Large Scale Genomic Association Studies in Fruit Fly and Human

Dace Ruklisa

Magdalene College

University of Cambridge

A dissertation submitted to
the University of Cambridge for the degree of

Doctor of Philosophy

January 4, 2011

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Acknowledgements

Before mentioning many important and interesting people that I have met in Cambridge, I would like to express my gratitude to my first mentor in science Janis Barzdins. My presence in computational biology would be impossible without the advice of my masters thesis supervisor Juris Viksna. At last there are serious doubts that my mind would be turned towards biological problems and doing a PhD without the assistance of my mother Maija.

I am indebted to the European Bioinformatics Institute for immediately throwing me from purely theoretical research environment into the melting pot of exciting biology. First and foremost, it is the achievement of Ewan, who had plenty of enthusiasm, more data sets than I could possibly analyse within 4 years and indestructible faith in people with mathematical background, even when the author of this thesis was sent to do experiments with her own hands. Overall it has been a mind broadening experience of working side by side with people from very diverse research backgrounds and different ways of attacking open problems in science.

Discussions within my research group and their support have been important both for shaping ideas and getting over the valleys of tedious technical work. Many thanks to Markus, Mikhail, Daniel, Alison, Michael, Sander, Albert, Andre, Petra and Marcel for creating nice and creative atmosphere in the office throughout my PhD time.

Many subtle adjustments to my PhD track were made after annual meetings with my Thesis Advisory Committee. Thanks to Anne-Claude Gavin for updating me about the newest research in network analysis, to Manolis Dermitzakis for helping me to put my discoveries

into the framework of population genetics and to Wolfgang Huber for suggestions of how to improve statistical models.

The association study for early developmental phenotypes in fruit fly was done in close collaboration with EMBL Heidelberg and particularly the Ephrussi group. Weeks spent in Heidelberg were extremely inspirational, intense and productive, both due to the large amount of research seminars and the novelty of the experience of working with microscopes and doing experiments. Thanks to Anne, Imre, Virginie, Ronald, Andrea, Proteeti and Sanjay for hosting me in the lab and, last but not least, to Anna for guiding me through the experimental procedures.

For most of my PhD I have been a member of the Marie Curie Actions programme and the Natural Killer Cell Defense Consortium. This grant came both with generous funding for my research and many opportunities to participate in workshops together with wet lab biologists, extensively discussing T cells, and thus encouraging my understanding of type 1 diabetes. Thanks to Bernard de Bono for coordinating and supervising the European Bioinformatics Institute segment of this programme.

Another important source of funding for my research has been a grant from European Rat Tools for Functional Genomics (EURATools). It is a bit of a pity that I did not find epistatic interactions for heart disease related phenotypes in rats. Nevertheless a large part of association study methodology and tools that were successfully used later were developed while being supported by this grant (thanks to Xose Fernandez for managing it).

Association studies with dense SNP maps are unimaginable without the plethora of matrix operations necessary for fitting statistical models, requiring days and even weeks of computing time. I am extremely grateful to the EBI Systems group, especially Rich Boyce and Gavin Kelman, for setting up a distributed computing cluster ("Hadoop cluster") that for many months had a single privileged user.

Throughout the years of PhD, I have received a lot of support and constructive criticism of methodological aspects of my research from my former colleagues in Riga. Thanks to Karlis Cerans, Juris Viksna, Karlis Podnieks and Viesturs Zarins for their readiness to discuss questions of computational biology, willingness to organize research seminars and invitations to "The Eagle" whenever they happen to be in Cambridge.

Important contributions to my profound appreciation of statistical models have been made by an extremely strong group of Magdalene college mathematicians, who have a habit to spontaneously gather for dinners at 6.30pm at Ramsay hall. Thanks to Arum, Ashok, Omri, Tamara and Jonathan for their encouraging idealism and belief in the usefulness of deeper theoretical understanding of statistics. I have to confess that this thorny though interesting path has sometimes unexpectedly led to exciting books in functional analysis.

The life at the European Bioinformatics Institute would not have been nearly as enjoyable as it was without its well organized PhD student community. It has been a true pleasure to drink coffee at 9.05am and then again at 14.55pm along with discussions ranging from the early symphonic works of Benjamin Britten to the models explaining the role of Calmodulin 1. Cocktail parties and dinners in the company of Melanie, Julia, Anika, Heidi, Nenad, Joern, Steve, Jacky, Angela, Afonso, Sigi and Felix will be greatly missed in my next academic stop.

A PhD is always a mixture of creativity and endurance. An important role in both phases has been played by my closest people, Sophia, Arina, Egita, Magdalena, Alexander and my parents Maija and Indulis.

Abstract

Genome wide association studies are very important to systematically bridge the knowledge of genotype and phenotype. With the recent increase in the genetic marker density for many organisms novel opportunities for a fine mapping of genetic contributions to complex traits emerge. This comes together with new challenges towards the modeling of phenotypes and also computing required to fit many statistical models.

I developed methods to construct large composite models for genotype to phenotype association that used multiple markers. Larger models should provide more accurate picture of joint effects of various loci and also their relative importance for explaining the phenotype.

Oocyte development in *Drosophila melanogaster* is an interesting model of developmental processes and their timing and can be easily studied using microscopy. I collected a set of phenotypes for oogenesis and defined traits for further association study via analysis of microscope images of various developmental stages. Descriptors of nurse cells were extracted from image channels showing the patterns of antibody staining (DAPI, Actin, *oskar*).

I then performed a genome wide association study of these phenotypes using my tools to characterize genetic components in oocyte development in fruit fly. This is one of the very few studies for fruit fly exploiting dense SNP maps (approximately 6,000,000 entities) and also incorporating multiple traits.

My methods were applied for reanalysis of the human case/control data for type 1 diabetes (T1D) from the WTCCC. T1D is a common

disease with rather complicated aetiology, therefore a good candidate for building composite models. Random sampling approaches allowed further investigations of these complex joint models.

Contents

List of Figures

Chapter 1

Introduction

1.1 Traits and Heritability

Each individual can be characterized in many ways, for example, by height, heart mass or having or not having influenza. If we look at a population of individuals we can observe variation in all of these traits: individual heights differ, some people have influenza, while others do not and there are variable measurements of heart mass reported by clinical sources. Some of this variation is continuous, such as height and heart mass, and some is discrete, such as having or not having an influenza. It is possible to categorize variation underlying a trait into two broad groups: genetic variation and environmental variation. The genetic component of a variation constitutes heritable part of a trait, while the environmental component consists of influences upon a trait that are changing from place to place and from living condition to living condition. As the name implies genetic factors are composed of a combination of genes, where each gene occupies a particular site on the genome called locus. Different forms of a gene are called alleles. A set of particular alleles for all loci of an individual is referred to as the genotype of individual. If a gene influences a trait as a genetic factor, each type of allele changes the value of a trait.

The relative importance of genetic versus environmental factors can be very different between traits. For example, influenza is caused by a type of virus, whose occurrence and spread strongly depends on environmental factors, such as climate. Other traits have a tendency to be passed from one generation to another or are especially pronounced in some population whose members have been living in the same place for a long time. Several diseases can serve as example of this type of trait, such a cystic fibrosis that is known to be largely determined by the presence of a particular allele and is nearly always inherited. In between highly heritable and purely environmentally determined characteristics there is a large complex zone of traits whose heritable and environmental determinants require careful elucidation. For example, heart mass is determined both genetically and environmentally. In all of these situations determining the heritable component is indispensable for understanding risk factors of disease (elevated blood pressure, rheumatoid arthritis) or knowing that disease can be eliminated by removal of environmental factors (influenza) or by being able to cultivate favourable properties (height of maize).

The major part of this thesis is devoted to understanding the genetic components underlying various traits: the first of these are the traits characterizing early development in fruit fly and the second is type 1 diabetes. The study of early development in fruit fly focuses on oogenesis and particularly nurse cell development throughout oogenesis. Several stages of oogenesis were studied and phenotyped using optical microscopy and image analysis techniques. Afterwards, statistical analysis was carried out to search for loci on the genome associated with developmental traits. Yet another study was carried out to discover genetic factors affecting human disease, particularly type 1 diabetes (T1D). Case/control data from the Wellcome Trust Case Control Consortium (WTCCC) were reanalysed using novel approaches to yield more insights into aetiology of T1D.

The remainder of the introduction is structured in the following way. In the next section of introduction I will talk about the nature of genetic components contributing to various traits in more detail. Then in section 1.3 I will discuss the experimental designs and types of populations used to elucidate heritability. Section 1.4 is devoted to introducing techniques of statistical modeling and model selection used to analyse various populations. A section encompassing an overview of association studies for *Drosophila* phenotypes follows (1.5). A discussion of factors affecting dorsal-ventral axis specification during oogenesis is included in section 1.6. Section 1.7 is dedicated to the description of the genetic components influencing type 1 diabetes. The introduction ends by a statement of the main questions asked in the thesis (section 1.8).

1.2 Genetic Components Influencing Traits

In eukaryotes, genes are organized into chromosomes, where each chromosome represents a single linkage group (Sturtevant et al., 1919). After chromosomes were established as the vectors of heredity Avery showed via studies of the virulence of bacteria that heritability is conveyed by DNA (Avery, 1915). This discovery was followed by the elucidation of the DNA structure and the postulation that DNA is a double helix made from two polymers of nucleic acids arranged in an antiparallel manner, where nucleotides from opposite strands form base pairs (Watson and Crick, 1953). Thus, each chromosome corresponds to a single double stranded DNA molecule. In most species there are two chromosomes of the same type, therefore each unit of heritability (locus) on a chromosome has two different alleles. A violation to this principle occurs when sex chromosomes are considered, because there can be two different sex chromosomes united in a single pair, where the combination of chromosome types determines sex. Chromosomes that do not determine sex are called autosomes. The number of chromosomes can vary widely among species; for example, human has 24 types of chromosomes arranged in 23 pairs, while fruit fly has 4 different chromosomes (2, 3, 4 and X). An offspring receives one autosome of each type plus one sex chromosome from each parent. In such a way it is possible to produce 4 different offspring genotypes from a fixed pair of parents. More variation and randomness is introduced into the assortment of offspring's genotype by recombination, which can exchange continuous part of a chromosome coming from one grandparent for that of another grandparent.

In classical genetics, the heritability of a trait is often studied experimentally by breeding experiments, if they are feasible for the species under study. The necessary condition for these experiments to be successful is having two parents with different phenotypes so that the alleles potentially responsible for occurrence of a trait are different as well. First a generation of individuals, called F_1, is obtained by crossing both parents. Often F_1 individuals are crossed (sibling-sibling mating) and phenotypes of offsprings (F_2 generation) are assessed. If all offspring in F_1 generation yield phenotype equivalent to one of the parents' phenotypes, the trait observed is called dominant, while trait unobserved is claimed to be recessive. If further offspring from generation F_2 yield two different phenotypes in proportions $1 : 3$, it is clear (or was clear for Gregor Mendel) that one of the two alleles affecting the trait is recessive, while another is dominant. Then the quarter of offspring with recessive trait corresponds to the allele combination aa, while the dominant trait is produced by all other allele combinations (Aa and AA). These experiments lead to postulation of principle of independent segregation that implies that dominant and recessive alleles are independently transmitted and segregate independently, thus leading to sudden reemerging of the recessive trait. This kind of breeding experiment is suitable for dealing with phenotypes affected by a single gene. However, it is rarely the case, because many traits are affected by several genes. A common assumption is that for a complex phenotype, the genetic effect would be the sum of several independent (main) effects of different loci and epistatic interactions of loci. Epistasis is defined as the interaction between genes, where certain allele at one locus is altering or masking the effect of alleles of another gene (Cordell, 2002). Loci that affect such a complex phenotype are called quantitative trait loci (QTLs), while the phenotype is referred to as a quantitative trait. Often gene-environment interactions contribute to phenotype as well, though they are difficult to study, because phenotypes have to be collected for each of different environments, and comprehensive control of the environment can be challenging (Falconer and Mackay, 1996).

Typically studies of heritability involve two major design decisions: first is that of choice of population under study, while second is that of choice of statistical models. In the above mentioned Mendelian case, the population would be two different parents and their offspring. Current experimental designs include selection of individuals affected by a disease versus a selection of individuals without disease (case-control studies), animal strains created by controlled breeding (e.g. inbred lines, populations selected for a particular trait), sets of families and sibling pairs to determine how relatives affected by a disease differ from those unaffected (family based studies) or sample from a population living in a certain environment, to observe the time of onset of a disease (prospective studies).

Each study design has to be supported by models that allow disentangling genetic part of a trait from the environmental part. The simplest models deal with a single locus, while more complicated traits require multi-locus modeling techniques. Typically many loci would be tested on each chromosome to detect which of them is the most important determinant of a trait. Therefore an important part of the modeling is the assessment of the significance of association between the alleles of a locus and a trait.

1.3 Elucidation of Heritability

1.3.1 Studies of Experimentally Designed Populations

One of the simplest experimental designs used to produce a population with certain favourable properties exploits backcross (BC) populations. To obtain such a population two phenotypically and genetically distinct founders are crossed. Offspring from the F_1 generation are collected and then crossed back to one of the parents. If the founders come from inbred lines, the backcross population obtained thus has more genetic diversity than any of the two founder populations. The new BC population is enhanced to develop traits similar to the founder not used to backcross F_1 (donor line). As a result, BC population is more amenable to genotype and phenotype studies, because there is larger variation in alleles and possible improvement in phenotypic properties towards desirable traits. BC lines are widely studied with an aim to identify QTLs in plants (tomato, rice, soybean) (Moncada et al., 2001; Tanksley and Nelson, 1996; Tanksley et al., 1996; Wang et al., 2004); they are used to cultivate certain characteristics of plants like yield, height and maturity. BC lines are important to introduce more genetic variation into inbred lines of domesticated species. For some studies advanced backcross lines have been developed by continued backcrossing of BC populations to the same founder thus obtaining BC_2 and BC_3 generations. This method solves the problem of introducing too many disadvantageous alleles from donor line, because their frequency is reduced due to negative selection and recombination in more advanced backcross generations (Tanksley and Nelson, 1996). Backcross populations are not as good in detecting recessive alleles and epistatic interactions in comparison with other study designs and other types of allele effect upon phenotype (Tanksley and Nelson, 1996).

Alternative types of breeding studies are based on selective breeding to produce two highly divergent populations in terms of the trait under study. Usually two different inbred lines serve as founders in these studies. After obtaining F_2 progeny, selection favouring one or both extremes of trait begins. First, F_2 individuals are tested for the phenotype, then a fixed percentage of individuals exhibiting the highest scores of trait are selected; also a certain number of lowest scoring individuals can be selected in parallel. Animals from selected subpopulations are mated within the selected population, thus a new high and new low population is obtained. Then individuals from high population are scored for phenotype and a fixed percentage with highest scores are selected for the next breeding. The lowest scoring individuals from low line are also selected for further breeding. Usually sibling-sibling matings are avoided. The selection from high and low lines is repeated for a few generations until both lines become highly divergent for a phenotype.

Such a selection should highlight associated loci due to change of allele frequencies observed in selected and unselected populations. Mainly QTLs exhibiting additive effects are becoming evident thus, because other models would not have allele frequencies changing in concordance with phenotype divergence. The method is difficult to use for studying multiple traits for obvious reasons. Nevertheless, genetic correlation of related traits can be effectively studied with this method. Besides, analysis requires a small number of generations (usually of order 10) to be contrasted to demands for establishing recombinant inbred lines (see next section).

Applications of selective breeding studies include mapping QTLs responsible for ethanol preference in mice (Belknap et al., 1997). Another study encompassing selected lines of animals lead to the discovery of several QTLs affecting growth related traits in chicken (Jacobsson et al., 2005).

