gene expression database for the molecular pharmacology of cancer." *Nature Genetics*, 24(236-244).

Scion Image version 4.0.2. (1998-2002). "Scion Image for windows." Scion Corporation.

Schlötterer, C., and Harr, B. (2002). "Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*." *Molecular Ecology*, 11, 947-950.

Schneider, S., Roessli, D. and Excoffier, L. (2000). Arlequin: A software for population genetics data analysis. Geneva, Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva. Version 2.000.

Schüpbach, T., and Wieschaus, E. (1998). "Probing for gene specificity in epithelial development." *International Journal of Developmental Biology*, 42(3), 249-55.

Schweitzer, R., Howes, R., Smith, R., Shilo, B. Z., and Freeman, M. (1995). "Inhibition of *Drosophila* EGF receptor activation by the secreted protein Argos." *Nature*, 376(6542), 699-702.

Schwendemann, A., and Lehmann, M. (2002). "Pipsqueak and GAGA factor act in concert as partners at homeotic and many other loci." *Proc Natl Acad Sci U S A*, 99(20), 12883-12888.

Shastry, B. S. (1999). "Recent developments in the genetics of schizophrenia." *Neurogenetics*, 2(3), 149-54.

Shilo, B. Z. (2003). "Signaling by the *Drosophila* epidermal growth factor receptor pathway during development." *Experimental cell research*, 283.

Shimomura, K., Low-Zeddies, S. S., King, D. P., and al., e. (2001). "Genome wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in Mice." *Genome Research*, 11, 959-980.

Simcox, A. (1997). "Differential requirement for EGF-like ligands in *Drosophila* wing development." *Mechanical Development*, 62(1), 41-50.

Simon, M. A. (2000). "Receptor tyrosine kinases: specific outcomes from general signals." *Cell*, 103, 13-15.

Slatkin, M. (1985). "Rare alleles as indicators of gene flow." *Evolution*, 39, 53-65.

Small, S., Blair, A., and Levine, M. (1992). "Regulation of the *even-skipped* stripe 2 in the *Drosophila* embryo." *EMBO journal*, 11, 4047-4057.

Sokal, R., and J., R. F. (1995). *Biometry*, W. H. Freeman and Company, New York.

Spradling, A. C., Stern, D., Beaton, A., Rhem, E. J., Laverty, T., Mozden, N., Misra, S., and Rubin, G. M. (1999). "The Berkeley *Drosophila* Genome Project gene disruption project: Single P-element insertions mutating 25% of vital *Drosophila* genes." *Genetics*, 153, 135-177.

Stam, L. F., and Laurie, C. C. (1996). "Molecular dissection of a major gene effect on a quantitative trait: The level of Alcohol Dehydrogenase expression in *Drosophila melanogaster*." *Genetics*, 144(4), 1559-1564.

Stark, J., Bonacum, J., Remsen, J., and DeSalle, R. (1999). "The evolution and development of Dipteran wing veins: a systematic approach." *Annual Review of Entomology*, 44, 97-129.

Steinmetz, L., Sinha, H., Richards, D., Spiegelman, J., Oefner, P., McCusker, J., and Davis, R. (2002). "Dissecting the architecture of a quantitative trait locus in yeast." *Nature*, 416, 326-330.

Steppan, S. J., Philips, P. C., and Houle, D. (2002). "Comparative quantitative genetics: the evolution of the G matrix." *Trends in Ecology and Evolution*, 1-8.

Stoll, M., and al, e. (2001). "A genomic systems biology map for cardiovascular function." *Science*, 294, 1723-1726.

Su, T. T., and O'Farrell, P. H. (1998). "Size control: cell proliferation does not equal growth." *Current Biology*, 8(19), R687-9.

Sucena, E., and Stern, D. L. (2000). "Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*." *Proc Natl Acad Sci U S A*, 97(9), 4530-4534.

Syvanen, A.-C. (2001). "Accessing genetic variation: genotyping single nucleotide polymorphisms." *Nature Reviews Genetics*, 2, 930-942.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." *Genetics*, 123, 585-595.

Teeter, K., Naeemuddin, M., Gasperini, R., Zimmerman, E., White, K. P., Hoskins, R., and Gibson, G. (2000). "Haplotype dimorphism in a SNP collection from *Drosophila melanogaster*." *Journal Experimental Zoology*, 288(1), 63-75.

Thompson, D. W. (1961). *On growth and form*, Cambridge University Press, Cambridge.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice." *Nucleic Acids Research*, 22, 4673-4680.

Thompson, J. N., Jr. (1974). "Studies on the nature and function of polygenic loci in *Drosophila*. II. The subthreshold wing vein pattern revealed in selection experiments." *Heredity*, 33(3), 389-401.

Thompson, J. N., Jr. (1974). "Studies on the nature and function of polygenic loci in *Drosophila*. I. Comparison of genomes from selection lines." *Heredity*, 33(3), 373-87.

Thompson, J. N. (1975). "A test of the influence of isoallelic variation upon a quantitative character." *Heredity*, 35, 401-406.

Thompson, J. N. J., Toney, J. V., and Schaefer, G. B. (1980). "Pattern compensation in *Drosophila* wing vein development." *Heredity*, 44, 93-102.

Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M., and Bergelson, J. (2003). "Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*." *Nature*, 423(6935), 74-77.

Tiong, S., Nash, D., and Bender, W. (1995). "*Dorsal wing*, a locus that affects dorsoventral wing patterning in *Drosophila*." *Development*, 121(6), 1649-1656.

True, J. R., Edwards, K. A., Yamamoto, D., and al., e. (1999). "*Drosophila* wing melanin patterns form by vein-dependent elaboration of enzymatic prepatterns." *Current Biology*, 9(23), 1382-1391.

True, J. R., Weir, B. S., and Laurie, C. C. (1996). "A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*." *Genetics*, 142(3), 819-37.

Tsukiyama, T., Becker, P. B., and Wu, C. (1994). "ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor." *Nature*, 367, 525-532.

Ungerer, M. C., Halldorsdottir, S. S., Modliszewski, J. L., Mackay, T. F. C., and Purugganan, M. D. (2002). "Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*." *Genetics*, 160(3), 1133-1151.

Vincent, J.P. (1998). "Compartment boundaries: where, why and how?" *International Journal of Developmental Biology*, 42, 311-315.

Vogelstein, B., Lane, D., and Levine, A. J. (2000). "Surfing the p53 network." *Nature*, 408(6810), 307-310.

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). "The segment polarity network is a robust developmental module." *Nature*, 406(6792), 188-92.

Waddington, C. H. (1940). "The genetic control of wing development in *Drosophila*." *Journal of*

Genetics, 41, 75-139.

Waddington, C. H. (1957). *The strategy of the genes, a discussion of some aspects of theoretical biology*, George Allen & Unwin Ltd., Bristol.

Wang, S.-H., Simcox, A., and Campbell, G. (2000). "Dual role for *Drosophila* epidermal growth factor receptor signaling in early wing disc development." *Genes and Development*, 14, 2271-2276.

Wang, W., Thornton, K., Berry, A., and Long, M. Y. (2002). "Nucleotide variation along the *Drosophila melanogaster* fourth chromosome." *Science*, 295(5552), 134-137.

Wasserman, J. D., Urban, S., and Freeman, M. (2000). "A family of *rhomboid*-like genes: *Drosophila rhomboid-1* and *roughoid/rhomboid-3* cooperate to activate EGF receptor signaling." *Genes and Development*, 14, 1651-1663.

Wayne, R. K., and Simonsen, K. L. (1998). "Statistical tests of neutrality in the age of weak selection." *Trends in Ecology and Evolution*, 13(6), 236-240.

Weber, K., Eisman, R., Higgins, S., Morey, L., Patty, A., Tausek, M., and Zeng, Z. (2001). "An analysis of polygenes affecting wing shape on chromosome 2 in *Drosophila melanogaster*." *Genetics*, 159, 1045-1057.

Weber, K. E. (1990a). "Selection on wing allometry in *Drosophila melanogaster*." *Genetics*, 126(4), 975-89.

Weber, K. E. (1990b). "Increased selection response in larger populations. I. Selection for wing-tip height in *Drosophila melanogaster* at three population sizes." *Genetics*, 125(3), 579-84.

Weber, K. E. (1992). "How small are the smallest selectable domains of form?" *Genetics*, 130(2), 345-353.

Weber, K. E., Eisman, R., Morey, L., Patty, A., Sparks, J., Tausek, M., and Zeng, Z. B. (1999). "An analysis of polygenes affecting wing shape on chromosome 3 in *Drosophila melanogaster*." *Genetics*, 153, 773-786.

Weigmann, K., Cohen, S. M., and Lehner, C. F. (1997). "Cell cycle progression, growth and patterning in imaginal discs despite inhibition of cell division after inactivation of *Drosophila* Cdc2 kinase." *Development*, 124(3555-3563).

Weir, B. S. (1996). *Genetic data analysis II*, Sinauer Associates Inc. Publishers, Sunderland Massachusetts.

Weir, B. S., and Hill, W. G. (2002). "Estimating F-statistics." *Annual Review of Genetics*, 36, 721-750.

White, K. P., Rifkin, S. A., Hurban, P., and Hogness, D. S. (1999). "Microarray analysis of *Drosophila* development during metamorphosis." *Science*, 286(5447), 2179-2184.

Wilkins, R. C., and Lis, J. T. (1998). "GAGA factor binding to DNA via a single trinucleotide sequence element." *Nucleic Acids Research*, 26, 2672-2678.

Woods, R. E., Sgro, C., Hercus, M., and Hoffmann, A. (1999). "The association between fluctuating asymmetry, trait variability, trait heritability, and stress: a multiply replicated experiment on combined stresses in *Drosophila melanogaster*." *Evolution*, 53(2), 493-505.

Wright, S. (1969). *Evolution and the genetics of populations*. Vol. 2 of 4, Chicago. University of Chicago Press

Yang, H. P., and Nuzhdin, S. V. (2003). "Fitness costs of Doc expression are insufficient to stabilize its copy number in Drosophila melanogaster." *Molecular Biology and Evolution*, 20(5), 800-804.

Yu, K., Sturtevant, M. A., Biehs, B., Francois, V., Padgett, R. W., Blackman, R. K., and al., e. (1996). "The Drosophila *decapentaplegic* and *short gastrulation* genes function antagonistically during adult wing vein development." *Development*, 122, 4033-4044.

Zapata, C., Alvarez, G., Rodqiguez-Trelles, F., and Maside, X. (2000). "A long-term study on seasonal changes of gametic disequilibrium between allozymes and inversions in *Drosophila* subobscura." *Evolution*, 54(5), 1673-1679.

Zecca, M., and Struhl, G. (2002). "Subdivision of the *Drosophila* wing imaginal disc by EGFR-mediated signaling." *Development*, 129, 1357-1368.

Zeng, Z.-B. (1994). "Precision mapping of quantitative trait loci." *Genetics*, 136, 1457-1468.

Zeng, Z. B., Liu, J., Stam, L. F., Kao, C. H., Mercer, J. M., and Laurie, C. C. (2000). "Genetic architecture of a morphological shape difference between two *Drosophila* species." *Genetics*, 154(1), 299-310.

Zimmerman, E., Palsson, A., and Gibson, G. (2000). "Quantitative trait loci affecting components of wing shape in *Drosophila melanogaster*." *Genetics*, 155(2), 671-83.

Zurovcova, M., and Ayala, F. J. (2002). "Polymorphism patterns in two tightly linked developmental genes, *Idgf1* and *Idgf2*, of *Drosophila melanogaster*." *Genetics*, 162, 177-188.

Appendices

Appendix A. List of primers used in PCR and sequencing of *EGFR*.

Fragment	Name	Direction	Location	Primer
1	S1	Forward	5377	TGCCCGTGTTTCAGTTTCCCAA
1	1A-i	Forward	5456	CGGGCAAATTAACATCGGGT
1	1ib	Reverse	6277	GAAGACATATTAGTGCAC
1	1B	Reverse	6502	TCGGTATCTGTCGGATGCT
1	an1	Reverse	6584	AATCTTTGTCCACAGCAGCCCCCTT
2	Y2	Forward	30079	AAAGCCTTCGGACGACTCTTGTGG
2	S2	Forward	30099	GTGGCTCGTAATGTGAAACT
2	2B-i	Reverse	31419	AAGAGGTGAGCCACAGGGCA
2	an2	Reverse	32188	TCTCCCGTCTCCCATTA
3	3x	Forward	35266	GAGACTTTCACCAGCGG
3	S3	Forward	35319	TCCAAACTCACAAGATAGCC
3	3y	Forward	35973	CATCTTGGTGAAAAGGC
3	i3	Forward	36049	TCTGGAATGCGGTTGCCTAT
3	3Ai	Forward	37041	CTGTCTTTGGTTCGTTCCTCTTC
3	an3	Reverse	37218	AGCGGTGAGTTGGAGTTAGA
3	3A	Forward	37362	CTACGCGAGCGGCTAAAAC
3	3nA	Reverse	38171	ACTTCTCCTCCTCCACGG
3	3Bi	Reverse	38378	CGTGGCACTTGGGACACTCG
4	4x	Forward	38905	GGATGTCTATGCCAACTACAC
3	3B	Reverse	39000	TACCCGGTGATCTCCTTC
4	4Ai	Forward	39556	TTCAGTGCCTACAAGTTTGA
5	S5	Forward	40623	GATTCCAGTGGCCATTAAGG
4	4z	Reverse	40661	CCTGTGGACTTGAGCA
4	4Bi	Reverse	40727	AGATTAACGTGCTCCACAGA
5	5ci	Forward	41210	AGCCGGAGATTTGTTCGCT
5	i5	Reverse	42108	TCTAAGATAACAGCCAGCAAAG
5	5Ai	Reverse	42120	CACGGGCTCCTATCTAAG
5	an5	Reverse	42830	TTCAGTAGGCATAACTTGGC

Fragment, refers to the PCR products, 1-5. Region 3 was amplified in two parts, using an3 and 3A as PCR primers.
Location in reference to Genebank record 17571116 and flybase record FBgn0003731.

Appendix B. Descriptive statistics of inbred lines by population and sex.

Trait	Pop	Sex	Mean	Std	Skew	Kurt
B1	UC	F	-0.0072	0.0168	0.2848	-0.2812
	UC	M	0.0055	0.0163	0.3553	-0.1198
	WE	F	-0.0059	0.0152	0.2135	0.2986
	WE	M	0.0067	0.0156	0.1960	0.3986
B2	UC	F	0.0005	0.0112	-0.4409	0.5939
	UC	M	-0.0049	0.0105	-0.4897	1.2524
	WE	F	0.0043	0.0104	-0.1791	0.1241
	WE	M	-0.0019	0.0099	-0.0720	0.0916
B3	UC	F	0.0029	0.0082	0.3360	-0.2405
	UC	M	-0.0031	0.0083	0.3843	-0.1398
	WE	F	0.0036	0.0074	-0.0014	0.1394
	WE	M	-0.0034	0.0076	0.0736	0.4722
C1	UC	F	-0.0011	0.0157	-0.2221	-0.0480
	UC	M	-0.0035	0.0159	-0.2330	-0.0533
	WE	F	0.0019	0.0143	0.0588	0.0186
	WE	M	0.0006	0.0143	-0.0457	0.0867
C2	UC	F	-0.0030	0.0072	0.3287	0.5209
	UC	M	0.0023	0.0079	0.2504	0.2037
	WE	F	-0.0026	0.0072	0.0945	0.2213
	WE	M	0.0030	0.0076	0.1759	0.6204
C3	UC	F	0.0017	0.0057	0.1428	0.0461
	UC	M	-0.0017	0.0060	0.1734	0.1018
	WE	F	0.0017	0.0057	-0.0670	0.2299
	WE	M	-0.0017	0.0056	-0.1265	0.5843
D1	UC	F	0.0055	0.0224	0.2340	0.1163
	UC	M	-0.0063	0.0224	0.2635	-0.0696
	WE	F	0.0055	0.0226	0.2938	0.8187
	WE	M	-0.0050	0.0208	-0.0473	0.0852
D2	UC	F	-0.0022	0.0144	0.0944	0.3517
	UC	M	0.0004	0.0140	0.0134	0.5591
	WE	F	-0.0012	0.0136	-0.0121	0.4171
	WE	M	0.0022	0.0134	0.0679	0.2378
D3	UC	F	-0.0012	0.0130	0.0289	0.6927
	UC	M	0.0004	0.0115	-0.0727	0.5887
	WE	F	-0.0003	0.0130	-0.2267	0.3845
	WE	M	0.0007	0.0127	-0.3377	0.7700
W1	UC	F	0.0043	0.0175	-0.0701	-0.3838
	UC	M	-0.0070	0.0175	-0.2341	-0.0011
	WE	F	0.0064	0.0157	-0.1549	0.1484
	WE	M	-0.0049	0.0156	-0.2391	0.2001
W2	UC	F	0.0042	0.0156	0.0502	0.2186
	UC	M	0.0020	0.0152	0.1074	0.3790
	WE	F	-0.0002	0.0143	-0.1667	0.1301
	WE	M	-0.0032	0.0146	-0.2944	0.1285

Appendix B. (Continued)

Trait	Pop	Sex	Mean	Std	Skew	Kurt
W3	UC	F	-0.0041	0.0110	0.3093	0.0679
	UC	M	0.0041	0.0111	0.3790	0.2735
	WE	F	-0.0039	0.0111	0.2495	-0.0096
	WE	M	0.0039	0.0107	0.2680	0.0476
W4	UC	F	0.0019	0.0097	-0.0054	-0.0710
	UC	M	-0.0026	0.0097	0.1848	-0.1268
	WE	F	0.0028	0.0098	0.1918	0.2085
	WE	M	-0.0023	0.0094	0.0033	0.0373
W5	UC	F	0.0007	0.0097	0.3289	0.6753
	UC	M	0.0028	0.0092	0.3447	0.9144
	WE	F	-0.0027	0.0084	0.2101	0.0767
	WE	M	0.0007	0.0079	0.2297	0.1435
W6	UC	F	-0.0034	0.0076	-0.2256	-0.1026
	UC	M	0.0022	0.0073	-0.2438	-0.1701
	WE	F	-0.0027	0.0078	0.1223	0.2136
	WE	M	0.0034	0.0075	0.1037	0.6884
W7	UC	F	0.0001	0.0059	-0.1044	0.6306
	UC	M	0.0001	0.0055	-0.0127	0.4413
	WE	F	-0.0001	0.0059	0.0354	0.4211
	WE	M	-0.0001	0.0058	0.0370	0.9031
W8	UC	F	0.0005	0.0043	0.1674	0.2414
	UC	M	-0.0003	0.0046	0.0904	0.3265
	WE	F	0.0001	0.0042	-0.0592	-0.0284
	WE	M	-0.0002	0.0044	-0.0851	0.1356
W9	UC	F	0.0004	0.0043	-0.3997	0.4171
	UC	M	-0.0006	0.0042	-0.3550	0.1288
	WE	F	0.0004	0.0038	-0.1569	0.2871
	WE	M	-0.0003	0.0036	-0.2379	0.3166
T Area	UC	F	21.1296	1.9249	-0.1087	0.3309
	UC	M	16.7303	1.6640	0.0786	0.1802
	WE	F	20.6487	1.7967	-0.4022	0.2447
	WE	M	16.7084	1.4653	-0.4507	0.5385
B Area	UC	F	8.4196	0.7780	-0.1633	0.0656
	UC	M	6.7967	0.6857	-0.0719	-0.0959
	WE	F	8.2677	0.7267	-0.3068	0.2844
	WE	M	6.8244	0.6050	-0.3557	0.5722
C Area	UC	F	6.4194	0.6887	0.0671	0.0508
	UC	M	5.0801	0.5979	0.1907	-0.0766
	WE	F	6.2871	0.6326	-0.2725	0.0445
	WE	M	5.0869	0.5155	-0.2820	0.1678
D Area	UC	F	6.2907	0.6021	-0.0488	0.7302
	UC	M	4.8535	0.4884	0.2697	0.8299
	WE	F	6.0939	0.5900	-0.2618	0.1149
	WE	M	4.7971	0.4652	-0.3208	0.2463
L1	UC	F	9.3444	0.4654	-0.3190	0.4798
	UC	M	8.1517	0.4269	-0.1498	0.1071
	WE	F	9.2798	0.4439	-0.5745	0.5773
	WE	M	8.1716	0.3894	-0.5906	0.7166

Appendix C. Mixed model ANOVA's of phenotypes.