In some cases, establishing of inbred lines is not feasible either because it is too time consuming for some animal species or selection histories of both founders are too divergent. Sometimes these problems can be solved by crossing two genetically and phenotypically distinct individuals each coming from a different outbred population to establish a new population. Typically QTLs would be mapped for a F_2 generation resulting from a cross between two outbred population representatives. Information about genotypes of F_1 individuals and founders is necessary in search for QTLs as well. Crosses derived from outbred population individuals have been successfully used to study growth rate and fat deposition in pigs, where founders represent European wild boar and Large White pig (Knott et al., 1998). Similar studies of pig intercrosses (Chinese Meishan and European domestic; Iberian and European white domestic; Iberian and Landrace) have been carried out by (Ovilo et al., 2000; Perez-Enciso et al., 2000; Rohrer and Keele, 1998a,b). Commercial pig populations have been examined for QTLs related to back fat and meat quality traits by (Evans et al., 2003). In general, studies of commercial populations are more complicated than studies of intercrosses between outbred populations because many loci can become fixed due to artificial selection imposed on commercial lines. However, in the example mentioned above, major QTLs were still segregating. Preselection of the most informative animals in terms of their phenotypes is sometimes necessary to overcome limitations of fixed alleles (Chatziplis and Haley, 2000).

1.3.2 Recombinant Inbred Lines

A recombinant inbred (RI) line is an inbred line whose genome is a mosaic of the genomes of the founder lines. To create a RI line two highly divergent strains are crossed until F_2 generation is obtained and then brother-sister mating is done for several generations until most loci are fixed for some homozygous allele combination. After this point RI lines are established and their (almost) exact genotypes can be reproduced infinitely by further matings among siblings.

RI lines have the favourable properties of having fixed genotype that can be reproduced over next generations. They are widely used in genotype/phenotype studies, because many phenotypic measurements can be taken for the same genotype, including those for multiple traits.

Using such lines is beneficial for observing more distinct or extreme phenotypes as there are no heterozygotes and thus no intermediate phenotypes. Also the task of model selection for QTL mapping becomes considerably simpler without the need to fit recessive and dominant models. RI lines are good for studying traits that are sensitive to environmental variation as well (Soller and Beckmann, 1990). However, many homozygous mutations can be lethal and thus remain unobserved (complementation tests can sometimes solve the problem as in (Fanara et al., 2002)). Also it is not possible to gain more detailed information about the nature of QTL's influence on phenotype as we do not know the intermediate phenotype of heterozygote. Besides, one has to be careful in selection of line founders, because only variation due to the alleles that are segregating in parental strains can be captured.

Variance of a trait in RI population is twice the variation that can be obtained in F_2 population and four times the potential variance of a backcross (BC) population (Kearsey and Pooni, 1996). Heritability itself is a bit less than twofold greater, but still more noticeable in RI lines (Belknap, 1998).

1.3.3 Linkage Studies

There are situations when experimental design of populations with specific properties is not feasible, most notably human studies. In such cases different approach towards studying heritability has to be pursued. Some methods devised for dealing with non-experimental populations use information about genotypes within family (linkage studies), while others use genotypic information sampled from a population in a certain way (association studies).

Linkage studies search for genome regions that are shared by individuals affected by a disease within a family. The expected frequency of genome region in the absence of association with a disease is estimated against the background of family structure and compared to the observed allele frequency for individuals having a disease. Siblings, parents and other family members take the role of controls for individuals with a disease. It is difficult to fine-map variation responsible for a trait because there are not as many recombination events happening within the number of generations studied (typically information about genotype is available only for a few generations).

Linkage studies are important when dealing with rare alleles that are associated with a high risk of a disease and are appropriate for situations when a disease or trait has high penetrance (otherwise there is a very small proportion of affected pedigree). The advantage of this type of studies is their immunity towards population stratification because all tests are conditioned on the alleles and their frequency of healthy relatives.

Linkage studies are limited to cases where alleles of founders are segregating. It means that we can estimate a disease penetrance only when both parents are heterozygous for the same loci, otherwise we cannot spot events of recombination and heritability. Therefore a large proportion of genetic variation cannot be studied with this design, because of lack of heterozygosity. Sometimes this type of study cannot be carried out due to a late onset of a disease and difficulty in collecting information about relatives.

Genome wide linkage analysis was first successfully done for Huntingdon disease, allowing the identification of the chromosomal location of a crucial mutation (Gusella et al., 1983). Several linkage studies have helped to elucidate a locus on chromosome 7 as being connected to cystic fibrosis (McConkie-Rosell et al., 1989; Tsui et al., 1985); mutations of several alleles of *CFTR* turned out to be a major causative factor for classic cystic fibrosis. There are now over 400 such rare diseases with known alleles that cause disease, and at least 2,000 other diseases with a clear monogenic inheritance pattern being actively studied.

1.3.4 Association Studies

Association studies search for loci associated with a trait or disease within a population. Ideally, a random sample from a population has to be taken and genotype/phenotype scored. Types of association studies include case/control and prospective studies. In the first situation a certain number of cases of a disease are taken and compared to a number of controls (individuals without disease). In the second situation a number of individuals are observed over time with an aim to detect the onset of a disease. A separate category of association studies deals with continuous (quantitative) traits.

Association studies can be divided in two types: candidate SNP analysis and indirect association studies. Candidate SNP analysis directly tests whether a SNP is related to an increased disease risk, while indirect association tests assume that risk polymorphism is located somewhere where it is in strong linkage disequilibrium (LD) with some genotyped SNP. Linkage disequilibrium is defined as the strength of non-random association between alleles at different loci (Futuyma, 1997). If two alleles at two different loci arose at roughly the same historical time, and they are both on the same chromosome and have little (if any) recombination between them then they are described as being in linkage disequilibrium. Then one of the alleles can be reliably inferred from the other allele. The requirement that the recombination between the loci is rare means that the loci must be relatively close to each other on the genome, though this distance varies due to the variation of recombination rates.

The applicability of direct and indirect association studies is dependent on the organism studied and properties of its genome. For example, *Drosophila melanogaster* genome exhibits very short linkage disequilibrium, therefore dense SNP maps and direct association studies are necessary.

Association studies have to be performed with greater marker densities than linkage studies, because of many recombination events taking place within a population, when measured in the context of the whole population. It has been estimated from coalescent simulations that for most human populations, strong linkage disequilibrium can be observed over approximately 3 kilobases (kb) (Kruglyak, 1999). It means that having maps with as few as 500,000 SNPs is needed for whole genome association scans.

The genome wide scan of SNPs to detect associations was first proposed by (Risch and Merikangas, 1996). It is also argued that such association studies are more powerful that linkage approaches (Risch and Merikangas, 1996). This hypothesis was confirmed by population simulations in (Long and Langley, 1999).

Originally there were two hypotheses about the origins and mode of operation of genotype and phenotype associations. The first called ”rare variant hypothesis” postulates that a common disease is caused by the summation of effects of low frequency alleles, where each of them affects disease risk independently. The second called ”common disease common variant hypothesis” asserts that there should be at least one common variant affecting a common disease. Whole genome association studies are well suited for finding common variants affecting the disease, while rare variants can be discovered by resequencing of candidate genes for many individuals (Bodmer and Bonilla, 2008). If a rare variant is discovered to be associated with disease, it is very unlikely that the true association is due to some other variant in close linkage disequilibrium. If a common variant is discovered, it is quite probable that it is not immediately functionally relevant to the trait studied and other variants within linkage disequilibrium have to be checked.

One of the most comprehensive association studies for human disease has been done by WTCCC (Wellcome Trust Case Control Consortium, 2007). Within this study genome wide scans to search for common variants causative of 7 common diseases, most of them autoimmune diseases like type 1 diabetes, have been performed. The study exploits approximately 500,000 SNPs, 3000 shared controls and 2000 cases for each disease.

Extra care has to be taken for the population stratification when performing association studies, because unlike in linkage studies genotypes of individuals are not conditioned upon their parents' genotypes, i.e., their population ancestry. For example, non-synonymous SNPs were tested for association with T1D and inflation of p-values over 10% was found to be due to confounding by population structure (Clayton et al., 2005). Adjustments for population structure present in case/control studies can be done by genomic control, i.e., rescaling traditional chi-square test results by a constant factor (Devlin and Roeder, 1999). Other approaches reconstruct the population structure by assigning population labels to individuals (Satten et al., 2001).

An important component of association studies is meta-analysis of reported associations. Such an analysis is done with an aim to estimate the ratio of false-positive results, assess the success of replication of the first study and estimate the publication bias, along with other parameters. Replication studies for reported associations and thus meta-analysis of them are necessary for two reasons. First, each association study reports loci that are brought to forefront by stochastic fluctuations due to individuals genotyped and genotyping errors, and are randomly selected from a pool of associations of comparable strength. Second, there can be hidden factors present that are not taken into account in the original association study. A special case of such hidden factor is population stratification. Therefore it is important to test reported associations with independent sample of individuals. A meta-analysis of many loci connected to human disease, especially those associated with type 2 diabetes, bipolar disorder and schizophrenia, was done in (Lohmueller et al., 2003).

1.4 Modeling Prerequisites

1.4.1 Model Types

Various approaches can be adopted to express the connection between genotype and phenotype. The additive model was first suggested by R.A.Fisher, who proposed to represent genotype for an individual at certain locus with one variable:

$$X_a = 2 \cdot I(x = AA) + I(x = Aa) \tag{1.1}$$

where X_a is the code for an additive effect assuming values 0, 1 and 2 and corresponding to the count of allele A within diallelic locus. $I(\cdot)$ stands for an indicator function yielding 1 if the condition in the brackets is satisfied (in the current case 1 is returned, if locus x has a particular genotype). The additive effect is essentially the sum of main effects of both alleles, where it is implicitly assumed that phase is not important and effect of allele A is indistinguishable in either position. The dominance effect is usually understood as the interaction of both alleles within a diallelic locus in such a way that one occurrence of allele A is enough to annihilate the impact of the alternative allele a: $X_d = 1 - I(x = aa)$. Then the full genotypic model for a single locus and its impact upon phenotype Y would read like this:

$$Y = \mu + \beta_1 X_a + \beta_2 X_d + \varepsilon \tag{1.2}$$

where β_1 is the magnitude of an additive effect and β_2 is the size of a dominance effect. Then μ is the population mean of the phenotype under study for individuals with genotype aa and ε is random variation of phenotype that cannot be explained by effects of the alleles considered. This type of model is used for continuous traits (continuous Y). If we want to model discrete traits, like disease status, we have to transform the phenotype via monotone link function $g(\cdot)$ (Davison, 2003):

$$\eta = g(E(Y)) = \mu + \beta_1 X_a + \beta_2 X_d \tag{1.3}$$

For binary outcome Y, the meaning of $E(Y)$ is the probability that the phenotype is one of the outcomes, for example, probability of individual having a disease. Various link functions are used in literature to discover associations, among them logit, probit, logarithm, log-log and complementary log-log functions. The performance of these transformations in disease prediction has been analysed in (Cordell et al., 2001). It was shown that the differences between results obtained are not large, especially between logit and probit functions. Further extensions of the generalized linear model model (1.3) necessary to incorporate epistatic interactions are discussed in (Cordell and Clayton, 2002).

1.4.2 Interval Mapping

An important aspect to decide when performing association studies is how to search for associations and how to model the effect of several loci upon phenotype simultaneously. One of the first methods proposed for systematic search for associated loci was interval mapping (Lander and Botstein, 1989). Within this approach an unknown genotype of a QTL located between two known markers is expressed as a function of two flanking marker genotypes and frequency of recombination between markers and marker and QTL. It is implicitly assumed that there is at most one recombination event between both markers, because 2 recombinations cannot be distinguished from 0. Interval mapping has been approximated by regression in (Haley and Knott, 1992).

The power of discovery of QTLs with interval mapping under various conditions (marker spacing, magnitude of allele effect) was studied in (Darvasi et al., 1993). It turns out that the power of interval mapping and t-test applied for a single marker is not very different (slight advantage towards interval mapping) (Darvasi et al., 1993; Haley and Knott, 1992). Interval mapping suffers more from non-normally distributed traits than direct marker tests (Darvasi et al., 1993).

Soon afterwards it was acknowledged that it is not sufficient to study one QTL at a time, because effects of other QTLs can interfere with the estimation of a single QTL effect (Haley and Knott, 1992; Knapp, 1991; Martinez and Curnow, 1992). For this reason several QTLs were incorporated in a single model to make the effect estimation for multiple loci simultaneous (Haley and Knott, 1992; Jansen, 1992; Knapp, 1991; Martinez and Curnow, 1992). However, it is computationally difficult to model several QTLs with the approach resembling interval mapping. Yet another approach was proposed that exploits linear regression, where phenotype is regressed on marker genotypes, without an explicit reference to QTLs between them (Cowen, 1989; Stam, 1991). At last it was suggested to merge the idea of regression upon marker genotypes and the idea of interval mapping for a QTL into a single model including several markers and only one QTL (Jansen, 1993; Zeng, 1993, 1994). This method was called composite interval mapping (CIM). Under this approach a set of background markers is first selected. Then a QTL is added to the model in a way similar to interval mapping. Selected markers take the role of other QTLs, while one specific QTL is studied. Then a significance of QTL can be assessed by referring to chi-square distribution (it is asymptotic distribution for log likelihood ratio for a model with QTL versus model with markers, but without QTL) (Jansen and Stam, 1994). Alternative methods to composite interval mapping based on Bayesian models were proposed in (Satapogan et al., 1996; Sillanpaa and Arjas, 1998).

Composite interval mapping was further extended into the multiple interval mapping (MIM) method that incorporates epistasis along with main effects (Kao et al., 1999). Epistatic interactions are described by Cockerham's scales. This approach has the property that all of the components describing epistasis are orthogonal. Thus economy in model parameters is achieved (4 instead of 9 unknown coefficients for each interaction). The model includes multiple QTLs, their genotypes expressed via flanking marker genotypes and recombination rates. A set of QTLs is selected by stepwise selection procedures (Kao et al., 1999). Parameters are estimated by a method proposed by (Kao and Zeng, 1997) and is based on maximum likelihood estimates. The approach chosen for parameter estimation avoids expectation maximization for all of the unknown QTL genotypes, therefore it is much more efficient computationally.

An interesting question regarding QTL mapping is how sure we can be about the discovered location of a QTL. The answer to this question has impact on follow up studies, where QTLs are fine mapped and candidate genes are analysed. This problem has been studied by estimating the resolving power for QTL location, where resolving power is defined as 95% confidence interval for location, when a map is infinitely dense (Darvasi and Soller, 1997). An alternative method for obtaining confidence intervals via bootstrapping was provided in (Visscher et al., 1996).

Initial models were built with sparser marker maps in mind, therefore they distinguished the notion of marker and the notion of QTL that is located somewhere between two markers and whose location has to be fine tuned by additional search. As genetic maps become denser, substitution of QTLs by genotyped loci yield more and more accurate results, while helping to avoid multidimensional genome search. Such a progress naturally leads towards some simplification of models mentioned above, because genotypes of all loci can be included in the model directly without the need to express any of them indirectly via recombination rates and LD with known markers.

1.4.3 Testing Significance of Allele Effect

Each type of population and study design chosen to elucidate heritability requires different type of statistical tests to determine the significance of the effect of specific alleles.

Statistical tests based on likelihood ratio methods for outbred populations are developed in (Haley et al., 1994). The approach is based on least squares method applied to all markers within a linkage group. Before least squares method is used, probabilities of origins of alleles for loci in terms of outbred grandparents' alleles are deduced. An assumption is made that the outbred populations analysed are fixed for alternative alleles.

Studies of populations constructed by selective breeding are more complicated, as no interval mapping method directly applies here. Adjustments of LOD scores that are applicable as significance thresholds for detecting QTLs in later generations of selection are proposed in (Falconer, 1989; Lander and Botstein, 1989).

Typically the power to detect QTLs is lost over generations (Darvasi and Soller, 1995). However, using generations beyond F_2 is advantageous as more recombination events can take place thus affecting linkage disequilibrium. Therefore the location of QTLs can be mapped with much higher precision in these populations (Darvasi and Soller, 1995).

When RI lines are used, single marker tests are more powerful than interval mapping, because RI lines exhibit a lot of recombination in any interval between two loci. The power for detecting QTLs in F_2 populations is analysed in (Darvasi and Soller, 1992). For RI lines, similar studies have been carried out by (Belknap et al., 1996).

One of the most widely used tests for linkage studies within families is transmission/disequilibrium test (TDT) (Spielman and Ewens, 1996). The model was first developed in (Spielman et al., 1993). The first version of TDT expressed genotypes of affected individuals conditioned on parents genotypes, thus requiring the existence of full information about parents. Mathematical models for linkage studies are elaborated in (Curtis, 1997; Falk and Rubinstein, 1987; Spielman et al., 1993). Tests using conditioning on genotypes of unaffected siblings instead of parents were developed later (Spielman and Ewens, 1998). This version of TDT is suitable for studying diseases with late onset. Further developments of the model, elaborating on the type of conditioning in case of missing data and availability of other relatives, include (Boehnke and Langefeld, 1998; Rabinowitz and Laird, 2000). In (Lange and Laird, 2002), many of TDT test types are put in a single statistical framework, thus unifying various approaches for conditioning. Asymptotic properties of TDT tests under null and alternative hypothesis can be assessed for rather broad classes of tests, thus allowing to obtain a measure of allele significance.

In most of the cases, applications of linkage studies are limited to testing simple additive, dominance or recessive models describing association with a trait. More complicated epistatic models and genotype-environment interactions cannot be traced by this approach, because there are too few individuals included in a single family to observe all combinations of interacting entities. Also non-segregating loci can become an obstacle for discovery of epistatic interactions.

In the case of epistasis, each allele separately has less pronounced effect on phenotype, probably also less penetrance, which is not a favourable situation for performing linkage studies.

The most typical statistical test used in association study setting together with additive, dominance or recessive models is the chi-square test with one degree of freedom. If phenotypes are binary, the Cochran-Armitage trend test can be used to test the association as well. This type of test has the advantage of not relying on the assumption of Hardy-Weinberg equilibrium (Devlin and Roeder, 1999). Trend test is immune for the effects of population stratification and cryptic relatedness of individuals, that in turn can lead to an excess of homozygotes due to correlation between both alleles (Devlin and Roeder, 1999). Under the same violations of conditions chi-square test p-values can become inflated. In this aspect the Cochran-Armitage test approaches tests used for family based studies.

Typically, independence of individuals within association study is assumed when performing statistical tests. Cases when this assumption does not hold are examined by (Devlin and Roeder, 1999) and the method based on Bayesian outlier model is proposed to detect associations under these circumstances, where associations themselves take the role of outliers (Devlin and Roeder, 1999). Still, having a completely independent sample leads to larger power in association discovery.

Indirect association studies are sometimes done with haplotypes instead of genotypes of single loci. Some conditions under which performing association studies with unphased genotype lead to more powerful tests than those with phased haplotype are discussed in (Clayton et al., 2004).

A regression model for association studies with haplotypes was proposed in (Zaykin et al., 2002). Modeling techniques for case/control studies with haplotype data have been explored in (Epstein and Satten, 2003; Stram et al., 2003; Zhao et al., 2003). The study of (Lin and Zeng, 2006) puts together several of the above mentioned techniques into a unified modeling approach, that can incorporate environmental variables and is not based on Hardy-Weinberg equilibrium assumption.

To attain more power when performing association studies with haplotypes, it has been suggested that the haplotype determination is united with estimation of haplotype impact on phenotype. Most popular approaches solving this simultaneous task incorporate methods based on expectation maximization and Newton-Raphson algorithm (Tregouet et al., 2002; Zaykin et al., 2002). Also, stochastic expectation maximization has been shown to be useful and efficient in this context due to its ability to avoid local minima and saddle points (Tregouet et al., 2004).

1.4.4 Tests and Multiple Loci

Specific attention has to be paid to the multiple testing problem within the association study context, because typically many loci would be tested during the search for an association. It means that p-values from tests performed for various loci cannot be taken literary, but instead have to be converted into a genome wide significance level.

Typically, adjustments of p-values are made by Bonferroni correction, which divides the p-value obtained by testing specific locus by the number of loci tested. In such a way genome wide significance of association is calculated from p-values of separate tests. For dense marker maps Bonferroni correction leads to rather stringent and conservative thresholds for declaring associations, because markers can be quite correlated and corresponding tests are not completely independent.

Earlier practice in estimating genome wide significance for linkage studies used LOD score 3 (base 10 logarithm of odds - equivalent to log-likelihood ratio of model corresponding to hypothesis of association against model for null hypothesis of no association) (Lander and Kruglyak, 1995). This score is approximately equivalent to p-value of 10^{-4} obtained from asymptotic distribution of log-likelihood ratios (chi-square distribution).

In case of Bayesian models correction is done via prior distributions. The equivalent of the association p-value is obtained by multiplying the evidence from genotype and phenotype information in favour of a chosen association model by the prior probability of the model (Gelman et al., 2004). To make adjustment for multiple tests, all prior probabilities of models can be chosen to be equal to the proportion of SNPs that are expected to be genuinely associated with a phenotype. This procedure resembles Bonferroni correction in some aspects, but is determined by the expected associations instead of the number of SNPs genotyped or models tested (Stephens and Balding, 2009). Prior probabilities can be much more flexible when correcting for multiple tests than Bonferroni correction, because they can be relatively larger for the models with loci showing previous evidence of association and smaller for models with loci of no known functional relevance to phenotype studied.

Permutation tests were suggested for assessment of significance threshold by (Churchill and Doerge, 1994). According to this method real phenotype values are permuted 1000 times (typically), and from each permutation a locus yielding the highest score of association is determined. Those loci whose level of association with the real phenotype is above the 95th quantile of the best permutation scores are claimed significant. This method produces significance thresholds that take into account the correlation of loci and other properties specific to a phenotype and genotype set. Therefore thresholds established by permutation tests are usually more accurate than those yielded by Bonferroni correction and are not as sensitive towards violation of various model assumptions.

A method similar to permutation tests is based on empirical estimation of false discovery rate. Again 1000 phenotype permutations are analysed and association scores for all models (loci) are obtained. Then a constant threshold is chosen and the number of permuted phenotype models scoring above the threshold versus the total number of models above the line is reported as false discovery rate (Benjamini and Hochberg, 1995). The threshold that gives significantly more associations with the real phenotype than associations with permuted phenotypes is used to report discovered associations.

Yet another alternative to permutation tests is cross validation (Hastie et al., 2008). The purpose of cross validation is not to give empirical p-value for associations, but rather to perform validation of models used to detect associations by trying to estimate the test error (error that will be obtained by analysing independent data set using the same model). Cross validation is typically used to determine the optimum number of loci to be included in large models and to estimate the most plausible parameters for large models. For example, the model selection part of interval mapping would be amenable to cross validation. N-fold cross-validation divides the data set in N non overlapping parts. Models are built with diminished data sets, where one of the parts is excluded. Then each of the models is used to estimate the prediction error for the part of data that was not considered during model building. Once the mean prediction error reaches minimum and stabilizes, it is clear that the model is saturated and optimum configuration is reached. Splitting into non-overlapping parts guarantees that the variance of the data sets and thus resulting models is high enough, therefore a small number of partitions is sufficient to determine prediction error and its variance (N can be as small as 5 or 10).

1.4.5 Model Selection Strategies

1.4.5.1 Stepwise Model Selection

The necessity to build models incorporating several loci was already discussed in the context of interval mapping. However, as soon as we are aiming at models with multiple markers or QTLs, we have to solve the problem of selecting entities to put into them. The first proposals included using backward elimination algorithm for selecting markers for interval mapping. It means that at the beginning all markers are included in the model. Then markers are removed one or several at a time until no further improvement in the model can be achieved by an extra removal (Davison, 2003). Goodness of models can be evaluated and compared by Akaike's information criterion (AIC) (Davison, 2003). However, the backward elimination algorithm is suitable only for sparse marker maps with few entities, otherwise the initial model becomes too large for simultaneous estimation of effects of all possible alleles. Also there is a high risk of matrices describing alleles being singular, especially in the presence of highly correlated loci.

For denser marker maps, other stepwise techniques have been suggested. Some methods begin with an "empty" model containing only intercept, but no loci, and gradually augment the model with more and more loci (forward stepwise regression). Some stepwise model selection strategies incorporate loci removal alongside adding. Stepwise methods can be described as greedy algorithms that traverse a path through sets of loci.

One of the disadvantages of the stepwise modeling approach is its deterministic nature that can lead to complete ignorance of loci yielding only slightly smaller model likelihoods in comparison with the best locus. In addition theoretical considerations indicate that models selected by adding and removing loci can soon overfit the data and thus are limited in the number of loci they can incorporate (Hastie et al., 2008).

1.4.5.2 Model Selection Based on Coefficient Shrinkage

The next class of model selection methods performs the shrinkage of coefficients describing the magnitude of allele effects by introducing penalty on the total size of genotype effects. By imposing extra limitations, it is possible to reduce some coefficients of effects to zero and deal with correlated loci by not allowing too many similar loci to have too much share in the final model (Hastie et al., 2008). Shrinkage of coefficients to zero is a continuous version of discrete backward elimination described in the previous section. Let us consider a quite generic model describing association between continuous phenotype and genotypes of several loci:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot X_{ij} + \varepsilon_i \tag{1.4}$$

where y_i is a phenotype for individual i, X_{ij} is a coefficient describing type of allele effect for individual i and locus j and β_j is the magnitude of effect of locus j, while β_0 is the population mean of the phenotype. Then ε_i is the random phenotypic variation for individual i that is not captured by the genetic component. Examples of X_{ij} include but are not limited to additive and dominance effects X_a and X_d defined in the Section 1.4.1. The task of all shrinkage methods is to estimate coefficients β so that the selection of loci is performed simultaneously via shrinkage. An example of shrinkage method is ridge regression. According to this approach coefficients are estimated from the following equation:

$$\beta^{ridge} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1.5}$$

where $(y_i - \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j)^2$ is the sum of squares of the model in (1.4), but $\sum_{j=1}^{p} \beta_j^2$ is the penalty term with complexity parameter λ controlling the amount of shrinkage. A shrinkage method of similar flavour is lasso, which imposes different penalty to the sum of squares estimate (Tibshirani, 1996):

$$\beta^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.6}$$

An interesting approach that unites benefits of ridge regression and lasso uses elastic net penalty:

$$\beta^{elNet} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \tag{1.7}$$

where α defines the mixture of both types of coefficient shrinkage (Zou and Hastie, 2005). For all of these models λ has to be chosen via some form of cross validation to achieve the lowest prediction error.

All shrinkage methods need a preselected set of features to start with. Only then can the coefficient adjustment take place. The dimensionality of association study problems is prohibitive for use of ridge regression or lasso for all genome loci directly (arguments are similar to the case of backward elimination algorithm). However, these methods can be useful for a post-processing of a smaller set of selected loci that already have shown some association with phenotype. For example, by a post-processing we can select a single representative locus from each association region.

1.4.5.3 Boosting and Forward Stagewise Linear Regression

One of the disadvantages of stepwise model selection methods are their greedy nature. We can easily imagine a situation when there are a few SNPs yielding a very strong association with the phenotype and a bigger pool of moderately associated SNPs. By running a greedy algorithm we will always be favouring a few strong associations up to the point that we might miss interesting moderate associations.

One of the approaches to building models, where each locus has fairer representation, is boosting. It is still a greedy model selection method, however the search path through loci is longer and more loci are considered on the path towards the solution. The main principle of this method is to combine many simple models to build a bigger and more accurate predictor. The approach is based on gradually adding small models to the composite model in such a way that each term is added only partially. In each step, the best of the small models is selected for adding. Before adding, the impact of the current best model is scaled by a small coefficient. Coefficients of the previous model terms are not recalculated. This approach can be compared to the importance sampling, because a new term gets a share in the big model that is proportional to the size of its effect upon phenotype. Conditioning on previous model terms characteristic for this approach can bring to the front more modest influences that explain the remaining variation in the outcome. It can be a useful property in elucidating QTLs.

Suppose that we are trying to fit a model $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot X_{ij} + \varepsilon_i$ by least squares, where the meaning of variables is the same as in equation (1.4). Then the forward stagewise regression is performed in the following way (Hastie et al., 2008):

```
1. Set $\alpha_k^* = 0, k = 1, \dots, p$.
2. for $m = 1$ to $M$ do
    (a) $(\beta^*, k^*) = argmin_{\beta_k,k} \sum_{i=1}^{N}(y_i - \sum_{l=1}^{p} \alpha_l X_{il} - \beta_k X_{ik})^2$
    (b) Update $\alpha_{k^*} = \alpha_{k^*} + \eta \cdot sign(\beta^*)$
3. return $\sum_{k=1}^{p} \alpha_k \cdot X_{ij}$
```

In each step the feature that explains the residual best is selected. So far the algorithm corresponds to stepwise regression. However, only a small portion of the total effect of this feature (characterized by η) is added to a model instead of refitting coefficients completely.

One of the most popular cases of boosting is AdaBoost method that is widely used for classification (Collins et al., 2002). Boosting is also done with trees serving as small models.

1.4.5.4 Probabilistic Model Selection

Probabilistic model selection is important for introducing more variation into the process of model building and its result. This can be crucial in genome wide studies if there are several loci with a similar effect on the phenotype. In such a case we would like to see them all represented in the final model, proportionally to their relative importance. This problem can be partially solved by boosting techniques described in the previous section, however the deterministic nature of stagewise algorithms can still be an obstacle towards obtaining an accurate picture of loci affecting phenotype.

A large class of probabilistic model selection methods exploit subsampling of individuals. It means that each step of model building is accomplished by using only a random subset of individuals and measurements. One of the simplest methods for probabilistic model selection is bagging, where simple models, each built from a different subsample, are just averaged. Bagging is sometimes used together with stepwise model selection to detect associations as in (Valdar et al., 2006, 2009). Here a forward selection is applied to various subsets of individuals and the resulting models combined by taking summary statistics, which are treated as the evidence for association. Also the random forest algorithm is based on averaging simple models (Breiman, 2001). In this case the role of simple models is played by decision trees, where each tree is built using a subsample of individuals and several subsamples of loci.

Random forests have been used in several association studies, where case/control data were analysed and SNP-SNP interactions were detected (Bureau et al., 2005; Lunetta et al., 2004). However, this method is used much more widely in the classification of gene expression profiles (Diaz-Uriarte and Alvarez de Andres, 2006; Huang et al., 2005; Shih, 2005). Another successful use of random forests is related to the prediction of protein-protein interactions from various features, including expression profiles, sequence data and GO categories (Qi et al., 2006).

Boosting methods can be converted into probabilistic algorithms by introducing subsampling of individuals. Under this approach the next small model is selected by using a subset of individuals. Then the model that performs the best on a given subsample is added to an aggregate model. An example of this method is the stochastic gradient boosting algorithm (Friedman, 2002).

An alternative approach to probabilistic model selection can be based on Monte Carlo simulations (Kirkpatrick, 1984). For example, simulated annealing can be performed over the search space of various loci (Heath, 1997). When using this approach, composite model gets slightly altered in each step by adding, removing or swapping loci. The choice of action is probabilistic and is based on Metropolis-Hastings algorithm. In such a way we can hope to find a global optimum of function (i.e. the best composite model) in the search space while avoiding local minima.

1.5 Association Studies for *Drosophila*

1.5.1 Traits and QTLs

QTL mapping in *Drosophila* has a rather long history, partly due to feasibility of producing various crosses and ability to maintain recombinant inbred lines. However, the number of phenotypes on which these studies have been focused is more limited. One of the first and most studied traits in the context of QTL mapping is abdominal and sternopleural bristle number (Gurganus et al., 1998; Long et al., 1998, 2000; Robin et al., 2002). Further studies have dealt with longevity (Geiger-Thornsberry and Mackay, 2004; Vieira et al., 2000), wing shape (Mezey et al., 2005; Pallson and Gibson, 2004), competitive fitness (Fry et al., 1998) and starvation stress resistance (Harbison et al., 2005; Nuzhdin et al., 2007).