	B1	F	*P*	B2	F	*P*	B3	F	*P*
Pop		0.15	ns		8.2	**		0.04	ns
Sex		1800.05	****		816.51	****		1647.51	****
Pop*Sex		0.26	ns		1.38	ns		6.69	*
$V_{line(pop)}$		0.172			0.067			0.039	
$V_{sex*line(pop)}$		0.006			0.003			0.002	
$V_{rep(pop\ line)}$		0.014			0.005			0.003	
$V_{sex*rep(pop\ line)}$		0.001			0.001			0.001	
$V_{residual}$		0.066			0.033			0.019	
	C1			C2			C3		
Pop		4.07	*		0.34	ns		0.03	ns
Sex		45.25	****		1313.04	****		765.01	****
Pop*Sex		3.02	ns		0.63	ns		0	ns
$V_{line(pop)}$		0.146			0.035			0.017	
$V_{sex*line(pop)}$		0.004			0.002			0.001	
$V_{rep(pop\ line)}$		0.006			0.003			0.001	
$V_{sex*rep(pop\ line)}$		0.001			0.000			0.000	
$V_{residual}$		0.066			0.017			0.014	
	D1			D2			D3		
Pop		0.14	ns		0.88	ns		0.06	ns
Sex		585.12	****		109.32	****		30.65	****
Pop*Sex		1.65	ns		3.89	*		1.35	ns
$V_{line(pop)}$		0.318			0.113			0.081	
$V_{sex*line(pop)}$		0.014			0.005			0.003	
$V_{rep(pop\ line)}$		0.015			0.007			0.005	
$V_{sex*rep(pop\ line)}$		0.004			0.002			0.001	
$V_{residual}$		0.148			0.064			0.072	

Variance components multiplied by 1000.
Significance: "ns" $P > 0.05$, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Appendix C. (continued)

	W1	F	*P*	W2	F	*P*	W3	F	*P*
Pop		1.14	ns		6.58	*		0.1	ns
Sex		1543.19	****		95.01	****		1136.6	****
Pop*Sex		0.12	ns		2.04	ns		0.92	ns
$V_{line(pop)}$		0.208			0.150			0.075	
$V_{sex*line(pop)}$		0.006			0.005			0.004	
$V_{rep(pop\ line)}$		0.011			0.008			0.006	
$V_{sex*rep(pop\ line)}$		0.001			0.001			0.000	
$V_{residual}$		0.054			0.056			0.037	
	W4			W5			W6		
Pop		0.14	ns		6.87	***		1.7	ns
Sex		569.81	****		279.79	****		1423.56	****
Pop*Sex		2.17	ns		9.43	***		1.7	ns
$V_{line(pop)}$		0.060			0.045			0.033	
$V_{sex*line(pop)}$		0.003			0.002			0.002	
$V_{rep(pop\ line)}$		0.003			0.004			0.003	
$V_{sex*rep(pop\ line)}$		0.001			0.001			0.000	
$V_{residual}$		0.028			0.023			0.021	
	W7			W8			W9		
Pop		0.01	ns		0	ns		0.01	ns
Sex		0.06	ns		46.22	****		131.73	****
Pop*Sex		0	ns		5.51	*		1.76	ns
$V_{line(pop)}$		0.014			0.010			0.008	
$V_{sex*line(pop)}$		0.001			0.000			0.000	
$V_{rep(pop\ line)}$		0.001			0.001			0.001	
$V_{sex*rep(pop\ line)}$		0.000			0.000			0.000	
$V_{residual}$		0.017			0.007			0.007	

Variance components multiplied by 1000.
Significance: "ns" P > 0.05, * P < 0.05, ** P <0.01, *** P <0.001, **** P < 0.0001.

Appendix C. (continued 2)

	B Area	F	*P*	C Area	F	*P*	D Area	F	*P*
Pop		0.34	ns		0.5	ns		4.26	*
Sex		9337.61	****		9704.06	****		11121.5	****
Pop*Sex		30.04	****		27.19	****		29.47	****
$V_{line(pop)}$		0.175			0.131			0.092	
$V_{sex*line(pop)}$		0.013			0.008			0.009	
$V_{rep(pop\ line)}$		0.097			0.079			0.065	
$V_{sex*rep(pop\ line)}$		0.012			0.008			0.008	
$V_{residual}$		0.191			0.143			0.121	
	T Area			L1					
Pop		1.34	ns		0.04	ns			
Sex		11705.4	****		14380.9	****			
Pop*Sex		34.02	****		19.69	****			
$V_{line(pop)}$		0.912			0.055				
$V_{sex*line(pop)}$		0.071			0.004				
$V_{rep(pop\ line)}$		0.701			0.045				
$V_{sex*rep(pop\ line)}$		0.080			0.005				
$V_{residual}$		1.180			0.078				

Significance: “ns” P > 0.05, * P < 0.05, ** P <0.01, *** P <0.001, **** P < 0.0001.

Appendix D. Variance components by population and sex.

Trait	Variance Component	UC F	UC M	WE F	WE M
B1	V_L	0.2224	0.1987	0.1523	0.1639
	V_R	0.0169	0.0152	0.0208	0.0207
	V_E	0.0556	0.0640	0.0620	0.0629
B2	V_L	0.0897	0.0749	0.0660	0.0579
	V_R	0.0038	0.0038	0.0088	0.0086
	V_E	0.0329	0.0329	0.0338	0.0306
B3	V_L	0.0472	0.0483	0.0338	0.0364
	V_R	0.0018	0.0026	0.0048	0.0044
	V_E	0.0176	0.0189	0.0178	0.0188
C1	V_L	0.1869	0.1961	0.1218	0.1290
	V_R	0.0051	0.0050	0.0083	0.0079
	V_E	0.0638	0.0635	0.0717	0.0637
C2	V_L	0.0337	0.0402	0.0358	0.0369
	V_R	0.0019	0.0029	0.0032	0.0041
	V_E	0.0168	0.0185	0.0161	0.0184
C3	V_L	0.0184	0.0214	0.0166	0.0153
	V_R	0.0016	0.0016	0.0021	0.0019
	V_E	0.0133	0.0143	0.0136	0.0144
D1	V_L	0.3660	0.3825	0.3390	0.2814
	V_R	0.0050	0.0131	0.0305	0.0186
	V_E	0.1522	0.1405	0.1535	0.1372
D2	V_L	0.1376	0.1249	0.1106	0.1060
	V_R	0.0052	0.0079	0.0135	0.0121
	V_E	0.0654	0.0596	0.0644	0.0614
D3	V_L	0.0889	0.0661	0.0905	0.0883
	V_R	0.0050	0.0027	0.0079	0.0061
	V_E	0.0749	0.0629	0.0762	0.0703

B1-D3 Variance components multiplied by 1000.
V_L, V_R and V_E : Variance components for Line, Replicate and Error.

Appendix D. (Continued)

Trait	Variance Component	UC F	UC M	WE F	WE M
W1	V_L	0.2558	0.2597	0.1829	0.1856
	V_R	0.0133	0.0110	0.0174	0.0154
	V_E	0.0494	0.0505	0.0533	0.0497
W2	V_L	0.1840	0.1781	0.1388	0.1384
	V_R	0.0056	0.0063	0.0138	0.0121
	V_E	0.0535	0.0516	0.0530	0.0589
W3	V_L	0.0814	0.0837	0.0782	0.0747
	V_R	0.0050	0.0058	0.0096	0.0081
	V_E	0.0357	0.0362	0.0378	0.0338
W4	V_L	0.0678	0.0650	0.0645	0.0596
	V_R	0.0019	0.0032	0.0052	0.0039
	V_E	0.0264	0.0272	0.0281	0.0273
W5	V_L	0.0655	0.0586	0.0417	0.0353
	V_R	0.0024	0.0024	0.0047	0.0051
	V_E	0.0250	0.0231	0.0235	0.0211
W6	V_L	0.0377	0.0315	0.0356	0.0331
	V_R	0.0018	0.0023	0.0052	0.0046
	V_E	0.0198	0.0197	0.0212	0.0199
W7	V_L	0.0162	0.0136	0.0148	0.0159
	V_R	0.0009	0.0006	0.0015	0.0013
	V_E	0.0177	0.0153	0.0189	0.0165
W8	V_L	0.0110	0.0129	0.0100	0.0106
	V_R	0.0008	0.0009	0.0012	0.0015
	V_E	0.0069	0.0080	0.0064	0.0073
W9	V_L	0.0113	0.0100	0.0081	0.0070
	V_R	0.0005	0.0007	0.0008	0.0007
	V_E	0.0076	0.0073	0.0059	0.0060
T Area	V_L	1.3942	1.1097	0.9268	0.6136
	V_R	0.8455	0.7676	1.2503	0.8959
	V_E	1.4987	0.8822	1.0752	0.6492
B Area	V_L	0.2560	0.2023	0.1908	0.1232
	V_R	0.1126	0.1133	0.1603	0.1249
	V_E	0.2423	0.1531	0.1775	0.1176
C Area	V_L	0.2002	0.1630	0.1297	0.0904
	V_R	0.0928	0.0843	0.1451	0.0985
	V_E	0.1809	0.1079	0.1317	0.0797
D Area	V_L	0.1317	0.0865	0.1110	0.0680
	V_R	0.0839	0.0673	0.1217	0.0838
	V_E	0.1549	0.0854	0.1177	0.0665
L1	V_L	0.0791	0.0708	0.0512	0.0415
	V_R	0.0493	0.0506	0.0797	0.0645
	V_E	0.0904	0.0605	0.0681	0.0470

W1-W9 Variance components multiplied by 1000.
V_L, V_R and V_E : Variance components for Line, Replicate and Error.

Appendix E. Genetic correlations between shape and size.