The general experience of QTL mapping in *Drosophila* indicates that many QTLs are sex and environment specific. A typical example of a trait exhibiting genotype-environment interaction is longevity (Nuzhdin et al., 1997; Vieira et al., 2000). The study of longevity encompassing five different environments (varying temperature, heat shock and induced starvation stress) revealed that all statistically significant QTLs are sex or environment specific, while there are no important main effects of genotype outside the environmental context (Vieira et al., 2000). Even more, allele effects on longevity can be antagonistic for different sexes or different environments, including antagonistic pleiotropic effects for some environments. Thus a following hypothesis was made: mutations with beneficial effects on survival or fertility early in life can have deleterious effects later (Vieira et al., 2000). Another example of temperature and sex specific QTLs are the QTLs affecting abdominal and sternopleural bristle number in *Drosophila* (Gurganus et al., 1998; Mackay and Lyman, 1998).

Olfactory guidance behaviour is rooted in sex specific genetic architecture (Fanara et al., 2002). Initially one broader region of QTL that was not sex specific was found. Then complementation tests were performed and as a result the region was decomposed into two linked QTLs that are located nearby and have sex specific effect (regions did not overlap) (Fanara et al., 2002).

Many epistatic effects have been found for various traits of *Drosophila* as well. For example, bristle number, lifespan and wing shape are influenced by epistatic interactions of several loci. On the contrary, size and shape of posterior genital arch can be explained almost solely by additive QTLs (Liu et al., 1998). Wing shape QTLs exhibit many pleiotropic effects. Interestingly enough, some QTLs in epistatic interactions alter not only the magnitude of phenotype but also shift emphasis to a different trait, like in the case of various wing shape phenotypes (Mezey et al., 2005).

QTL mapping for wing shape is practically the only study attempting to dissect a composite trait by decomposing it into a set of well defined measurements. Different studies characterize wings by slightly different measurements: traits range from general dimensions of wings to subtler markers of vein intersections and relative positions (Mezey et al., 2005; Weber et al., 1999, 2001; Zimmerman et al., 2000). Probably the most accurate method is to define landmarks at the vein junctions and normalize measurements between landmarks relatively to a wing centroid (Pallson and Gibson, 2004). Then 18 relative warp measurements can be taken, which is sufficient to capture all possible vein measurement combinations.

Usually QTL mapping in *Drosophila* is done with RI strains. RI lines are used in studies of wing shape, longevity, sensory bristle number, olfactory behaviour and many other (Mezey et al., 2005; Nuzhdin et al., 1997; Zimmerman et al., 2000).

Several QTLs have been mapped for *Drosophila* using artificial selection for a trait. With this approach, a heat shock gene cluster (8 genes in total) and also a cluster of genes encoding insulin-like proteins were found to be responsible for stress resistance in *Drosophila* (Nuzhdin et al., 2007).

1.5.2 Towards Finer Mapping

There are not so many association studies for *Drosophila*, where associated loci would be mapped to a level of separate polymorphisms. One example of such a fine mapping is the study of *Egfr* region that was previously implied to be associated with wing shape characteristics (Pallson and Gibson, 2004). Many flies from nearly isogenic fly lines were sequenced for this region. Several associations were found, where most of them represented synonymous substitutions in the long exons.

In *Drosophila* fine mapping of QTLs, narrowing an associated region down to a single gene is usually done by complementation tests. According to this procedure, the allele of interest is crossed to a deficiency stock containing deficiency chromosome with null allele or alternatively a mutant allele and a balancer chromosome. Then the phenotype is assessed for 4 different genotypes of F_1 generation. If both alleles of interest yield considerably different phenotypes when combined with deficiency chromosome, then failure to complement is declared (comparison is relative to both parental phenotypes). This finding can imply either allelic effect of locus within parental strain or epistasis between parental allele and either deficiency or balancer chromosome.

Complementation tests have been successfully used to establish the relevance of QTLs to specific pathways, for example, relatedness of wing shape QTLs to Hedgehog and Decapentaplegic pathways (Mezey et al., 2005). In this case, a complementation test is done with alleles of the discovered QTL and alleles of pathway genes. If the QTL fails to complement most pathway members, then it is highly probable that the QTL belongs to the same pathway. Complementation tests are crucial in assessing whether there is some natural variation at the locus. It is possible to construct artificial allele mutations for some loci not exhibiting natural variation, therefore it is necessary to validate associations discovered by some experiments. 16 candidate genes potentially affecting longevity were studied and failure to complement was discovered in 4 cases confirming natural variation for these genes (Geiger-Thornsberry and Mackay, 2004). Other cases where deficiency complementation mapping has been used for fine mapping of QTLs include several studies of bristle number phenotypes (Gurganus et al., 1999; Long et al., 1996).

Complementation test requires the existence of a mutant allele stock, which can be serious obstacle for study of some genes. In addition, deficiency chromosomes can uncover around 100 genes and due to the broadness of the region uncovered, it is difficult to prove that QTL is really allelic to the candidate gene. Alternatives for fine mapping in those cases are LD mapping to narrow down the candidate gene list or using denser SNP maps.

P-element insertion is yet another method to study quantitative traits. Insertions have to be performed in a common isogenic background and for each type of insertion a trait has to be measured for a lot of individuals. Significant variation resulting from certain *P*-element insertions was observed in sensory bristle number and olfactory behaviour thus proving the role of specific candidate genes in the aetiology of the trait (Fanara et al., 2002; Norga et al., 2003). Novel genes potentially involved in neural development were discovered via study of the number of sensory bristles by *P*-element insertions (Norga et al., 2003). After performing many insertions those that were significantly changing bristle numbers were further scanned for insertion sites. One third of genes found had a characterized role in neural development, among them *Notch* and *Egfr* pathway genes (Norga et al., 2003).

P-element insertions can be used to test epistatic interactions as well. To achieve this goal, insertion lines have to be crossed in all possible combinations and then double heterozygotes have to be assessed for their phenotype. If double heterozygote phenotypes can be predicted from marginal effects of heterozygosity at each locus, then the effect is additive. In such a way an epistatically interacting network of several loci was discovered for olfactory behaviour (Fanara et al., 2002).

Many QTLs responsible for longevity are found through transgenic experiments (Aigaki et al., 2002; Tower, 2000). At least 19 QTLs involved in longevity were found by [Mackay, 2002]. This study was followed up by identification of plausible candidate genes within associated regions (Geiger-Thornsberry and Mackay, 2004).

With the availability of full resequencing of polymorphisms for *Drosophila* and detailed information about alleles, the issues of fine mapping of QTLs become less acute. Now it is possible to suggest association regions more precisely without extra experimental effort. The short LD exhibited by the *Drosophila* genome is very helpful here. However, the necessary correction for multiple testing for a huge amount of SNPs becomes more severe and conservative. Therefore having larger amount of genetically divergent strains phenotyped becomes an essential prerequisite for success in association studies.

1.6 Dorsal-Ventral Axis Specification during Oogenesis of *Drosophila melanogaster*

1.6.1 Stages of Oogenesis

The ovary of *Drosophila melanogaster* consists of 10 to 20 ovarioles held together by a peritoneal sheath. Individual egg chambers in different developmental stages are lined up linearly within ovariole. As the ovariole is traversed towards anterior end egg chambers exhibit younger and younger developmental stages.

There are three major cell types present in any egg chamber. The chamber is surrounded by a layer of follicle cells (see Fig.1.1). The most prominent component of an egg chamber is the oocyte with a rather large nucleus. There are 15 nurse cells located in the interior of a chamber, all of them connected to oocyte with ring canals. The complex of oocyte and nurse cells arises from mitosis with incomplete cytokinesis (Spradling, 1993).

Oogenesis is divided in 14 stages where late stages (from 10 onwards) are split into smaller units (Ashburner et al., 2005; King, 1970). The process of oogenesis can be characterized by several major changes in the architecture of egg chamber. At the beginning, the oocyte is the same size as any of the nurse cells. Both types of cells grow with similar rate until stage 8, when yolk formation begins and the oocyte expands to occupy a bigger and bigger portion of the egg chamber. By stage 9 it already takes one third of an egg chamber, but in stage 10 it corresponds

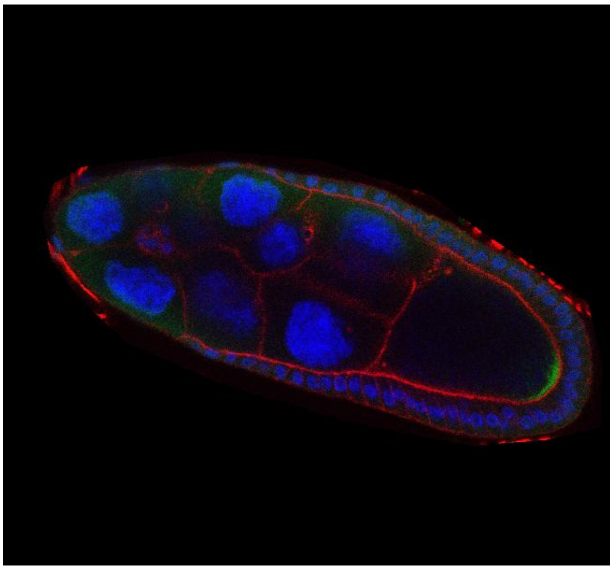


Figure 1.1: Microscope image of an egg chamber representing developmental stage 9.

to a half of the volume. Nurse cells gradually migrate towards the anterior end of the egg chamber as the development progresses. Approximately at stage 12, fully grown nurse cells shrink, nuclei condense to a spherical shape and their apoptosis takes place (Ashburner et al., 2005). Migration of nurse cells towards the anterior end of egg chamber is paralleled by migration of follicle cells: in stage 9, part of follicle cells migrate from posterior part of oocyte to anterior part of oocyte to begin centripetal migration at anterior end of oocyte afterwards.

1.6.2 Determination of Dorsal-Ventral Axis

The above mentioned migration of cells is a part of a bigger process that determines the formation of dorsal-ventral axis in the embryo. Here an important role is played by protein and mRNA localization (Bashirullah et al., 1998; St Johnston, 2005). Necessary proteins and mRNAs are usually synthesized in nurse cells, then transported into the oocyte through ring canals and at last localized with the help of large transport complexes (Wilhelm et al., 2000). The first stage of the journey of mRNA (the transport from nurse cells) is facilitated by dynein

(Bullock and Ish-Horowicz, 2001; Clark et al., 2007). The transport mechanism used for delivering mRNAs from nurse cells into the oocyte during early developmental stages is parallel to that used for apical localization of transcripts in blastoderm embryos (Bullock and Ish-Horowicz, 2001). Both types of localization depend on *Egl* and *BicD* function (they show similar distributions during oogenesis and embryogenesis) (Bullock and Ish-Horowicz, 2001; Clark et al., 2007). mRNA localization is followed by a local translation.

Fully functional mRNA transport is dependent on intact microtubule cytoskeleton as proved by treating and thus destroying microtubules by colchinine and colcemid (Clark et al., 1994; Zimyanin et al., 2008). During stage 7 microtubule organizing centre migrates from the posterior pole of the oocyte to the anterior end in such a way that microtubule plus ends become directed towards the posterior cortex (Brendza et al., 2000; Clark et al., 1997). It has been shown that localization of several important mRNAs, for example *oskar* and *Staufen*, is achieved through biased random walk towards the posterior end of oocyte. The movement bias itself is ensured by specific microtubule orientation (microtubule plus ends tend to assume narrower angle towards the posterior pole) (Zimyanin et al., 2008). Other prerequisites for successful localization, apart from working microtubule network, are functional follicle cell-to-oocyte signaling and correct follicle cell fate determination (anterior or posterior). It is interesting to note that Actin does not have a direct impact on localization of some mRNAs, like *oskar* and *bicoid* (Pokrywka and Stephenson, 1995; Zimyanin et al., 2007).

The main mRNA determinants of the anterior-posterior axis are *bicoid*, *oskar* and *gurken*. The *bicoid* mRNA has to be localized at the anterior end as it determines the anterior morphogen and thus the formation of head and thorax (Ephrussi and Lehmann, 1992; Lasko, 1999; St Johnston et al., 1989). In contrast *oskar* mRNA travels from the anterior end, where it gets exported from nurse cells, to the posterior pole. This process happens during stages 8-10 (Ephrussi et al., 1991; Kim-Ha et al., 1991). The importance of *oskar* lies in the fact that it determines the formation of primordial germ cells and abdomen in an embryo. The formation of the dorsal side of the embryo is dependent on correct Gurken signaling at the anterior part of the oocyte (Gonzalez-Reyes et al., 1995; Roth et al., 1995). *gurken*, *bicoid* and *oskar* mRNA translation has to be local to ensure normal course of development.

1.6.3 Phenotypes of Disrupted Axis Specification

There are several auxiliary proteins working in close collaboration with three main axis definers. Absence of them causes severe disruptions of developmental processes.

Mago Nashi (Mago) is crucial for establishing communication between follicle cells and the oocyte (Micklem et al., 1997; Newmark et al., 1997). Failure in the communication disrupts formation of dorsal-ventral and anterior-posterior axes as well as localization of *oskar* mRNA (Ephrussi and Lehmann, 1992).

Staufen protein colocalizes with *oskar* mRNA in the oocyte. Its role is to participate in transport of *oskar* mRNA towards the posterior pole (Ephrussi et al., 1991; Kim-Ha et al., 1991). Afterwards, Staufen activates *oskar* translation (Micklem et al., 2000). Another task of Staufen is to take part in *bicoid* mRNA localization and translation (Micklem et al., 2000). Mago has an additional role in anchoring the Staufen-*oskar* complex at its target location (Mohr et al., 2001).

The Tsunagi protein colocalizes with Mago and interprets the follicle cell-to-oocyte signal necessary to shape oocyte axes. Studies of *tsu* gene mutations imply that the Tsunagi protein is crucial for proper *oskar* localization at posterior pole, because Tsunagi helps in anchoring of mRNA (Mohr et al., 2001). Export, localization and quantity of *gurken* mRNA is affected by *tsu* mutations as well.

Mutations in Barentsz protein ensure that *oskar* is held at the anterior end of the oocyte due to incapacitated microtubule dependent transport to the posterior pole (van Eeden et al., 2001). Under normal conditions, Barentsz is supposed to be localized at the posterior pole at stage 9 together with *oskar*, as their localization is interdependent. Later, during stage 10, Barentsz disappears from the posterior pole after helping in *oskar* localization (van Eeden et al., 2001).

Mutations in heavy chain of kinesin I are known to disrupt *oskar* localization by interfering with active transport along microtubules (Brendza et al., 2000). Also Staufen migration to the posterior pole is inhibited in null mutants of heavy chain of kinesin I (Brendza et al., 2000). Despite being disruptive for transport along microtubules these mutations are not destructive for microtubules themselves.

Localization of Tropomyosin II is required for the head development due to participation in polarization of follicle cells (Erdélyi et al., 1995).

It has been demonstrated that splicing of *oskar* mRNA at a first exon-exon junction is crucial for *oskar* localization (Hachet and Ephrussi, 2004). The specific sequence at the first intron plays no role in the success of localization. It implies that there is a structural role for proteins like Barentsz, Y14, Mago Nashi heterodimer, eIF4AIII in the *oskar* localization (Hachet and Ephrussi, 2004). *oskar* is mislocalized to the anterior end if there is a point mutation in any of the genes producing proteins mentioned above (Hachet and Ephrussi, 2001; Mohr et al., 2001; Palacios et al., 2004).

1.7 Studying Heritability of Human Disease: the Case of Type 1 Diabetes

1.7.1 Type 1 Diabetes and Its Main Genetic Determinants

Type 1 diabetes (T1D) is a common autoimmune disease with complex aetiology. Autoimmune diseases are diseases where the immune system attacks the host rather than foreign pathogens, and are characterized by the defects in T cell selection and B cell selection (Gorodezky et al., 2006). T1D is distinguished from other autoimmune diseases by the lack of insulin-producing pancreatic β-cells. Depletion of β-cells is due to T cell mediated autoimmune destruction mechanism (Roep, 2003). Destruction is done via Th1 (CD4+) cells activating Tc (CD8+) cells that in turn attack β-cells (Gorodezky et al., 2006).

T1D has a tendency to co-occur among family members. For example, concordance in twins is 30-50% (Dorman and Bunker, 2000; Roep, 2003). This implies existence of genetic factors affecting disease and also the comparatively large amount of loci involved in T1D aetiology. There is also a multitude of environmental factors affecting the onset of T1D, including diet and viral infection

(Gorodezky et al., 2006).

The first genetic factors affecting T1D were found via candidate gene studies. Examples include HLA, INS, CTLA4, PTPN22 and IL2RA genes (Bell et al., 1984; Bottini et al., 2004; Lowe et al., 2007; Nistico et al., 1996). Chromosome 11p15 containing insulin locus INS was discovered to be associated with T1D by (Bell et al., 1984; Bennett et al., 1995; Vafiadis et al., 1997). PTPN22 gene on chromosome 1p13 was identified as causative factor for T1D by (Bottini et al., 2004; Smyth et al., 2004). CTLA4 gene on chromosome 2q31 was identified by (Anjos et al., 2004; Kristiansen et al., 2000; Ueda et al., 2003).

The biggest contributor to T1D heritability is the MHC region located on chromosome 6p21 (Barrett et al., 2009). It has been estimated that the MHC region comprises around 50% of genetically induced risk (Gorodezky et al., 2006). Various polymorphisms in the MHC region on chromosome 6p21 containing HLA genes were fine mapped by (Cucca et al., 2001b; Noble et al., 1996). Associated variants turned out to be clustered on genes HLA-DRB1, HLA-DQA1 and HLA-DQB1. Also HLA-DPB1 gene from class II was later found to be associated with T1D (Bugawan et al., 2002; Cucca et al., 2001a; Erlich et al., 1996; Noble et al., 2000). More detailed study of HLA class I alleles in family samples has led to fine mapping of 6 independent allele effects after correcting for strong linkage disequilibrium observed in the HLA region and discovery that half of them affect the age of onset for T1D (Valdes et al., 2005). Other alleles of HLA genes, like DR3/DR4 genotype, are connected with an early onset of T1D as well (Demaine et al., 1995; Fujisawa et al., 1995; Valdes et al., 1999).

A genome wide study incorporating 7 common diseases, among them autoimmune diseases like T1D, Crohn's disease (CD) and rheumatoid arthritis (RA) was done by the WTCCC (Wellcome Trust Case Control Consortium, 2007). Case/control sample consisting of approximately 2000 cases for each disease and 3000 shared controls was used in the study (Wellcome Trust Case Control Consortium, 2007). In total 24 association signals were discovered for various diseases including 7 associations with T1D. Five of six previously reported T1D loci were replicated, the exception being the INS gene. Two of the loci associated with T1D were discovered through multi-locus analysis exploiting haplotype information from HapMap and imputed alleles of non-genotyped SNPs. The following chromosome bands were spotted for associations with T1D: 1p13, 6p21 (MHC region), 12q13, 12q24, 16p13, 4q27 and 12p13. Combined analysis of RA and T1D yielded yet another locus at 10p15. P-values for chromosome 18p11 initially were considered to be only suggestive for association with T1D, however this region showed strong association with CD and suggestive association with RA, therefore it was included into the reported list of associations.

Replication study of regions reported in WTCCC study confirmed associations on chromosomes 12q13, 12q24, 16p13 and 18p11 (Todd et al., 2007). Studies of linkage disequilibrium within associated regions almost always led to confirmation of the causal variant reported previously.

Other genome wide studies have helped to elucidate genetic factors affecting T1D as well. Region 18q22 was first found to be associated with T1D by a genome wide association study of nonsynonymous SNPs (Todd et al., 2007). Especially strong association was shown for the CD226 gene. Another genome wide study discovered the IFIH1 region on chromosome 2q24 (Smyth et al., 2006). This association was immediately replicated in a family study of proposed candidate gene (Smyth et al., 2006). Further genome wide studies have discovered at least 20 additional risk factors for T1D, excluding those coming from MHC region and previously discovered by candidate gene studies (Concannon et al., 2008; Cooper et al., 2008; Fung et al., 2009; Hakonarson et al., 2007; Smyth et al., 2006; Todd et al., 2007; Wellcome Trust Case Control Consortium, 2007). Most of the loci found thus have moderate effect upon the outcome of disease.

Meta-analysis of three datasets (a sample from WTCCC study, a sample of T1D cases from the study of Genetics of Kidneys in Diabetes, a sample from the National Institute of Mental Health study (Baum et al., 2008; Cooper et al., 2008; Wellcome Trust Case Control Consortium, 2007)) highlighted 41 association region for T1D, half of them being novel findings (Barrett et al., 2009). A replication study was carried out for the proposed regions involving over 4000 cases, the same amount of controls and over 4000 trios from families with multiple affected offspring. 18 regions were found significant after the replication study, including chromosome 1q32.1 encompassing immunoregulatory cytokine genes IL10, IL19 and IL20. Interactions with MHC region were tested in the same study. Five potential interactions were found, where in 4 cases the risk associated with interacting allele is reduced in the presence of the MHC risk allele (Barrett et al., 2009).

Yet another meta-analysis using the same three datasets was carried out by (Cooper et al., 2008). Associations with 10 previously known loci were confirmed, while no new loci emerged from the study apart from association evidence for 4q27 region encompassing genes IL2-IL21 (Cooper et al., 2008). After genotyping independent set of cases and controls (over 6000 individuals each) and almost 3000 families for top scoring SNPs of the meta-analysis, 4 new loci were discovered on chromosomes 6q15, 10p15, 15q24 and 22q13 with candidate genes BACH2, PRKCQ, CTSH and C1QTNF6.

1.7.2 Fine Mapping of Associated Loci

Studies of linkage disequilibrium and the building of composite models incorporating several loci are very important to fine map variants within association region. The study of linkage disequilibrium is important to distinguish whether the reported SNP is really causative for a disease or if it is just brought to the forefront by a linked variant. Composite models are indispensable to discern independent variants within region by testing whether newly discovered variant can add anything important to the explanation of a disease by known SNPs. For example, linkage disequilibrium studies indicated that ITPR3 gene polymorphisms are actually linked to the MHC region and HLA-DQB1 gene (Qu and Polychronakos, 2007). Besides, ITPR3 gene variants did not contribute significantly to a composite model accounting for the effect of variants within HLA-DQB1.

The significance of IL2RA locus was first indicated by (Qu et al., 2007b; Vella et al., 2005). First fine mapping leading to discovery of two different groups of causal polymorphisms within IL2RA gene was done by (Lowe et al., 2007). Stepwise logistic regression was used to test whether there is another SNP in the region that could add to the T1D risk explanation (Lowe et al., 2007). Two SNPs (ss52580101 and rs11597367) were found to be sufficient for the explanation of T1D risk associated with the region under study. Further fine mapping of variants within the IL2RA gene identified two independent effects, one being a 5

SNP haplotype and another being a single SNP unrelated to the haplotype (Qu et al., 2009). The discovered haplotype turned out to be in strong LD with a previously reported SNP, while the SNP represented a novel association.

1.7.3 Associations Shared by Type 1 Diabetes and Other Diseases

It has been studied whether variants associated with T1D are pleiotropic for several autoimmune diseases. For example, in depth study of variants in 6p21 region has been carried out to determine whether there are shared alleles that increase risk for both T1D and multiple sclerosis (MS) (Alcina et al., 2009). Also the IL2RA gene was searched for overlapping causative variants for T1D and MS after accumulating the evidence for IL2RA association with T1D (Lowe et al., 2007; Qu et al., 2007b; Vella et al., 2005) as well as with multiple sclerosis (Alcina et al., 2009). It was found that causative variants are different for both diseases despite them being located in close proximity. IL2RA is also connected with Graves' disease (Brand et al., 2007).

Systemic lupus erythematosus and T1D share at least 3 association regions: PTPN22, HLA and CTLA4 (Qu et al., 2007a). IRF5 gene variants have been tested for associations with both diseases as well. Two IRF5 gene variants are associated with systemic lupus erythematosus, one variant being located downstream from IRF5 and affecting the expression of this gene and another variant at intron 1 being responsible for alternative splicing (Graham et al., 2006; Sigurdsson et al., 2005). Both variants were tested for association with T1D in a family based study, but no significant risk was found (Qu et al., 2007a).

Another study of potential pleiotropic mechanisms for T1D and other diseases has been carried out for the PTPN22 gene on chromosome 1p13 (Bottini et al., 2006). Association of PTPN22 with T1D disease susceptibility was first found by (Bottini et al., 2004; Ladner et al., 2005). Associations with the same gene have been found for rheumatoid arthritis (RA) (Begovich et al., 2004; van Oene et al., 2005), Graves' disease (Smyth et al., 2004; Velaga et al., 2004), systemic lupus erythematosus (Kyogoku et al., 2004; Wu et al., 2005), generalized vitiligo (Canton et al., 2005) and other autoimmune diseases. A pathway involving several polymorphisms in PTPN22 gene and explaining disease associations was proposed by (Bottini et al., 2006).

A shared polymorphism increasing risk for several autoimmune diseases has been found within CD226 gene on chromosome 18q22 (Hafler et al., 2009). Ser^{307} allele of rs763361 is associated with T1D, MS and to a weaker extent with RA and autoimmune thyroid disease. All of these diseases are T cell mediated, where T cells target different organs and tissues in each case. Forward logistic regression was used to test whether there are other variants within the region apart from Gly307Ser that can contribute to explanation of disease risk (answer was negative).

Polymorphisms within gene CTLA4 on chromosome 2q31 are responsible both for increased T1D risk (Marron et al., 2000; Ueda et al., 2003) and Graves' disease risk (Heward et al., 1999). One common polymorphism affecting both diseases is reported in (Ueda et al., 2003).

1.8 Aims of Work

Early development in fruit fly is practically neglected in association study literature. However, it has been studied widely by qualitative instead of quantitative ways thus highlighting many important pathways of mRNA and protein localization. Severe mutations usually get the most of attention, like mislocalization patterns leading to a failure to hatch for an embryo. I attempt to characterize developmental patterns during oogenesis as continuously varying traits. Phenotypes are scored from microscope images highlighting the crucial components of oocyte. Such an approach should be more precise in identification of subtle variation in developmental patterns.

Oogenesis phenotypes are further subjected to genome wide association study. This study is among the first ones to exploit a map of roughly 3 million different loci for *Drosophila melanogaster*. Such a density allows to gain all the benefits from using nearly isogenic lines. Besides, it is one of the first systematic scans for loci affecting developmental patterns. The aim is to find good candidate SNPs for associations that could be subjected to further biological verification.

Many reviews of association studies admit how difficult it is to chase the loci with modest effects on the phenotype. The idea behind the reanalysing of human case/control data from WTCCC is to test whether a specific class of model selection methods can bring the SNPs of modest impact from background to foreground.

Firstly, stepwise model selection methods are applied to the WTCCC dataset. Conditional scans of genome have the potential of increasing the power of association discovery, because modest associations are enhanced after taking into account the biggest associations. Afterwards I turn to the probabilistic model selection methods. Probabilistic approaches should produce rather high variation among selected entities and therefore have the potential of highlighting new loci of modest effect upon phenotype.

At the end of these studies it should be possible to decide whether the model selection approaches pursued are suitable for extracting more information (associations) from case/control data. Another aim of the study is to compile a list of new candidate SNPs. However, the first question is, how closely previously reported associations can be replicated when using various model selection methods.

The thesis is organized as follows. The second chapter specifies experimental techniques used in collection of *Drosophila melanogaster* phenotypes and also defines image analysis methods necessary for phenotype quantification. The third chapter describes models used in association discovery for fruit fly. Fourth chapter covers the results of fruit fly association study. Modeling techniques specific to analysis of human case/control data are explained in chapter 5. At last, results of association search in human are described in chapter 6.

Chapter 2

Phenotyping Oogenesis of *Drosophila melanogaster*

2.1 Describing Oogenesis Phenotypes

In literature one can find a set of well described phenotypes resulting from mutations in alleles of genes affecting the course of development of *Drosophila melanogaster*. Typically these genes transcribe mRNAs whose correct localization in oocyte is crucial for establishing the correct pattern of expression and sometimes also the correct signalling between various compartments. However, the phenotypes are detected almost exclusively in a qualitative way. This approach to phenotyping is suitable for the severeness of developmental defects observed (like failure of embryo to hatch) or obviousness of mislocalization of crucial compounds (mRNA scattered over the oocyte instead of being expressed only at posterior pole). A good example is the description of the phenotype corresponding to a mutation in *stau* allele: *oskar* is temporarily localized at the posterior pole during stage 9, but it does not get anchored and disappears before stage 10; meanwhile since stage 8 *oskar* exhibits high concentration at the anterior boundary; the mutant allele later results in defective abdominal patterning (Kim-Ha et al., 1991). Subtler, quantitative and continuous differences in oogenesis phenotypes have been studied to a much smaller extent. However, some of the discovered roles of specific proteins and mRNAs in dorsal-ventral axis specification naturally lead to formulations in terms of continuous traits. For example, it is known that the increase in *oskar* expression causes proportionally increased amount of Barentsz to locate at the posterior pole, which in turn implies significance of Barentsz in *oskar* transport (van Eeden et al., 2001). Here it is possible to imagine the presence of some *oskar* gene allele or the expression level of *oskar* being explanatory variables in the model describing the amount of Barentsz at the posterior pole.

Another aspect that has been missing from qualitative descriptions of phenotypes is the integration of general characteristics of the egg chamber into trait definition. Examples of such characteristics include shape, elongation and volume of oocyte or volume and amount of a certain protein in nurse cells. Nurse cells can serve as very good indicators of developmental course, because their volume and nuclei change considerably during various stages. For example, nurse cell nuclei have polyploidization level of 256-512 copies at stage 8. It reaches 512-1024 copies at stage 10 before cell contents gradually pour into oocyte during stage 11 (Ashburner et al., 2005). In addition many important proteins and mRNAs are synthesized in nurse cells and only then transported into the oocyte (Wilhelm et al., 2000). Therefore nurse cell characteristics can be viewed as crucial precursors or indicators of important oogenesis functions.

Shifting focus from studying specific pathway by phenotyping localization of few mRNAs or proteins to general developmental traits should facilitate finding loci responsible for the developmental process via a genome wide association study. Firstly, defining traits in a more generic fashion can save a lot of experimental effort that would be necessary to perform antibody staining and imaging of many different mRNAs. Secondly, it would be possible to easily identify loci having pleiotropic effect on various phenotypes. Thirdly, there would be no need to limit the focus to studying particular candidate genes via mutagenesis, but instead an unbiased genome wide search for interesting loci would be feasible. Such a systematic search for associations has now become possible due to the availability of fully resequenced lines of *Drosophila melanogaster* (Drosophila population genomics project, 2010).

I collected general traits of an egg chamber, such as the size of nurse cell nuclei, shape descriptors of nurse cell nuclei and relative localization of nuclei. The size of the nurse cell nuclei is almost equivalent to the size of the nurse cells.

Nuclei are measured instead of the cells themselves, because it is much easier to carry out a staining that highlights nuclei instead of a whole cell. These traits should provide broader information about the course of development and subtle differences created by various genetic factors.

I propose to treat oogenesis phenotypes from the continuous prospective and try to establish the association between continuous variation in developmental characteristics and genomic loci. Now the question arises about the right method of phenotypic measurement that could discern mutations to a high enough level of detail. Optical microscopy was proposed as a tool for a systematic screening of high-dimensional phenotypes. The trait measurements themselves were the products of image analysis performed on pictures of egg chambers, where several biological features indicating the developmental course are marked by staining. Traits showing more variation between strains than within a strain are good candidates for further genome wide study.