	B1	B2	B3	C1	C2	C3	D1	D2	D3	W1	W2	W3	W4	W5	W6	W7	W8	W9	T Area	B Area	C Area	D Area	L1
B1	0.96	0.16	0.24	-0.29	0.27	0.02	0.18	0.01	0.22	-1.00	-0.52	0.53	0.47	-0.27	-0.20	0.11	0.17	0.11	-0.45	-0.29	-0.41	-0.54	-0.77
B2	0.18	0.92	-0.15	-0.20	0.37	0.33	0.31	0.44	0.03	-0.28	-0.48	-0.90	-0.01	-1.13	0.22	0.09	-0.08	-0.01	0.12	0.19	0.04	0.07	0.17
B3	0.26	-0.23	0.95	0.10	0.05	0.24	0.20	-0.38	0.15	0.03	-0.10	0.35	0.77	-0.04	-0.99	-0.11	-0.83	0.03	-0.18	-0.54	0.19	-0.06	0.11
C1	-0.31	-0.21	0.02	1.03	0.05	-0.03	0.45	0.51	0.02	0.87	-0.94	0.24	0.24	0.37	0.29	-0.59	0.01	0.04	0.12	0.21	0.23	-0.20	0.40
C2	0.25	0.38	-0.03	0.01	1.01	0.20	0.38	0.44	-0.15	-0.49	-0.64	-0.51	-0.24	0.34	0.27	0.06	-1.12	0.27	0.36	0.05	0.86	0.03	-0.09
C3	0.02	0.28	0.31	-0.02	0.17	0.97	0.35	0.32	-0.03	-0.25	-0.46	-0.52	-0.26	0.19	-1.13	-0.16	0.16	-0.10	0.10	-0.23	0.32	0.25	-0.23
D1	0.15	0.25	0.17	0.46	0.31	0.34	0.91	0.07	0.02	-0.20	-0.59	-0.78	0.87	0.60	0.19	-0.06	0.09	-0.08	0.21	0.17	0.35	0.00	0.18
D2	-0.01	0.44	-0.42	0.49	0.47	0.31	0.02	0.95	-0.01	-0.06	-0.83	-0.42	-0.81	-0.02	-0.14	-0.86	0.19	-0.32	0.21	0.10	0.08	0.42	-0.14
D3	0.23	0.02	0.12	0.03	-0.08	-0.07	0.02	-0.01	0.91	-0.33	0.42	0.10	0.46	-0.31	0.18	-1.66	0.02	0.35	-0.26	-0.25	-0.28	-0.15	-0.22
W1	-1.00	-0.28	-0.02	0.89	-0.49	-0.23	-0.20	-0.06	-0.36	0.99	-0.04	0.17	-0.11	0.08	0.20	-0.02	-0.02	-0.05	0.18	0.24	0.15	0.07	0.72
W2	-0.50	-0.51	-0.04	-0.92	-0.58	-0.47	-0.54	-0.80	0.42	-0.04	1.02	0.04	-0.05	-0.02	0.06	0.02	0.04	-0.02	-0.06	-0.12	-0.24	0.27	0.07
W3	0.51	-0.87	0.38	0.21	-0.51	-0.49	-0.82	-0.42	0.09	0.18	0.06	0.94	0.17	-0.10	-0.21	-0.01	0.05	0.02	-0.54	-0.35	-0.54	-0.58	-0.38
W4	0.45	-0.04	0.71	0.21	-0.29	-0.23	0.90	-0.84	0.43	-0.13	-0.01	0.11	0.94	0.03	0.14	-0.02	-0.08	-0.05	-0.27	-0.17	-0.16	-0.45	0.11
W5	-0.29	-1.15	0.04	0.37	0.24	0.19	0.63	-0.07	-0.27	0.09	0.06	-0.09	0.10	0.89	-0.06	0.01	-0.03	0.03	0.30	0.08	0.53	0.20	0.00
W6	-0.18	0.28	-1.03	0.30	0.27	-1.13	0.21	-0.12	0.25	0.19	0.05	-0.25	0.17	-0.15	0.92	0.04	0.04	0.08	0.15	0.55	0.02	-0.32	0.27
W7	0.12	0.10	-0.05	-0.60	-0.01	-0.13	-0.04	-0.86	-1.72	0.02	-0.02	0.01	0.02	-0.01	-0.04	0.93	-0.03	0.02	0.08	0.21	0.12	-0.20	0.11
W8	0.22	-0.08	-0.74	0.09	-1.02	0.16	0.22	0.16	-0.03	-0.04	-0.06	0.02	0.03	0.06	0.04	0.03	1.02	-0.03	-0.20	0.26	-0.79	-0.03	-0.23
W9	0.17	-0.06	0.04	0.04	0.23	-0.15	-0.11	-0.30	0.36	-0.07	-0.02	0.11	-0.04	0.03	0.06	-0.02	0.01	0.90	0.00	0.00	0.00	0.00	0.00
T Area	-0.44	0.05	-0.03	0.03	0.28	0.05	0.22	0.10	-0.35	0.18	0.02	-0.51	-0.19	0.32	0.00	0.21	-0.21	0.00	1.09	2.62	2.52	2.78	2.80
B Area	-0.24	0.18	-0.43	0.15	-0.01	-0.28	0.20	-0.01	-0.29	0.21	-0.08	-0.33	-0.06	0.05	0.44	0.33	0.25	0.00	2.67	0.97	1.99	2.30	2.55
C Area	-0.43	-0.06	0.32	0.17	0.77	0.29	0.35	-0.02	-0.35	0.17	-0.14	-0.50	-0.09	0.57	-0.12	0.23	-0.78	0.00	2.71	2.04	1.04	2.21	2.33
D Area	-0.53	-0.03	0.11	-0.30	-0.04	0.21	-0.01	0.32	-0.26	0.06	0.34	-0.52	-0.38	0.25	-0.46	-0.07	-0.08	0.00	2.89	2.25	2.34	1.07	2.59
L1	-0.78	0.10	0.24	0.33	-0.21	-0.28	0.19	-0.29	-0.33	0.75	0.16	-0.35	0.21	0.02	0.15	0.28	-0.25	0.00	2.99	2.61	2.53	2.69	1.22

Calculated for sexes independently over populations, males above and females below diagonal.
On the diagonal are genetic correlations of the trait between sexes.

Appendix F. Genetic correlations and 95% confidence intervals for the 23 traits, estimated for females only.

	B1	B2	B3	C1	C2	C3	D1	D2	D3
B1	1.4								
B2	0.18	1.57							
	(0.04, 0.31)								
B3	0.26	-0.23	1.53						
	(0.13, 0.38)	(-0.35, -0.09)							
C1	-0.31	-0.21	0.02	1.52					
	(-0.43, -0.19)	(-0.33, -0.07)	(-0.11, 0.16)						
C2	0.25	0.38	-0.03	0.01	1.49				
	(0.12, 0.37)	(0.25, 0.49)	(-0.17, 0.1)	(-0.12, 0.15)					
C3	0.02	0.28	0.31	-0.02	0.17	1.87			
	(-0.12, 0.15)	(0.15, 0.4)	(0.19, 0.43)	(-0.16, 0.11)	(0.03, 0.3)				
D1	0.15	0.25	0.17	0.46	0.31	0.34	1.46		
	(0.02, 0.28)	(0.12, 0.37)	(0.03, 0.29)	(0.34, 0.56)	(0.18, 0.43)	(0.22, 0.46)			
D2	-0.01	0.44	-0.42	0.49	0.47	0.31	0.02	1.6	
	(-0.15, 0.12)	(0.33, 0.55)	(-0.53, -0.31)	(0.38, 0.58)	(0.36, 0.57)	(0.18, 0.43)	(-0.12, 0.15)		
D3	0.23	0.02	0.12	0.03	-0.08	-0.07	0.02	-0.01	1.89
	(0.09, 0.35)	(-0.12, 0.16)	(-0.02, 0.25)	(-0.11, 0.16)	(-0.22, 0.05)	(-0.21, 0.06)	(-0.12, 0.15)	(-0.14, 0.13)	
B Area	-0.24	0.18	-0.43	0.15	-0.01	-0.28	0.2	-0.01	-0.29
	(-0.36, -0.11)	(0.05, 0.31)	(-0.53, -0.31)	(0.01, 0.28)	(-0.14, 0.13)	(-0.4, -0.15)	(0.07, 0.33)	(-0.14, 0.13)	(-0.41, -0.16)
C Area	-0.43	-0.06	0.32	0.17	0.77	0.29	0.35	-0.02	-0.35
	(-0.53, -0.31)	(-0.2, 0.07)	(0.19, 0.43)	(0.03, 0.3)	(0.71, 0.82)	(0.16, 0.41)	(0.23, 0.47)	(-0.15, 0.12)	(-0.46, -0.23)
D Area	-0.53	-0.03	0.11	-0.3	-0.04	0.21	-0.01	0.32	-0.26
	(-0.62, -0.43)	(-0.17, 0.1)	(-0.02, 0.25)	(-0.42, -0.17)	(-0.17, 0.1)	(0.08, 0.34)	(-0.15, 0.12)	(0.2, 0.44)	(-0.39, -0.13)

Genetic correlations, and 95% confidence intervals in brackets below, as calculated with z-function (Sokal and Rohlf 1995).
Confidence intervals can only be estimated for correlations within bounds (-1 to 1), N=206.

Appendix F. (continued)

	B1	B2	B3	C1	C2	C3	D1	D2	D3
W1	-1	-0.28	-0.02	0.89	-0.49	-0.23	-0.2	-0.06	-0.36
		(-0.4, -0.14)	(-0.16, 0.12)	(0.86, 0.92)	(-0.59, -0.38)	(-0.36, -0.1)	(-0.33, -0.06)	(-0.19, 0.08)	(-0.48, -0.24)
W2	-0.5	-0.51	-0.04	-0.92	-0.58	-0.47	-0.54	-0.8	0.42
	(-0.59, -0.39)	(-0.6, -0.4)	(-0.17, 0.1)	(-0.94, -0.89)	(-0.66, -0.48)	(-0.57, -0.35)	(-0.63, -0.43)	(-0.85, -0.75)	(0.31, 0.53)
W3	0.51	-0.87	0.38	0.21	-0.51	-0.49	-0.82	-0.42	0.09
	(0.41, 0.61)	(-0.9, -0.83)	(0.26, 0.49)	(0.08, 0.34)	(-0.6, -0.4)	(-0.59, -0.38)	(-0.86, -0.78)	(-0.53, -0.3)	(-0.05, 0.22)
W4	0.45	-0.04	0.71	0.21	-0.29	-0.23	0.9	-0.84	0.43
	(0.34, 0.55)	(-0.18, 0.1)	(0.64, 0.78)	(0.08, 0.34)	(-0.41, -0.16)	(-0.36, -0.1)	(0.88, 0.93)	(-0.87, -0.79)	(0.31, 0.53)
W5	-0.29	-1.15	0.04	0.37	0.24	0.19	0.63	-0.07	-0.27
	(-0.41, -0.16)		(-0.1, 0.17)	(0.24, 0.48)	(0.1, 0.36)	(0.06, 0.32)	(0.54, 0.71)	(-0.2, 0.07)	(-0.39, -0.14)
W6	-0.18	0.28	-1.03	0.3	0.27	-1.13	0.21	-0.12	0.25
	(-0.31, -0.05)	(0.15, 0.4)		(0.18, 0.42)	(0.14, 0.39)		(0.08, 0.34)	(-0.25, 0.01)	(0.12, 0.37)
W7	0.12	0.1	-0.05	-0.6	-0.01	-0.13	-0.04	-0.86	-1.72
	(-0.02, 0.25)	(-0.03, 0.24)	(-0.18, 0.09)	(-0.68, -0.5)	(-0.15, 0.12)	(-0.26, 0)	(-0.17, 0.1)	(-0.89, -0.82)	
W8	0.22	-0.08	-0.74	0.09	-1.02	0.16	0.22	0.16	-0.03
	(0.09, 0.35)	(-0.21, 0.06)	(-0.79, -0.67)	(-0.05, 0.22)		(0.03, 0.29)	(0.09, 0.35)	(0.02, 0.29)	(-0.16, 0.11)
W9	0.17	-0.06	0.04	0.04	0.23	-0.15	-0.11	-0.3	0.36
	(0.04, 0.3)	(-0.19, 0.08)	(-0.09, 0.18)	(-0.09, 0.18)	(0.1, 0.36)	(-0.28, -0.02)	(-0.24, 0.03)	(-0.42, -0.17)	(0.24, 0.47)
T Area	-0.44	0.05	-0.03	0.03	0.28	0.05	0.22	0.1	-0.35
	(-0.55, -0.33)	(-0.09, 0.18)	(-0.17, 0.1)	(-0.1, 0.17)	(0.15, 0.4)	(-0.08, 0.19)	(0.08, 0.34)	(-0.04, 0.23)	(-0.46, -0.22)
L1	-0.78	0.1	0.24	0.33	-0.21	-0.28	0.19	-0.29	-0.33
	(-0.83, -0.72)	(-0.04, 0.23)	(0.11, 0.36)	(0.2, 0.45)	(-0.34, -0.08)	(-0.4, -0.15)	(0.06, 0.32)	(-0.41, -0.16)	(-0.44, -0.2)

Genetic correlations, and 95% confidence intervals in brackets below, as calculated with z-function (Sokal and Rohlf 1995).
Confidence intervals can only be estimated for correlations within bounds (-1 to 1), N=206.

Appendix F. (continued)

	W1	W2	W3	W4	W5	W6	W7	W8	W9
W1	1.27								
W2	-0.04	1.43							
	(-0.18, 0.09)								
W3	0.18	0.06	1.54						
	(0.04, 0.3)	(-0.07, 0.2)							
W4	-0.13	-0.01	0.11	1.45					
	(-0.26, 0)	(-0.15, 0.12)	(-0.03, 0.24)						
W5	0.09	0.06	-0.09	0.1	1.6				
	(-0.05, 0.22)	(-0.08, 0.19)	(-0.22, 0.05)	(-0.03, 0.23)					
W6	0.19	0.05	-0.25	0.17	-0.15	1.66			
	(0.06, 0.32)	(-0.08, 0.19)	(-0.37, -0.12)	(0.04, 0.3)	(-0.28, -0.01)				
W7	0.02	-0.02	0.01	0.02	-0.01	-0.04	2.28		
	(-0.12, 0.16)	(-0.16, 0.11)	(-0.12, 0.15)	(-0.12, 0.15)	(-0.15, 0.12)	(-0.17, 0.1)			
W8	-0.04	-0.06	0.02	0.03	0.06	0.04	0.03	1.71	
	(-0.18, 0.09)	(-0.2, 0.07)	(-0.11, 0.16)	(-0.1, 0.17)	(-0.07, 0.2)	(-0.09, 0.18)	(-0.11, 0.16)		
W9	-0.07	-0.02	0.11	-0.04	0.03	0.06	-0.02	0.01	1.69
	(-0.2, 0.07)	(-0.15, 0.12)	(-0.03, 0.24)	(-0.17, 0.1)	(-0.11, 0.17)	(-0.07, 0.2)	(-0.15, 0.12)	(-0.12, 0.15)	
T Area	0.18	0.02	-0.51	-0.19	0.32	0	0.21	-0.21	0
	(0.04, 0.3)	(-0.11, 0.16)	(-0.6, -0.4)	(-0.31, -0.05)	(0.19, 0.44)	(-0.14, 0.13)	(0.07, 0.33)	(-0.34, -0.08)	(-0.14, 0.14)
L1	0.75	0.16	-0.35	0.21	0.02	0.15	0.28	-0.25	0
	(0.68, 0.8)	(0.03, 0.29)	(-0.46, -0.22)	(0.08, 0.34)	(-0.11, 0.16)	(0.02, 0.28)	(0.14, 0.4)	(-0.37, -0.12)	(-0.14, 0.14)

Genetic correlations, and 95% confidence intervals in brackets below, as calculated with z-function (Sokal and Rohlf 1995).
Confidence intervals can only be estimated for correlations within bounds (-1 to 1), N=206.

Appendix F. (continued)

	W1	W2	W3	W4	W5	W6	W7	W8	W9
B Area	0.21	-0.08	-0.33	-0.06	0.05	0.44	0.33	0.25	0
	(0.08, 0.34)	(-0.22, 0.05)	(-0.45, -0.21)	(-0.19, 0.08)	(-0.09, 0.18)	(0.32, 0.54)	(0.2, 0.44)	(0.12, 0.37)	(-0.14, 0.14)
C Area	0.17	-0.14	-0.5	-0.09	0.57	-0.12	0.23	-0.78	0
	(0.03, 0.3)	(-0.27, -0.01)	(-0.6, -0.4)	(-0.22, 0.04)	(0.47, 0.66)	(-0.25, 0.02)	(0.09, 0.35)	(-0.83, -0.72)	(-0.14, 0.14)
D Area	0.06	0.34	-0.52	-0.38	0.25	-0.46	-0.07	-0.08	0
	(-0.08, 0.19)	(0.22, 0.46)	(-0.61, -0.42)	(-0.49, -0.26)	(0.12, 0.37)	(-0.56, -0.35)	(-0.21, 0.06)	(-0.22, 0.05)	(-0.14, 0.14)

	T Area	B Area	C Area	D Area	L1
T Area	3.15				
B Area	2.67	2.62			
C Area	2.71	2.04	2.76		
D Area	2.89	2.25	2.34	3.08	
L1	2.99	2.61	2.53	2.69	3.37

Genetic correlations, and 95% confidence intervals in brackets below, as calculated with z-function (Sokal and Rohlf 1995).
Confidence intervals can only be estimated for correlations within bounds (-1 to 1), N=206.

Appendix G. Key for genebank and working alignments.

GB	Work	GB	Work	GB	Work	GB	Work
5479	77	5922	511	6520	1195	30628	1786
5480	78	5924	513	6527	1212	30676	1834
5498	96	5929	518	6529	1217	30688	1846
5499	97	5954	543	6531	1219	30691	1849
5509	107	5963	552	6533	1221	30704	1862
5510	108	5967	556	6535	1223	30727	1885
5511	109	5987	576	30145	1280	30733	1891
5530	128	5993	582	30146	1281	30736	1894
5533	131	5998	587	30182	1317	30763	1921
5559	157	6034	623	30200	1335	30805	1963
5589	187	6058	647	30245	1380	30814	1972
5604	202	6063	652	30264	1400	30867	2025
5616	214	6065	654	30281	1438	30869	2027
5636	236	6073	662	30281	1440	30894	2054
5686	286	6077	666	30286	1443	30934	2094
5688	288	6081	670	30292	1449	30936	2096
5695	295	6085	674	30334	1491	30970	2130
5700	300	6099	688	30343	1501	30974	2134
5703	303	6135	724	30350	1508	30992	2152
5712	312	6212	804	30358	1516	30993	2153
5737	337	6178	830	30381	1539	30994	2154
5762	362	6271	923	30401	1559	31049	2207
5763	363	6326	983	30402	1560	31051	2209
5764	364	6326	992	30403	1561	31070	2231
5795	395	6326	997	30416	1574	31116	2277
5810	410	6326	1000	30456	1614	31160	2321
5833	433	6340	1015	30482	1640	31164	2325
5872	472	6343	1018	30484	1642	31175	2336
5895	495	6349	1024	30486	1644	31191	2352
5896	496	6364	1039	30505	1663	31200	2361
5897	497	6370	1045	30506	1664	31201	2362
5898	498	6371	1046	30508	1666	31210	2371
5899	501	6383	1058	30511	1669	31211	2372
5899	504	6410	1085	30538	1696	31215	2376
5899	506	6412	1087	30556	1714	31218	2379
5899	507	6420	1095	30563	1721	31220	2381
5900	508	6421	1096	30565	1723	31223	2384
5900	509	6512	1187	30589	1747	31238	2399
5900	510	6516	1191	30623	1781	31242	2403

GB: Location in Genebank record 17571116, (Flybase: FBgn000373)
Work: Location in the laboratory alignment.

Appendix G. (Continued)

GB	Work	GB	Work	GB	Work	GB	Work
31244	2405	31629	2808	35683	3709	36162	4189
31245	2406	31633	2812	35685	3711	36177	4204
31250	2411	31634	2813	35687	3713	36190	4217
31258	2419	31650	2829	35697	3723	36199	4226
31264	2425	31664	2843	35772	3798	36200	4227
31274	2435	31665	2844	35777	3803	36201	4228
31277	2441	31747	2926	35779	3805	36210	4237
31279	2443	31758	2937	35780	3806	36214	4241
31281	2445	31761	2940	35788	3814	36222	4249
31288	2452	31773	2952	35801	3827	36226	4253
31289	2453	31798	2977	35802	3828	36248	4275
31291	2455	31845	3024	35808	3834	36279	4306
31291	2457	31873	3053	35809	3835	36315	4342
31293	2462	31881	3061	35810	3836	36332	4359
31294	2463	31902	3086	35816	3842	36343	4370
31321	2490	31915	3106	35826	3853	36366	4393
31333	2502	31929	3120	35829	3856	36396	4423
31340	2509	31942	3133	35834	3861	36431	4458
31341	2510	31950	3141	35910	3937	36434	4461
31345	2514	31998	3189	35918	3945	36444	4471
31349	2518	32029	3220	35928	3955	36504	4534
31352	2521	32035	3226	35948	3975	36510	4540
31355	2524	32166	3358	35953	3980	36511	4541
31361	2530	35345	3371	35955	3982	36514	4544
31362	2531	35362	3388	36006	4033	36517	4547
31365	2534	35366	3392	36022	4049	36529	4559
31372	2541	35368	3394	36025	4052	36565	4595
31442	2611	35371	3397	36041	4068	36606	4636
31443	2612	35391	3417	36066	4093	36630	4660
31472	2641	35404	3430	36078	4105	36644	4674
31490	2659	35437	3463	36081	4108	36722	4752
31508	2677	35535	3561	36082	4109	36734	4764
31510	2679	35572	3598	36093	4120	36745	4781
31522	2691	35574	3600	36102	4129	36749	4785
31547	2716	35619	3645	36108	4135	36760	4796
31597	2766	35671	3697	36122	4149	36761	4797
31619	2798	35672	3698	36129	4156	36760	4798
31624	2803	35680	3706	36132	4159	36791	4829
31626	2805	35682	3708	36159	4186	36823	4861

GB: Location in Genebank record 17571116, (Flybase: FBgn000373)
Work: Location in the laboratory alignment.