After collecting of phenotypes a genome wide association study linking developmental traits and SNPs from 37 lines of *Drosophila melanogaster* was carried out. Lines obtained from Bloomington Drosophila Stock Center and donated by Trudy Mackay were used to study the relationship between genotype and phenotype.

2.2 Experimental Procedures

2.2.1 *Drosophila melanogaster* Stocks

37 fly lines from Bloomington Drosophila Stock Center deposited by Trudy Mackay were used: RAL-208, RAL-301, RAL-303, RAL-304, RAL-307, RAL-313, RAL-315, RAL-324, RAL-335, RAL-357, RAL-358, RAL-360, RAL-362, RAL-365, RAL-375, RAL-379, RAL-380, RAL-391, RAL-399, RAL-427, RAL-437, RAL-486, RAL-514, RAL-517, RAL-555, RAL-639, RAL-705, RAL-707, RAL-714, RAL-730, RAL-732, RAL-765, RAL-774, RAL-786, RAL-799, RAL-820, RAL-852. *Drosophila melanogaster* ancestors of these lines were sampled from Raleigh, North Carolina. All lines are isogenic. Homozygosity was confirmed for the majority of the genome in each strain (Drosophila population genomics project,

2010). Line RAL-208 was analysed and phenotyped, but later discarded, because the genotypic information did not meet quality requirements. Two lines (RAL-514, RAL-730) were not successfully phenotyped as the number of flies was too small and/or samples from the ovaries could not be collected. Thus, the total of 34 fly lines were fully analysed.

2.2.2 Antibody Staining of Ovaries

In total 7 female flies were dissected from each line and samples of ovaries taken. Ovaries were dissected in 1x PBS, then fixed in PBS by adding 4% formaldehyde for 20 minutes. Afterwards ovaries were blocked in PBT[1] (0.3% Triton) with 0.5% Bovine Serum Albumine and left for 1 hour. Then incubation with primary antibody (The Crude Rabbit anti Oskar) at room temperature in the blocking buffer took place. The Crude Rabbit anti Oskar antisera were diluted 1:3000 in blocking buffer. Ovaries were washed 2 times with PBT (0.3% Triton) and 0.5% Bovine Serum Albumine and then blocked in PBT (0.1% Triton) with 10% Normal Goat Serum for 2 hours. Afterwards the solution was incubated with secondary antibody (Goat anti Rabbit FITC) and phalloidin for 2 hours in PBT (0.1% Triton). Goat anti Rabbit FITC was diluted 1:500, but Rhoda mine conjugated phalloidin was diluted 1:200. Lastly ovaries were incubated with DAPI diluted at 1:2500 in 1x PBS for 5 minutes. Then ovaries were put into 100μl of mounting medium (2% N-propylgallate, 80% glycerol).

2.2.3 Collecting Images

Imaging was performed with the Leica confocal microscope in UV confocal scanning mode with 40x lens and 1.25 oil. At the beginning of work a profile of imaging parameters was set up and then used throughout the whole batch of samples. This profile defined image channels and their corresponding frequencies and intensities. DAPI was captured with blue channel, Actin channel was red and *oskar* fluorescence went to a green channel. Each picture was obtained as an average of 4 frames taken at regular depth intervals. The initial image depth was manually set at a plane of maximum *oskar* expression.

[1] Acronym PBT denotes 1x PBS with 0.1% TritonX

2.3 Extraction of Traits via Image Analysis

2.3.1 Main Steps in Feature Extraction

In Fig.2.1 we can see examples of collected egg chamber images representing developmental stage 9 and stage 10 respectively. The dark compartment is the oocyte with *oskar* mRNA localized at its posterior end (crescent in green colour). The blue layer around oocyte represents the follicle cells. The large blue objects at the anterior end of egg chamber are nurse cell nuclei. Red boundaries of egg chamber, oocyte, nurse cells and follicle cell layer come from Actin staining. The egg chamber in stage 10 is considerably bigger than the other in stage 9. Also, the prominence of the oocyte is increased from approximately 1/3 of egg chamber to 1/2 from stage 9 to stage 10. In stage 9 a follicle cell layer surrounds the oocyte as well as part of the nurse cell area, but later they migrate to fully encircle the oocyte (boundary between oocyte and nurse cells already begins to form in stage 10 as can be seen from corners). Nurse cells in stage 10 are slightly bigger and the staining of their nuclei is less bright in comparison with stage 9. *oskar* concentration at the posterior pole is the highest in stage 9, later it spreads out and the crescent becomes thinner. Of all the aspects of egg chamber captured I focused my attention mostly on nurse cells, disregarding the question of localization of specific proteins throughout this study. From a technical point of view nurse cells are a relatively easy target for a semi-automated image processing pipeline, therefore suitable for demonstrating the approach of feature extraction from microscope images.

The path from a microscope image to traits is outlined in Fig.2.2. While on the left side major image processing activities are depicted, the right side shows their outcomes. The process begins with a manual selection of the egg chamber from microscope image. This step is necessary because sometimes there are two or more egg chambers or even some eggs captured in a single picture. For example, in Fig.2.2 there is an egg chamber surrounded by two mature eggs. Then the blue channel is taken from egg chamber image. A single channel is sufficient for getting nurse cell nuclei descriptors as there is not much extra information about cells that could be obtained from *oskar* expression or Actin patterns. The

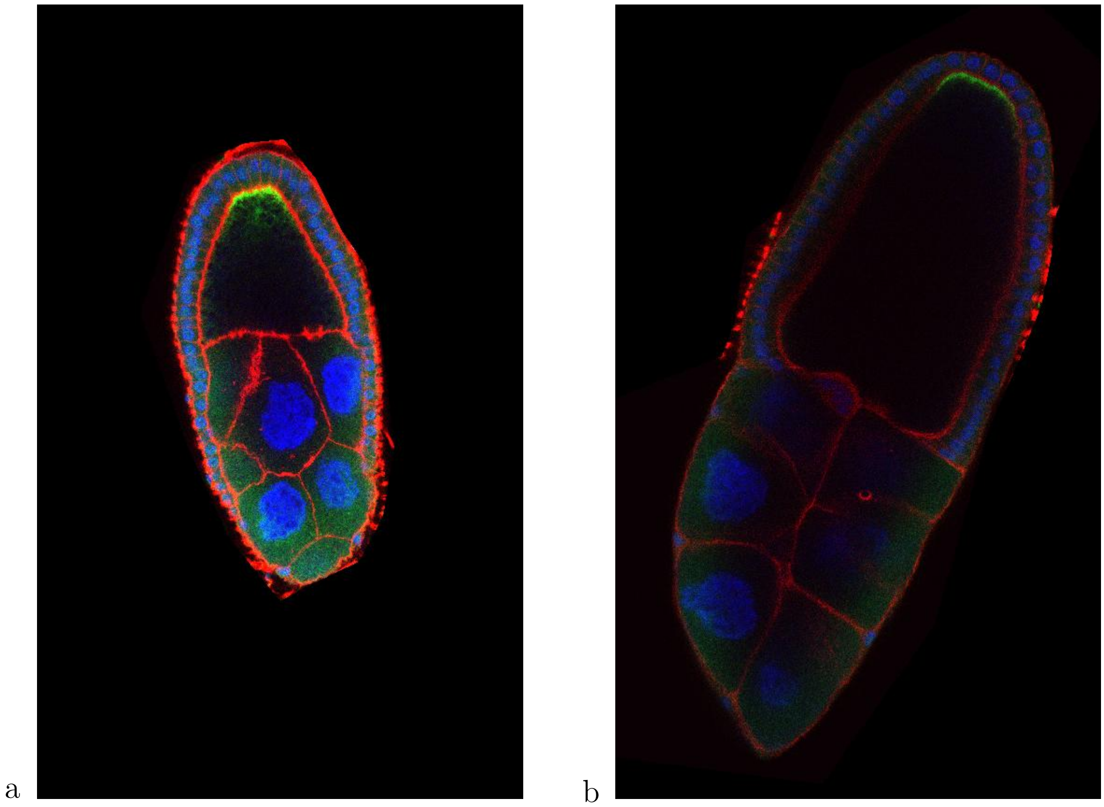


Figure 2.1: Microscope images of egg chambers: a. stage 9, b. stage 10.

blue channel is further manually segmented into nurse cells and follicle cells. This step facilitates identification of nurse cells, because follicle cells and nurse cells are very different in size and their identification requires completely different tuning of parameters. Another problem avoided by cell type separation is occasional merging of the follicle cell layer with nurse cells which are close to the anterior end of oocyte. At last automated identification of nurse cells can take place. Once the cells are detected original nurse cell image together with cell nuclei masks are used for trait scoring.

The next section is devoted to a more detailed description of image analysis necessary to automatically detect nurse cells, and is followed by more precise definitions of extracted traits.

2.3.2 Automated Cell Detection Pipeline

Automated detection of cells cannot be imagined without certain image analysis operations. The aim is to distinguish cells from an image background and to detect contours of their nuclei. The contour detection typically ends with obtaining cell nuclei masks. A mask is black and white image where white pixels correspond to the nuclei and black pixels indicate background (see Fig.2.6). Therefore the central component of the analysis pipeline is the binarization of the original image with an aim to classify regions into cells and background. The biggest problem in accomplishing this task is avoiding small spurious features coming from a noisy background to be classified as cells. As it can be seen from Fig.2.1 the background noise in egg chamber images is not very noticeable, but still certain precautions have to be taken. Typically original image would be processed by a flat filter before binarization begins with an aim to reduce non-homogeneity of texture and smooth all features. After binarization takes place noise can be reduced with morphological operations (O'Gorman et al., 2008). For example, if a pixel that is classified as a foreground (cell nucleus) is surrounded almost solely by background pixels, it is highly unlikely that it truly represents a cell nucleus. In such a case foreground pixel can be converted into background. Background pixels can be checked for their foreground neighbours in similar way. These background/foreground conversion operations are helpful in getting rid of small

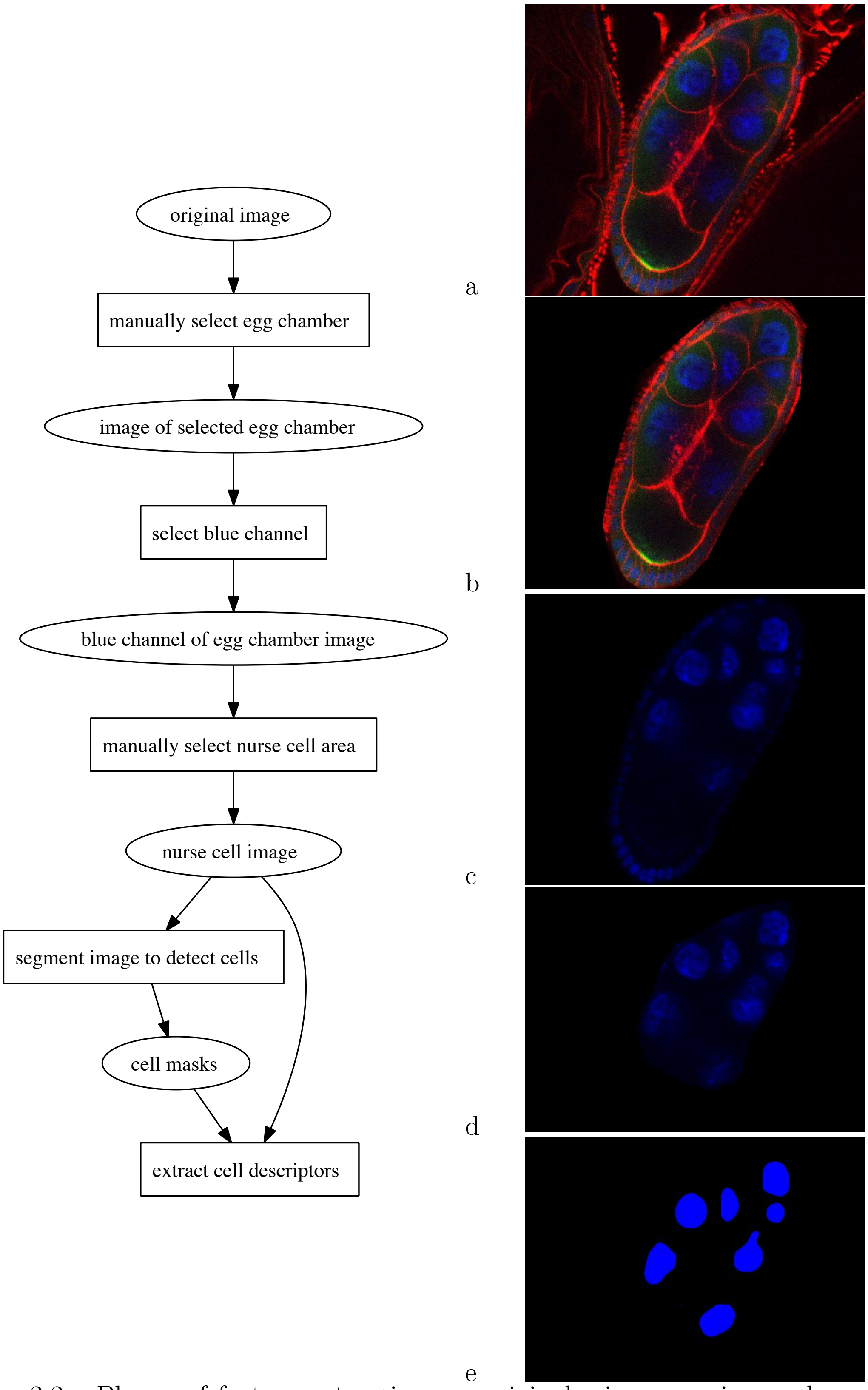

original image
manually select egg chamber
image of selected egg chamber
select blue channel
blue channel of egg chamber image
manually select nurse cell area
nurse cell image
segment image to detect cells
cell masks
extract cell descriptors
a
b
c
d
e

Figure 2.2: Phases of feature extraction: a. original microscope image, b. selected egg chamber, c. blue channel of egg chamber image, d. selected nurse cells, e. masks of cell nuclei.

spurious features, filling small gaps and also in smoothing of nuclei boundaries. One of the methods that combines both types of conversions is image opening. It starts with a round of eroding boundaries (foreground to background conversion) and then proceeds to background to foreground conversion. After smaller features were removed there were still some larger objects present that are not cells. To remove them I could use the knowledge of a typical size of a nurse cell and erase everything that is deviating too much from this size. A way how these operations were combined into a cell detection pipeline is depicted in Fig.2.3.

For some steps in image analysis pipeline there are several possible approaches. For example, the binarization step can be performed both with global thresholding algorithms and local thresholding methods. A global thresholding approach selects a single signal intensity level from intensity histogram and proclaims that everything below this level is background, while everything above is a foreground. A local thresholding method looks at a smaller neighbourhood and classify a pixel as a foreground if it is either surrounded by many similar foreground pixels or it is very different from its background neighbourhood. As the background of nurse cell images is very homogeneous and does not contain much noise, global thresholding could potentially work here. After experimenting with Moments method and other algorithms provided by plugin of Gabriel Landini (Landini, 2010) (algorithms not assuming bimodal intensity histogram were tested), it was discovered that global approaches are not favourable for stage 10 egg chambers. As the nurse cell nuclei staining is more dispersed in stage 10 and signals are weaker global thresholding has a tendency to identify only irregular parts of late stage cells or dismiss them altogether (Fig.2.4). Local thresholding does not have this problem as it is much more sensitive to the intensity changes. However, some spurious features can arise near the edges of the nurse cell segment, because there is a noticeable division between homogenous egg chamber background and black background. Fortunately, they can be easily recognized by their shape and compliance with the segment border and thus manually removed (see part d. of Fig.2.6).

After experimenting with several local thresholding methods the Bernsen algorithm was chosen as it produced the most accurate boundaries of cell nuclei and did not adjoin background regions with cell regions (see nurse cell segmentation

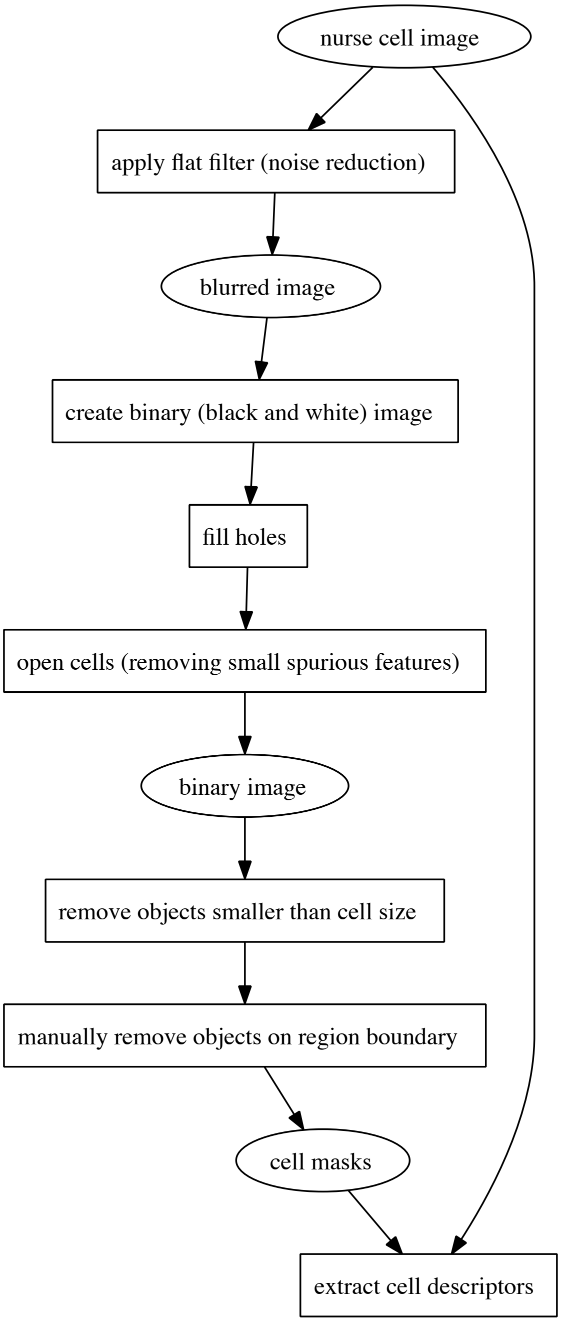

nurse cell image
apply flat filter (noise reduction)
blurred image
create binary (black and white) image
fill holes
open cells (removing small spurious features)
binary image
remove objects smaller than cell size
manually remove objects on region boundary
cell masks
extract cell descriptors

Figure 2.3: Image analysis pipeline for nurse cell nuclei identification.

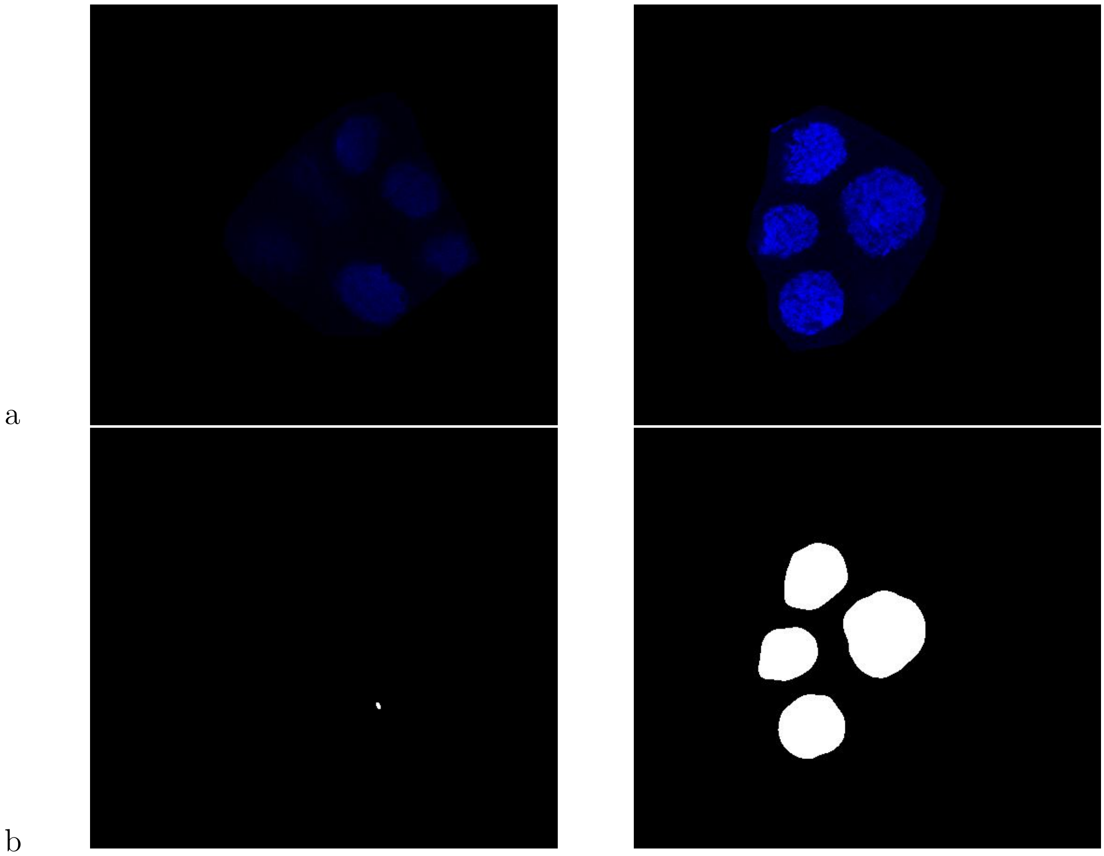


Figure 2.4: Moments method of global thresholding applied to cells of different stages: a. original nurse cell image, b. binarized image. Global thresholding methods treat cells of different stages and brightness in different way. Cells from the left image are practically lost after thresholding, while size of the cells from the right image is slightly exaggerated.

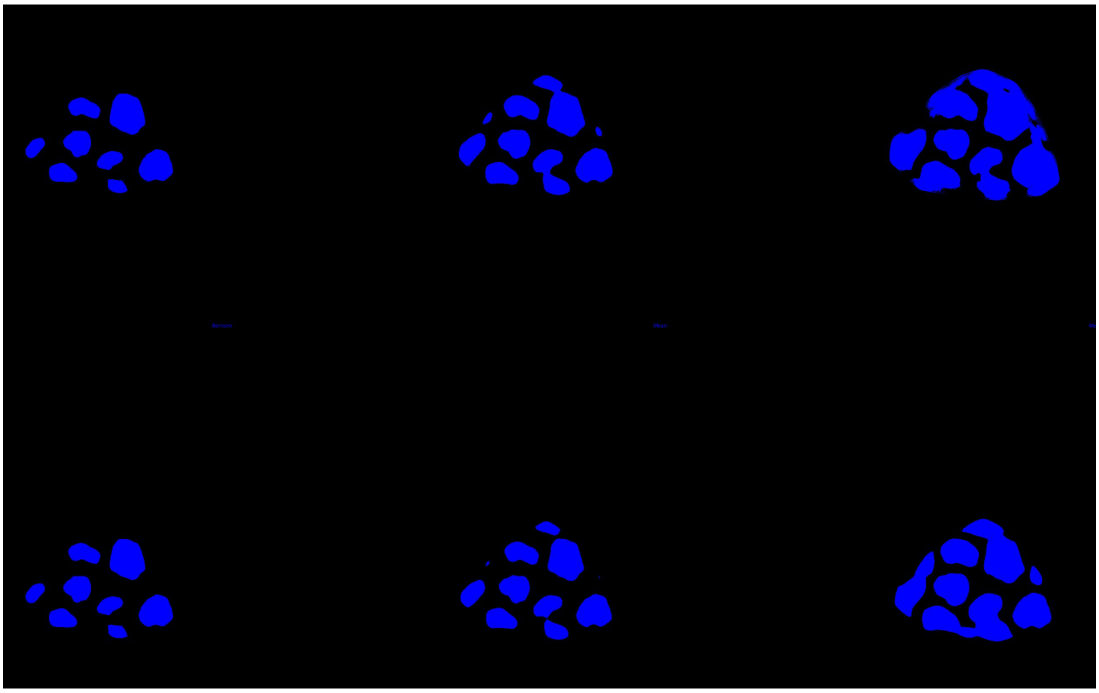


Figure 2.5: Comparison of six methods of local thresholding applied to the nurse cell image in Fig.2.6. Methods from left to right: Bernsen, Mean, Median, MidGrey, Niblack, Sauvola. In all cases mask with a radius 45 was used.

in 2.5). The use of medium gray level in the region as threshold (as specified by Bernsen method) is suitable in nurse cell case, because signal intensity within one cell nucleus has small variation. Thus, the whole nucleus can be easily distinguished from the average greyness of the region.

Tuning of parameters for various steps within pipeline led to following decisions. For initial noise removal Gaussian filter was used with $\sigma = 5.5$. Bernsen algorithm was applied with circular neighbourhood of radius 45 pixels and intensity difference 15 levels of gray scale. Then cellular opening (erosion followed by a dilation) was performed twice with threshold of 4 foreground pixels in neighbourhood and no requirements for connectivity. Objects of area less than 750 pixels were removed. All of these steps were performed by ImageJ tool, where local thresholding was done with the plugin created by Gabriel Landini (Landini, 2010).

The full course of image analysis and cell detection is illustrated by examples

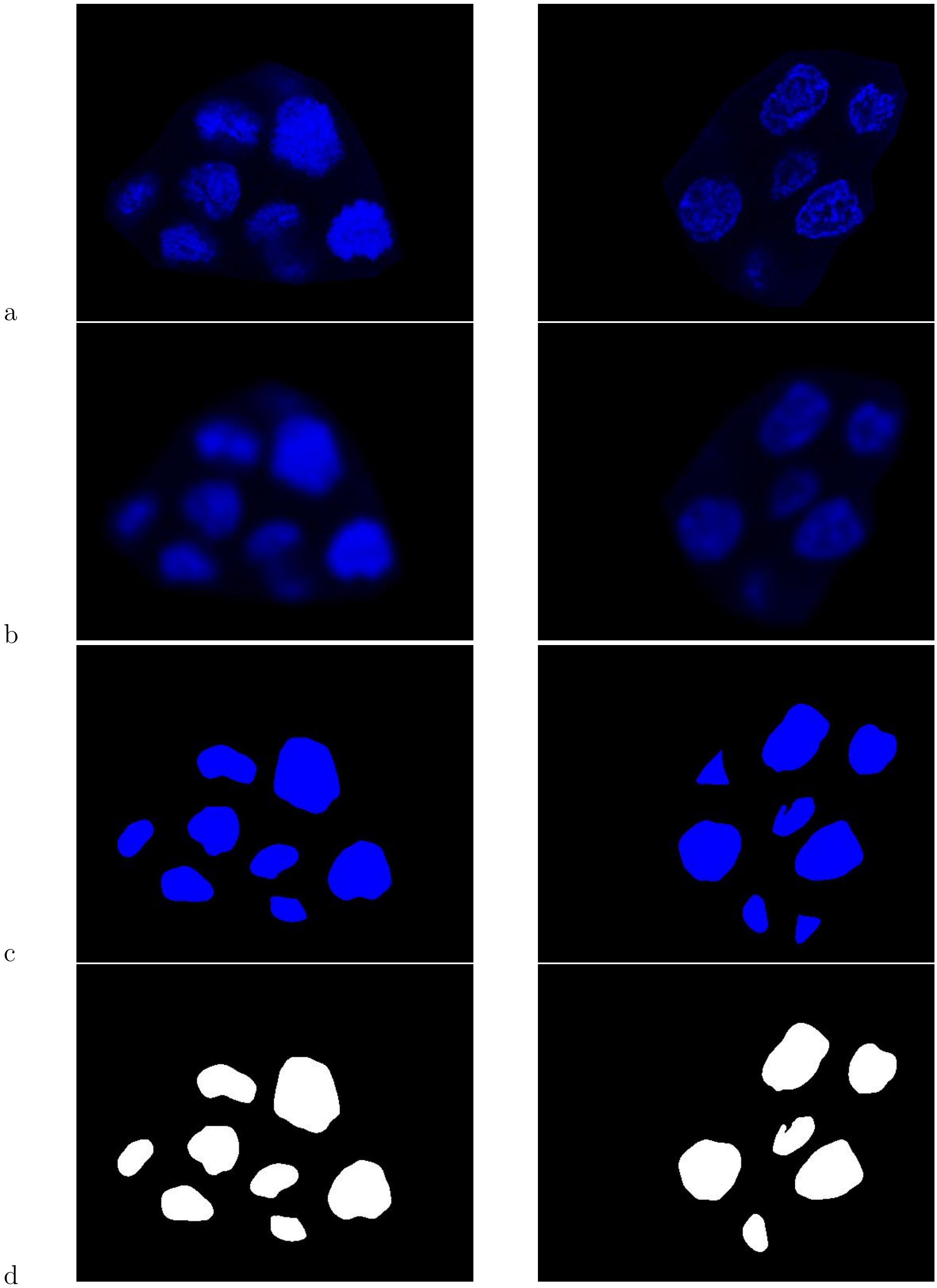


Figure 2.6: Phases of segmenting a blue nurse cell channel. There stages depicted are: a. original nurse cell image, b. blurred image (Gaussian filter, $\sigma = 5.5$), c. binary image, d. masks of cell nuclei; in the right example characteristic outliers at the segment boundary are removed.

in Fig.2.6. Screening of all processed images assured that the masks are correct apart from very few spurious objects that were manually removed. The images and their binary masks were later superimposed to obtain various characteristics of objects, like their size, curvature or moments.

2.3.3 Feature Extraction

After initial segmenting of images I extracted various descriptors of cell nuclei and egg chamber regions. These later served as traits in the study of association with genetic loci. 92 image features were collected in total, which can be divided into 3 categories: geometrical shape descriptors, moments and texture descriptors. Geometrical features provided basic characterizations of cell nuclei, including volume, perimeter and compactness. These are probably the most obvious things to measure and the easiest for subsequent interpretation. Moments characterized the distribution of signal intensity within object. The understanding of moments in image analysis is similar to how we would understand moments in statistics. If in statistics the first order moment corresponds to the mean of a variable, then in image analysis the first order moment is either area or total signal intensity (for a binary and non-binary image correspondingly). If in statistics moments characterize the underlying distribution from which random variables are drawn, then the image moments can describe signal distribution in various ways. In image analysis it is almost always a two dimensional case of moments, dimensions corresponding to x and y axis. Texture descriptors captured the relationship between pixels in a narrow neighbourhood. They provided information about local variations in signal intensity in contrast to global variations that are described by moments. For example, homogeneity of DNA distribution within cell nucleus can be studied thus. Another useful texture characterizer was the entropy or the measure of order of DNA within nucleus. I will proceed with precise definitions of features within each class. EBImage package for R was used for extraction of features from microscope images (Pau et al., 2010).

2.3.3.1 Geometrical Shape Descriptors

First I extracted simple geometrical descriptors of object shape. These were: coordinates of geometric center, area, perimeter, mean distance from center to perimeter (together with corresponding variation and largest difference between distances), effective radius (radius of a circle with the same area), acircularity (measure of deviation from circular shape) and compactness. I also studied the Fourier transform of a distance profile and reported 4 first frequency components. The distance profile shows how the distance from center to perimeter changes when the perimeter is scanned in clockwise direction. The higher is the frequency the smaller is interval in radians after which distance from center to perimeter is scanned. Fourier transform components are more detailed counterparts of acircularity and compactness.