Appendix G. (Continued)

GB	Work	GB	Work	GB	Work	GB	Work
36849	4887	37865	5901	38104	6177	38791	6864
36910	4948	37874	5910	38105	6178	38816	6889
36912	4950	37880	5916	38106	6179	38830	6903
36938	4976	37883	5919	38140	6213	38857	6930
36966	5004	37892	5928	38191	6264	38860	6933
36967	5005	37919	5955	38207	6280	38866	6939
37003	5041	37961	5997	38209	6282	38869	6942
37037	5075	37973	6009	38218	6291	38890	6963
37100	5138	37979	6015	38233	6306	38908	6981
37169	5207	37987	6024	38236	6309	38914	6987
37173	5211	37991	6028	38269	6342	38920	6993
37174	5212	37992	6029	38284	6357	38924	6997
37177	5215	37993	6030	38293	6366	38932	7005
37200	5238	37996	6033	38308	6381	38938	7011
37236	5274	38015	6052	38362	6435	38941	7014
37261	5299	38019	6056	38383	6456	38956	7029
37282	5320	38023	6060	38404	6477	39007	7080
37332	5370	38025	6062	38413	6486	39010	7083
37334	5372	38028	6065	38422	6495	39160	7233
37415	5453	38029	6066	38455	6528	39193	7266
37480	5518	38033	6070	38461	6534	39194	7267
37498	5536	38035	6072	38476	6549	39196	7269
37539	5577	38037	6074	38482	6555	39199	7272
37545	5583	38037	6075	38491	6564	39262	7335
37600	5638	38039	6081	38506	6579	39280	7353
37601	5639	38046	6088	38533	6606	39298	7371
37625	5663	38049	6091	38542	6615	39300	7373
37674	5712	38052	6094	38581	6654	39304	7377
37683	5721	38055	6096	38584	6657	39320	7393
37715	5753	38056	6098	38587	6660	39338	7411
37716	5754	38059	6101	38593	6666	39347	7420
37729	5767	38063	6106	38617	6690	39389	7462
37745	5783	38075	6148	38668	6741	39401	7474
37772	5808	38081	6154	38683	6756	39404	7477
37805	5841	38082	6155	38707	6780	39419	7492
37817	5853	38089	6162	38713	6786	39425	7498
37832	5868	38096	6169	38728	6801	39431	7504
37850	5886	38098	6171	38746	6819	39433	7506
37856	5892	38100	6173	38775	6848	39446	7519

GB: Location in Genebank record 17571116, (Flybase: FBgn000373)
Work: Location in the laboratory alignment.

Appendix G. (Continued)

GB	Work	GB	Work	GB	Work	GB	Work
39451	7524	40344	8421	41214	9291	42086	10163
39461	7534	40398	8475	41241	9318	42140	10217
39479	7552	40428	8505	41247	9324	42153	10230
39488	7561	40437	8514	41250	9327	42163	10240
39504	7577	40458	8535	41256	9333	42173	10250
39506	7579	40464	8541	41262	9339	42180	10257
39512	7585	40506	8583	41283	9360	42181	10258
39546	7623	40524	8601	41286	9363	42190	10267
39548	7625	40536	8613	41310	9387	42227	10308
39553	7630	40545	8622	41313	9390	42241	10322
39554	7631	40560	8637	41352	9429	42242	10323
39571	7648	40590	8667	41379	9456	42250	10331
39594	7671	40620	8697	41416	9493	42265	10346
39597	7674	40635	8712	41520	9597	42269	10350
39603	7680	40653	8730	41544	9621	42285	10366
39684	7761	40671	8748	41556	9633	42297	10378
39717	7794	40672	8749	41559	9636	42305	10386
39759	7836	40683	8760	41601	9678	42336	10418
39792	7869	40710	8787	41658	9735	42367	10449
39870	7947	40722	8799	41670	9747	42377	10459
39873	7950	40737	8814	41682	9759	42378	10460
39894	7971	40770	8847	41694	9771	42391	10473
39909	7986	40824	8901	41703	9780	42405	10487
39912	7989	40936	9013	41712	9789	42415	10497
39948	8025	40944	9021	41733	9810	42422	10504
39972	8049	40959	9036	41743	9820	42424	10506
40026	8103	40965	9042	41751	9828	42444	10530
40044	8121	40998	9075	41769	9846	42454	10540
40059	8136	41040	9117	41818	9895	42467	10553
40101	8178	41061	9138	41819	9896	42489	10575
40110	8187	41064	9141	41820	9897	42493	10579
40119	8196	41079	9156	41823	9900	42534	10620
40140	8217	41112	9189	41850	9927	42570	10656
40149	8226	41115	9192	41925	10002	42615	10701
40152	8229	41148	9225	41931	10008	42645	10731
40158	8235	41154	9231	41943	10020	42660	10746
40224	8301	41157	9234	42010	10087	42661	10747
40263	8340	41160	9237	42018	10095	42707	10793
40281	8358	41163	9240	42023	10100	42779	10865
40332	8409	41187	9264	42043	10120	42783	10869

GB: Location in Genebank record 17571116, (Flybase: FBgn000373)
Work: Location in the laboratory alignment.

Appendix H. Multi-trait associations for all 18 traits.

SNP		SNP x SEX		SNP x POP		SNP x SEX x POP	
Traits	Site	Traits	Site	Traits	Site	Traits	Site
7	10346	7	*00215	7	*10431	6	*01432
6	07462	5	10100	7	10418	6	02510
6	08187	4	01335	6	10322	6	02813
6	10449	4	02813	6	10460	5	*06096
5	03982	4	03982	5	*01229	4	*01441
5	04976	4	05997	5	*01432	4	*05230
5	06987	3	00108	5	02502	4	02406
5	08226	3	00131	5	06306	4	05997
5	10120	3	00513	5	07269	4	06060
4	01335	3	00654	5	07373	4	09156
4	01663	3	01095	5	10323	3	*00978
4	01891	3	02798	5	10449	3	*02214
4	02325	3	02808	4	*01080	3	*10431
4	02803	3	03133	4	*01202	3	01885
4	02805	3	03371	4	00497	3	02403
4	04068	3	04950	4	03133	3	02524
4	05841	3	06056	4	03598	3	03133
4	06342	3	06060	4	04660	3	03371
4	06660	3	06155	4	07272	3	04241
4	07233	3	07233	3	*00353	3	04559
4	07534	3	08799	3	*00857	3	04950
4	07971	3	09324	3	*00873	3	06033
4	07989	3	09339	3	*01403	3	08340
3	*05230	3	09678	3	*02435	3	08697
3	00501	3	10120	3	00078	3	09141
3	00654	3	10346	3	01015	3	09240
3	01087	3	10418	3	01085	3	09318
3	01560	3	10865	3	01095	3	09339
3	02502			3	04674	3	10120
3	04797			3	07083	3	10323
3	05853			3	07462	3	10346
3	06933			3	08226	3	10418
3	07492			3	08421	3	10449
3	07506			3	09156	3	10460
3	07869			3	09789		
3	08103			3	10865		
3	08121						
3	08178						
3	08196						
3	08340						
3	09240						
3	09264						
3	09846						
3	09897						
3	09927						

The number of traits each site affected at the 0.05 level, as assayed in analysis of variance. Counting all 18 shape parameters. The columns represent the four genetic terms in the model, SNP, and interactions with sex and population. Star (*) indicate insertion-deletion polymorphisms.

Appendix I. Multi-trait associations for all 9 orthogonal whole wing traits.

SNP		SNP x SEX		SNP x POP		SNP x SEX x POP	
Traits	Site	Traits	Site	Traits	Site	Traits	Site
4	07462	4	*00215	3	*01080	4	*01432
3	03982	3	10100	3	*01432	3	02813
3	04068	2	*06096	3	*10431	2	*00978
3	04976	2	00108	3	06306	2	*01441
3	06660	2	00131	3	10322	2	*05230
3	07534	2	00157	3	10418	2	*06096
3	08103	2	00654	2	*00353	2	02403
3	10346	2	01335	2	*01202	2	02510
2	*05230	2	01723	2	*01229	2	04241
2	00654	2	02808	2	*03087	2	04559
2	01087	2	02813	2	00303	2	06033
2	01335	2	03982	2	00497	2	06066
2	01560	2	04950	2	01085	2	08049
2	02325	2	05910	2	03133	2	09156
2	02502	2	05997	2	03598	2	09231
2	04275	2	06889	2	05853	2	09240
2	05841	2	08799	2	05997	2	09318
2	05997	2	09789	2	07083	2	10120
2	06033	2	10120	2	07269	2	10346
2	06178	2	10418	2	07373		
2	06342	2	10865	2	07462		
2	06933			2	08226		
2	06987			2	08541		
2	07233			2	08697		
2	07492			2	09789		
2	07506			2	10323		
2	07869			2	10449		
2	07971			2	10460		
2	07989						
2	08178						
2	08187						
2	08226						
2	09240						
2	09333						
2	09846						
2	09897						
2	10120						
2	10449						
2	10865						

The number of traits each site affected at the 0.05 level, as assayed in analysis of variance. Just the orthogonal shape parameters for whole wing are surveyed. The columns represent the four genetic terms in the model, SNP, and interactions with sex and population. Star (*) indicate insertion-deletion polymorphisms.

Appendix J. Descriptive statistics of Round robin crosses with WE.

Trait	Block	Mean	Std	Skew	Kurt
B1	A	-0.0060	0.0118	0.4965	1.8622
	B	-0.0075	0.0127	0.3914	0.3745
	C	-0.0051	0.0120	0.1911	0.7733
B2	A	0.0049	0.0083	-0.1388	-0.0124
	B	0.0051	0.0079	0.1005	0.1216
	C	0.0047	0.0082	0.0134	0.4427
B3	A	0.0048	0.0068	-0.3451	0.3512
	B	0.0027	0.0066	0.1990	-0.2268
	C	0.0041	0.0066	0.0869	-0.2480
C1	A	0.0007	0.0104	0.0062	-0.1280
	B	0.0023	0.0122	0.1752	0.8293
	C	0.0005	0.0110	0.0035	-0.0902
C2	A	-0.0025	0.0061	0.1946	0.9897
	B	-0.0032	0.0058	0.1285	0.1733
	C	-0.0019	0.0056	0.1706	0.1663
C3	A	0.0019	0.0048	-0.1705	1.3122
	B	0.0018	0.0045	0.0098	0.1423
	C	0.0017	0.0048	-0.0166	0.0649
D1	A	0.0059	0.0177	0.0655	-0.2077
	B	0.0070	0.0170	-0.2185	0.1341
	C	0.0077	0.0157	-0.3587	1.0222
D2	A	-0.0038	0.0101	-0.1377	0.3401
	B	-0.0023	0.0108	0.1907	0.3018
	C	-0.0036	0.0099	0.0519	-0.0437
D3	A	0.0024	0.0108	0.2394	0.1089
	B	0.0024	0.0110	0.2555	0.5571
	C	0.0014	0.0108	0.1570	0.4229

Block: Each line was mated 3 times as dam and 3 times as sire, organized in 3 blocks of crosses (A, B, C). The pairings were randomized and the 3 blocks scored in interleaved fashion in replicate.

Appendix J. (Continued)

Trait	Block	Mean	Std	Skew	Kurt
W1	A	0.0063	0.0111	-0.4296	0.8084
	B	0.0079	0.0125	-0.5116	0.4734
	C	0.0047	0.0117	-0.2474	0.3667
W2	A	-0.0006	0.0106	-0.0141	0.2216
	B	-0.0011	0.0123	-0.1122	0.0589
	C	-0.0002	0.0112	-0.0042	-0.0201
W3	A	-0.0042	0.0093	0.1662	-0.1508
	B	-0.0054	0.0088	0.1889	0.4287
	C	-0.0046	0.0088	0.2861	0.8495
W4	A	-0.0047	0.0079	0.0963	-0.4489
	B	-0.0036	0.0077	0.0625	-0.2416
	C	-0.0048	0.0073	0.1522	0.1200
W5	A	-0.0017	0.0065	-0.1332	-0.0063
	B	-0.0016	0.0064	-0.0093	0.1803
	C	-0.0010	0.0063	0.1419	0.0245
W6	A	-0.0034	0.0058	0.0669	1.6983
	B	-0.0022	0.0060	0.1309	0.3728
	C	-0.0025	0.0060	-0.0981	0.0036
W7	A	0.0007	0.0046	0.1364	0.1290
	B	0.0004	0.0048	-0.1085	0.4178
	C	0.0005	0.0045	0.1276	0.2184
W8	A	-0.0004	0.0037	-0.1054	0.8599
	B	0.0006	0.0035	-0.3397	1.9465
	C	-0.0004	0.0034	-0.2817	0.1408
W9	A	0.0015	0.0032	-0.0914	0.4037
	B	0.0009	0.0031	0.0521	0.1302
	C	0.0012	0.0032	0.2329	0.3803
T Area	A	22.1964	1.4627	-0.5428	0.8698
	B	21.7482	1.5951	-0.6544	1.2438
	C	22.0888	1.4786	-0.5325	0.7183
B Area	A	8.8362	0.5979	-0.4303	0.6598
	B	8.7447	0.6053	-0.6902	2.1173
	C	8.8174	0.5908	-0.4912	0.5799
C Area	A	6.8176	0.5386	-0.4760	0.5667
	B	6.6208	0.5922	-0.4198	0.2908
	C	6.7778	0.5210	-0.3197	0.5408
D Area	A	6.5426	0.4607	-0.3897	0.5949
	B	6.3827	0.5233	-0.3503	0.7217
	C	6.4937	0.4933	-0.3827	0.3198
L1	A	9.6674	0.3510	-0.6257	1.4069
	B	9.5699	0.3934	-0.8555	2.1829
	C	9.6268	0.3664	-0.6393	0.7996

Block: Each line was mated 3 times as dam and 3 times as sire, organized in 3 blocks of crosses (A, B, C). The pairings were randomized and the 3 blocks scored in interleaved fashion in replicate.

Appendix K. Descriptive statistics of Kenyan test crosses.

Trait	Pop	Sex	Mean	Std	Skew	Kurt
B1	Sam	F	-0.0006	0.0105	0.2891	0.1863
	Sam	M	0.0123	0.0109	0.1498	0.2458
	bs	F	0.0000	0.0113	0.3476	0.1665
	bs	M	0.0149	0.0113	0.2291	0.0240
	E1	F	-0.0057	0.0114	0.4799	0.7533
	E1	M	0.0117	0.0113	0.1673	-0.1197
B2	Sam	F	0.0051	0.0061	0.0546	0.0093
	Sam	M	0.0022	0.0064	-0.0099	0.0802
	bs	F	-0.0032	0.0076	0.1415	0.0362
	bs	M	-0.0086	0.0076	-0.0235	0.0137
	E1	F	-0.0086	0.0064	-0.0980	0.8342
	E1	M	-0.0108	0.0062	0.1471	0.5661
B3	Sam	F	-0.0004	0.0054	0.0910	-0.1293
	Sam	M	-0.0036	0.0054	0.0170	0.0509
	bs	F	0.0082	0.0055	-0.5283	0.9401
	bs	M	0.0022	0.0050	-0.0329	1.3399
	E1	F	-0.0013	0.0052	0.1135	0.4808
	E1	M	-0.0062	0.0049	-0.0860	0.6467
C1	Sam	F	0.0016	0.0109	-0.1144	0.1173
	Sam	M	-0.0040	0.0108	-0.2586	-0.0100
	bs	F	0.0096	0.0102	-0.0914	-0.0168
	bs	M	0.0113	0.0097	-0.2549	0.5046
	E1	F	-0.0019	0.0093	0.0255	0.6943
	E1	M	-0.0032	0.0101	0.0643	0.3735
C2	Sam	F	0.0012	0.0059	0.1297	-0.3570
	Sam	M	0.0060	0.0064	0.0433	-0.3894
	bs	F	-0.0030	0.0057	-0.0005	-0.2010
	bs	M	0.0025	0.0057	0.3715	0.2760
	E1	F	0.0009	0.0057	0.6931	1.2434
	E1	M	0.0067	0.0061	0.6960	1.7279
C3	Sam	F	-0.0007	0.0045	0.2940	0.5615
	Sam	M	-0.0030	0.0042	0.0105	0.0575
	bs	F	0.0046	0.0047	-0.0689	-0.0099
	bs	M	0.0019	0.0043	0.0072	0.7534
	E1	F	0.0039	0.0046	-0.0309	0.0643
	E1	M	0.0022	0.0044	-0.0154	0.0771
D1	Sam	F	0.0172	0.0139	-0.0576	-0.0128
	Sam	M	0.0116	0.0147	-0.0913	-0.1403
	bs	F	0.0090	0.0157	-0.0465	0.0594
	bs	M	0.0017	0.0152	-0.1617	-0.0543
	E1	F	-0.0081	0.0142	0.3117	0.3449
	E1	M	-0.0146	0.0147	0.0460	0.4136

Appendix K. (continued)

Trait	Pop	Sex	Mean	Std	Skew	Kurt
D2	Sam	F	0.0007	0.0100	-0.0181	-0.1597
	Sam	M	0.0044	0.0106	-0.0001	-0.0631
	bs	F	-0.0134	0.0101	0.1112	0.6631
	bs	M	-0.0025	0.0095	0.1717	0.2675
	E1	F	-0.0106	0.0097	0.1560	0.1217
	E1	M	0.0014	0.0098	0.1985	0.1917
D3	Sam	F	0.0000	0.0097	0.3448	0.6918
	Sam	M	-0.0046	0.0090	0.4248	0.5220
	bs	F	-0.0005	0.0096	0.1248	1.0284
	bs	M	-0.0025	0.0085	0.1136	0.2338
	E1	F	-0.0015	0.0093	0.0296	0.1420
	E1	M	-0.0056	0.0088	0.3372	0.4874
W1	Sam	F	-0.0019	0.0098	-0.1836	-0.2034
	Sam	M	-0.0168	0.0099	-0.0024	-0.2264
	bs	F	0.0099	0.0109	-0.1554	-0.1325
	bs	M	-0.0034	0.0109	-0.2740	0.3975
	E1	F	0.0037	0.0112	-0.3490	0.3955
	E1	M	-0.0136	0.0106	-0.1826	0.3915
W2	Sam	F	-0.0019	0.0097	-0.1248	-0.0011
	Sam	M	-0.0006	0.0094	-0.1318	-0.1511
	bs	F	-0.0072	0.0089	0.1551	0.0077
	bs	M	-0.0132	0.0092	-0.0623	0.2979
	E1	F	0.0063	0.0097	-0.2236	0.2171
	E1	M	0.0010	0.0103	-0.1654	0.3645
W3	Sam	F	-0.0075	0.0071	-0.1121	0.2901
	Sam	M	-0.0026	0.0079	-0.1788	-0.0030
	bs	F	0.0061	0.0087	-0.0040	0.2417
	bs	M	0.0127	0.0089	0.0116	0.0929
	E1	F	0.0045	0.0072	-0.0619	0.1737
	E1	M	0.0094	0.0072	0.1540	0.3231
W4	Sam	F	-0.0040	0.0063	-0.1549	-0.1911
	Sam	M	-0.0014	0.0068	-0.2431	0.0191
	bs	F	-0.0081	0.0061	0.3622	1.0863
	bs	M	-0.0007	0.0064	-0.0069	0.7071
	E1	F	0.0034	0.0058	0.0686	0.0740
	E1	M	0.0098	0.0066	0.1638	0.6708
W5	Sam	F	0.0013	0.0052	-0.1150	0.0356
	Sam	M	0.0016	0.0052	-0.1157	0.1605
	bs	F	0.0051	0.0058	-0.0311	0.5225
	bs	M	0.0061	0.0053	0.0874	0.2930
	E1	F	0.0054	0.0054	0.0556	0.1987
	E1	M	0.0046	0.0053	0.0506	0.0509
W6	Sam	F	0.0031	0.0065	-0.1780	-0.2567
	Sam	M	0.0059	0.0059	-0.0553	-0.3530
	bs	F	-0.0031	0.0065	-0.1194	-0.0522
	bs	M	0.0003	0.0058	0.0585	0.8742
	E1	F	-0.0008	0.0064	0.1222	0.0692
	E1	M	0.0011	0.0062	0.2221	0.0892

Appendix K. (continued)

Trait	Pop	Sex	Mean	Std	Skew	Kurt
W7	Sam	F	0.0001	0.0049	0.0791	0.5316
	Sam	M	-0.0007	0.0048	0.1518	0.4806
	bs	F	0.0018	0.0051	0.1352	0.9796
	bs	M	0.0000	0.0042	0.0476	0.9578
	E1	F	0.0015	0.0046	-0.0872	0.0862
	E1	M	-0.0009	0.0043	0.0116	0.2157
W8	Sam	F	0.0006	0.0032	0.0613	0.2425
	Sam	M	-0.0004	0.0034	0.2930	-0.0017
	bs	F	-0.0005	0.0033	-0.0006	-0.0690
	bs	M	-0.0004	0.0033	-0.1892	0.1946
	E1	F	-0.0007	0.0033	-0.4527	0.5858
	E1	M	-0.0008	0.0036	-0.0238	0.4435
W9	Sam	F	0.0007	0.0027	-0.1868	0.9103
	Sam	M	-0.0006	0.0027	-0.1563	0.1620
	bs	F	-0.0005	0.0032	-0.2715	0.1818
	bs	M	-0.0017	0.0030	-0.1255	0.1745
	E1	F	0.0024	0.0030	-0.0701	0.6689
	E1	M	0.0014	0.0029	0.0168	0.0378
T Area	Sam	F	23.8398	1.6501	-0.3176	0.1117
	Sam	M	18.9534	1.3917	-0.3980	0.0042
	bs	F	23.6099	1.9468	-0.5020	-0.3892
	bs	M	19.0624	1.6598	-0.4932	-0.3040
	E1	F	23.9465	1.6181	-0.4281	0.4000
	E1	M	19.1073	1.3597	-0.0579	-0.0791
B Area	Sam	F	9.6569	0.6671	-0.2323	0.2895
	Sam	M	7.7437	0.5689	-0.3552	0.0658
	bs	F	9.4220	0.7803	-0.3743	-0.2505
	bs	M	7.7446	0.6726	-0.4189	-0.2236
	E1	F	9.5072	0.6604	-0.3160	0.0894
	E1	M	7.6949	0.5769	-0.0336	-0.1512
C Area	Sam	F	7.3027	0.5777	-0.2503	-0.1171
	Sam	M	5.8030	0.4862	-0.3520	0.0223
	bs	F	7.4796	0.7134	-0.4802	-0.5713
	bs	M	5.9948	0.5969	-0.4465	-0.2287
	E1	F	7.6104	0.6043	-0.4342	0.4734
	E1	M	6.0317	0.4936	-0.0590	-0.1817
D Area	Sam	F	6.8801	0.5167	-0.1680	0.0267
	Sam	M	5.4067	0.4127	-0.2326	-0.1604
	bs	F	6.7082	0.5440	-0.4268	-0.0173
	bs	M	5.3230	0.4583	-0.4027	-0.1608
	E1	F	6.8289	0.4721	-0.2853	0.2644
	E1	M	5.3807	0.3884	-0.1164	0.1839
L1	Sam	F	9.9200	0.3687	-0.5231	0.5353
	Sam	M	8.6856	0.3474	-0.5608	0.2513
	bs	F	9.9856	0.4550	-0.5554	-0.1675
	bs	M	8.7596	0.4155	-0.5379	-0.2254
	E1	F	9.8389	0.3752	-0.4552	0.4106
	E1	M	8.5766	0.3462	-0.0998	0.2024

Appendix L. Mixed model ANOVA's of phenotypes in Kenyan test cross.

	B1	F	*P*	B2	F	*P*	B3	F	*P*
Cross		11.39	****		379.37	****		229.33	****
Sex		1815.42	****		203.42	****		573.85	****
C*S		15.86	****		21.55	****		15.66	****
Line		26.85	****		13.35	****		11.75	****
C*L		1.92	**		1.69	**		1.29	ns
S*L		3.30	****		2.16	**		3.81	****
C*S*L		1.07	ns		1.50	*		1.27	ns
$V_{rep(C*L)}$		0.0119			0.0040			0.0033	
$V_{sex*rep(C*L)}$		0.0031			0.0004			0.0010	
$V_{residual}$					0.0326			0.0166	
	C1			C2			C3		
Cross		208.45	****		62.92	****		150.64	****
Sex		68.05	****		777.77	****		140.35	****
C*S		47.85	****		3.87	*		4.71	**
Line		18.09	****		28.65	****		16.91	****
C*L		1.55	*		1.91	**		1.45	*
S*L		1.34	ns		3.79	****		1.63	*
C*S*L		1.39	ns		1.45	*		0.96	ns
$V_{rep(C*L)}$		0.0083			0.0025			0.0016	
$V_{sex*rep(C*L)}$		0.0007			0.0003			0.0002	
$V_{residual}$		0.0708			0.0193			0.0136	
	D1			D2			D3		
Cross		197.80	****		125.96	****		20.53	****
Sex		225.24	****		633.74	****		95.11	****
C*S		2.75	*		53.97	****		8.77	****
Line		9.62	****		15.83	****		16.35	****
C*L		1.34	ns		1.83	**		1.89	**
S*L		1.82	*		2.39	***		2.13	**
C*S*L		1.79	**		1.10	ns		1.35	ns
$V_{rep(C*L)}$		0.0316			0.0078			0.0058	
$V_{sex*rep(C*L)}$		0.0000			0.0019			0.0006	
$V_{residual}$		0.1499			0.0652			0.0597	

Variance components multiplied by 1000.
Significance: "ns" P > 0.05, * P < 0.05, ** P <0.01, *** P <0.001, **** P < 0.0001.

Appendix L. (continued)

	W1	F	*P*	W2	F	*P*	W3	F	*P*
Cross		138.66	****		223.04	****		238.75	****
Sex		2160.54	****		95.94	****		446.73	****
C*S		19.66	****		58.99	****		5.33	**
Line		26.62	****		18.11	****		9.95	****
C*L		1.27	ns		2.58	****		1.47	*
S*L		2.50	***		3.57	****		3.04	****
C*S*L		1.01	ns		1.89	**		1.14	ns
$V_{rep(C*L)}$		0.0127			0.0065			0.0100	
$V_{sex*rep(C*L)}$		0.0028			0.0007			0.0003	
$V_{residual}$		0.0517			0.0585			0.0383	
	W4			W5			W6		
Cross		215.90	****		95.18	****		74.49	****
Sex		659.99	****		0.02	ns		92.78	****
C*S		46.29	****		14.73	****		9.24	****
Line		12.75	****		14.97	****		21.82	****
C*L		1.57	*		2.48	****		1.79	**
S*L		3.62	****		1.92	*		1.79	*
C*S*L		1.28	ns		1.81	**		1.21	ns
$V_{rep(C*L)}$		0.0056			0.0018			0.0033	
$V_{sex*rep(C*L)}$		0.0012			0.0000			0.0005	
$V_{residual}$		0.0228			0.0214			0.0226	
	W7			W8			W9		
Cross		11.26	****		32.48	****		136.37	****
Sex		77.55	****		0.00	ns		77.09	****
C*S		9.08	****		15.27	****		0.70	ns
Line		15.41	****		40.60	****		8.73	****
C*L		1.57	*		1.77	**		1.63	*
S*L		2.17	**		2.42	***		1.58	ns
C*S*L		1.04	ns		1.32	ns		1.12	ns
$V_{rep(C*L)}$		0.0013			0.0005			0.0007	
$V_{sex*rep(C*L)}$		0.0000			0.0002			0.0002	
$V_{residual}$		0.0164			0.0061			0.0063	

Variance components multiplied by 1000.
Significance: "ns" P > 0.05, * P < 0.05, ** P <0.01, *** P <0.001, **** P < 0.0001.

Appendix L. (continued 2)

	B Area	F	*P*	C Area	F	*P*	D Area	F	*P*
Cross		28.50	****		39.18	****		37.97	****
Sex		9237.81	****		8828.04	****		9845.90	****
C*S		14.27	****		4.18	**		4.06	**
Line		28.05	****		31.76	****		22.18	****
C*L		2.20	***		2.11	***		2.55	****
S*L		2.27	**		2.41	***		1.80	*
C*S*L		1.11	ns		1.05	ns		1.21	ns
$V_{rep(C*L)}$		0.0540			0.0379			0.0280	
$V_{sex*rep(C*L)}$		0.0084			0.0062			0.0050	
$V_{residual}$		0.1473			0.1104			0.0871	
	T Area			L1					
Cross		26.53	****		31.81	****			
Sex		11185.23	****		12585.19	****			
C*S		6.46	***		2.45	ns			
Line		28.41	****		26.53	****			
C*L		2.28	****		2.00	***			
S*L		2.04	**		1.24	ns			
C*S*L		1.15	ns		1.10	ns			
$V_{rep(C*L)}$		0.3318			0.0199				
$V_{sex*rep(C*L)}$		0.0522			0.0034				
$V_{residual}$		0.8100			0.0469				

Significance: “ns” P > 0.05, * P < 0.05, ** P <0.01, *** P <0.001, **** P < 0.0001.

Appendix M. ANOVA tables of repeatability tests with Kenyan test cross.

Source	Trait	Site	F	P	Trait	Site	F	P
Cross	T Area	T31656C	2.71	0.07932	C1	C30200T	126.30	0.00000
Sex			583.59	0.00000			6.03	0.03191
SNP			0.27	0.61263			7.72	**0.01796**
Cross*SNP			0.87	0.42611			0.34	0.71329
SNP*Sex			0.11	0.74644			0.01	0.92609
Cross*Sex			0.39	0.68261			8.54	0.00095
Cross*SNP*Sex			0.15	0.86152			0.01	0.98581
Cross	C2	C31634T	31.52	0.00000	T Area	T40722C	NA	
Sex			163.72	0.00000				
SNP			0.41	0.53630				
Cross*SNP			1.31	0.28400				
SNP*Sex			0.21	0.65610				
Cross*Sex			0.72	0.49616				
Cross*SNP*Sex			0.02	0.97831				
Cross	D1	T39389C	322.57	0.00000	W9	5683del1	ND	
Sex			66.95	0.00000				
SNP			0.01	0.91484				
Cross*SNP			2.55	0.08472				
SNP*Sex			0.16	0.69573				
Cross*Sex			0.10	0.90797				
Cross*SNP*Sex			1.23	0.29830				
Cross	T Area	G30401A	NS		W7	C30505A	3.84	0.03115
Sex							19.74	0.00099
SNP							3.09	0.10652
Cross*SNP							2.98	0.06355
SNP*Sex							0.81	0.38757
Cross*Sex							1.87	0.16849
Cross*SNP*Sex							0.13	0.87752

Terms including SNP are highlighted if *p*-value is lower than 0.05.
NA: Site 40722 is not segregating in the Kenyan sample
ND: Area surrounding site 5683 was not sampled in the Kenyan population
NS: Site 30401 is segregating at low frequency Kenyan, not scored.