2.3.3.2 Moments

One of the simplest and most generalizable approaches to describing image is moment analysis. The moment m_{pq} of order pq is defined as:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) \, \mathrm{d}x \, \mathrm{d}y \tag{2.1}$$

Here x and y denote the location of a pixel and $f(x, y)$ is the signal intensity. The moment m_{00} corresponds to the area of the object, if applied to a binary image, and to the total signal intensity, if applied to the grayscale image (O'Gorman et al., 2008). I was not particularly interested in other moments apart from m_{00}, because they are not translation invariant. Not being translation invariant means that for higher order moments the location of a cell within image is equally important as other characteristics, like volume or elongation. However, moments of higher order could potentially have a use in characterizing relative localization patterns, like in the case of *oskar*.

The independence of moment from location is frequently achieved by using central moments as image descriptors. The central moment μ_{pq} is defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - x_c)^p (y - y_c)^q f(x, y) \, \mathrm{d}x \, \mathrm{d}y \tag{2.2}$$

where $x_c = m_{10}/m_{00}$ (normalized x coordinate of object centroid) and $y_c = m_{01}/m_{00}$ (normalized y coordinate of object centroid). These moments are used in calculation of major and minor axis, orientation of an object and its eccentricity. Central moments are not rotation invariant, therefore not very informative about pictures of egg chambers apart from measuring minor and major axes and eccentricity (elongation). If all egg chamber images were rotated towards their anterior-posterior axis, central moments would be more useful in describing relative localization patterns. However, their use in characterizing nurse cells would still be limited, because the orientation of cells relatively to anterior-posterior axis exhibit a lot of variance.

If a central moment is normalized by an area of object, one can obtain a scale invariant moment. More precise definition of a scale invariant moment is captured in the following equation:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+\frac{i+j}{2}}} \tag{2.3}$$

An example of scale invariant moment is the moment of inertia: $I = \eta_{20} + \eta_{02}$. In turn, η_{11} represents deviation from a circular shape, while η_{22} is the normalized second order moment of the same quantity. Unfortunately, these moments suffer from the lack of rotational invariance, therefore cannot be used in describing nurse cells. Nevertheless they are important in deriving rotational invariants.

There are several ways how rotation invariant image descriptors can be constructed. I used two different classes of rotation invariant moments for describing nurse cell nuclei, namely Hu's invariants (O'Gorman et al., 2008) and Zernike moments (Khotanzad and Hong, 1990; Kim and Kim, 2000). Both types of descriptors have the advantage of being capturers of shape features that are not dependent on cell nuclei size or volume. Thus, these moments can potentially elucidate the properties of cells that are not directly tied to their developmental stage. They rather describe the distribution of DNA within nucleus and characterize its shape.

Hu's invariants are derived from scale invariant moments by combining them and performing several normalizations as described in (O'Gorman et al., 2008). Each of these invariants is calculated in a different way and their generating functions do not comply with a single form of equation as it was in case of moments (therefore more detailed description is omitted here). I calculated all 7 Hu's invariants for original images and 2 first for binarized images.

Zernike moments are obtained under a rather different paradigm: each moment is a projection of an image on Zernike polynomials (Khotanzad and Hong, 1990). A set of polynomials is chosen in such a way that they form an orthogonal basis. It means that these polynomials are uncorrelated and projection on each of them captures a different aspect of image. The moment is defined in the following way:

$$A_{pq} = \frac{m+1}{\pi} \int_x \int_y f(x,y)[V_{pq}(x,y)]^* \,\mathrm{d}x\,\mathrm{d}y \qquad \text{where} \qquad x^2+y^2 \leq 1 \quad (2.4)$$

As before, $f(x,y)$ is image intensity function, but $[V_{pq}(x,y)]^*$ is a complex conjugate of a Zernike polynomial $V_{pq}(x,y)$ of order p and angular dependence q. For example, conjugate of $2+3i$ is $2-3i$. Further in the text there will be an alternative notation for Zernike moments used for the sake of simplicity, where a moment of order p and angular dependence q will be written as "V.p.q". I calculated Zernike moments up to order 12. The advantage of using these moments (especially, when Zernike polynomials are substituted with pseudo-Zernike ones) is their robustness against noise and small shape perturbations, i.e., small features do not gain unproportional significance in calculation of higher order moments (Xia et al., 2007).

2.3.3.3 Texture Descriptors

I characterized the texture of nurse cell nuclei by Haralick features (Haralick, 1979; Haralick et al., 1973). All of the Haralick features are based on gray level co-occurrence matrix. Assuming there are n gray levels distinguished, the co-occurrence matrix is defined as:

$$C = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \qquad (2.5)$$

where p_{ij} is a probability of seeing grayness of level j adjacent to a level i. Any fixed spatial configuration can stand for pixel adjacency. It means that all Haralick features characterize relationships between signal intensity of adjacent pixels in some way. I extracted 13 Haralick features: 2 homogeneity measures, contrast, correlation, entropy, variance, 2 information measures of correlation and average of the sum of neighbour intensities together with variance and entropy of both sum and difference of neighbour intensities. To avoid feature dependency from rotational angle I used the implementation that averages intensities of 4 neighbours before calculating probabilities.

The advantage of using this type of measurements is that most of them have a rather intuitive interpretation in terms of DNA distribution. For example, homogeneity or entropy measurements characterize the level of order within the nucleus. Also correlation measurements yield important information about the structure of the nucleus' interior. Haralick features have been successfully used in characterizing the patterns of protein localization and in discrimination between various cell types (Murphy et al., 2000).

2.3.4 Trait Variability within and between *Drosophila melanogaster* Lines

The first important question after collecting traits was whether there are any traits that are more variable among *Drosophila melanogaster* lines than within a single line. Traits having high variability among lines are the ones with the largest potential in association discovery, because having bigger than expected by chance variability at a line level implies the existence of a factor that can discriminate between lines. The discriminator can be either of genetic nature or it can be a cofactor, like sex, population origin or technical artefact of microscopy.

I measured between line variability by the Kruskal-Wallis test. The alternative hypothesis for this test is that at least two lines have significantly different medians, while the null hypothesis is that all line medians are the same. If the alternative hypothesis is accepted, it is not automatically clear which medians are the most distinct ones. However, this question is not particularly important here, because the aim is to assess between line variability, while more precise discrimination of lines is left for further analysis of individual SNPs.

Before performing Kruskal-Wallis test I averaged measurements for cells coming from the same egg chamber (image). This is necessary, because the test requires that all measurements from a line form independent random sample, but characteristics of cell nuclei from a single egg chamber can be correlated. As the Kruskal-Wallis test is non-parametric it does not require any further assumptions about the distribution of measurements. Therefore it should be robust enough to deal with a large pool of rather different traits.

An alternative approach to estimating the between line variability can be based on the measure of heritability. Heritability is assessed by calculating the proportion of the total variance of a trait that can be explained by the difference in alleles (Hartl and Clark, 2007). The trait variance can be decomposed into the genetic and environmental part by regression of a trait on the allele values. The estimation of a heritability requires the choice of a particular locus to analyze and the knowledge of its alleles for the individuals from the population. As the aim here is not exhaustive locus by locus analysis, but the initial assessment of between line variance for traits, I used the Anova test instead of a heritability measure. The Anova model was constructed so that the genotype information is not required. The phenotypic variance was modeled thus:

$$y_{jk} = \mu + L_j + \varepsilon_{jk} \tag{2.6}$$

where L_j denotes the trait variance attributable to a line j and μ is the population mean of a trait. The error term is denoted by ε_{jk}, which according to the model assumptions has to be normally distributed. Trait measurements of the same line have to represent an independent random sample, therefore measurements of nurse cell nuclei are averaged for each egg chamber before performing the Anova test to avoid trait correlation within images. Therefore y_{jk} denotes the average trait value for an egg chamber (line j and image k).

The variance component associated with L_j can be estimated in the following way:

$$MS_{line} = \frac{\sum_j n_j(\overline{Y}_j - \overline{Y})^2}{l-1} \tag{2.7}$$

where $\overline{Y}_j$ is the average of measurements corresponding to a line j and $\overline{Y}$ is the average of all measurements, while l denotes the number of lines and n_j is the number of trait measurements for a line j. The within line variance of a trait can be estimated in a similar way:

$$MS_{within} = \frac{\sum_j \sum_k (y_{jk} - \overline{Y}_j)^2}{n - l} \tag{2.8}$$

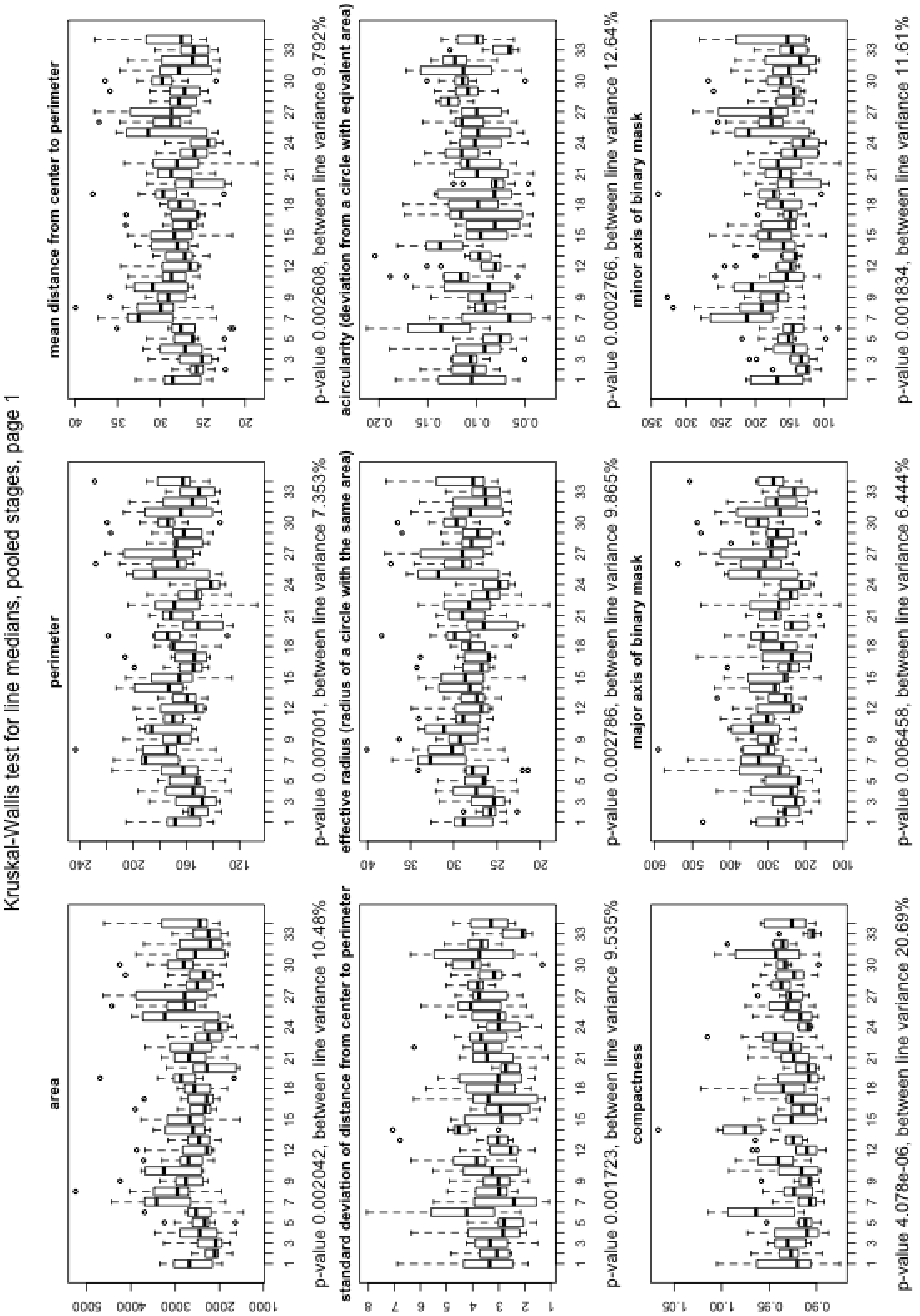
where $\overline{Y}_j$ is the average of phenotypic measurements for line j and n is the total number of measurements for a trait. Then the proportion of the total variance attributable to the between line variance can be expressed as:

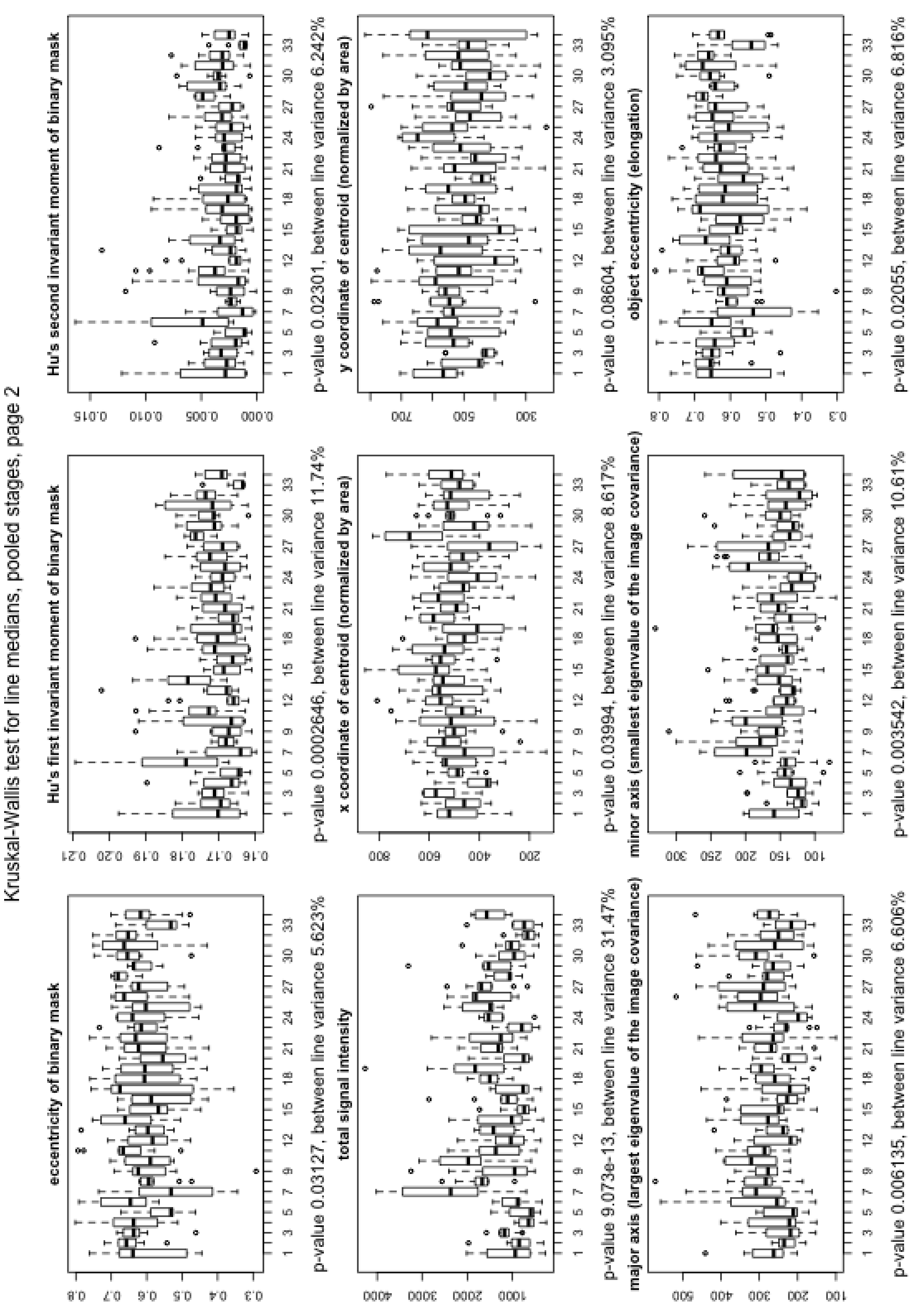
$$\frac{\sigma_l^2}{\sigma_p^2} = \frac{MS_{line} - MS_{within}}{MS_{line} - MS_{within} + n_0 \times MS_{within}} \tag{2.9}$$

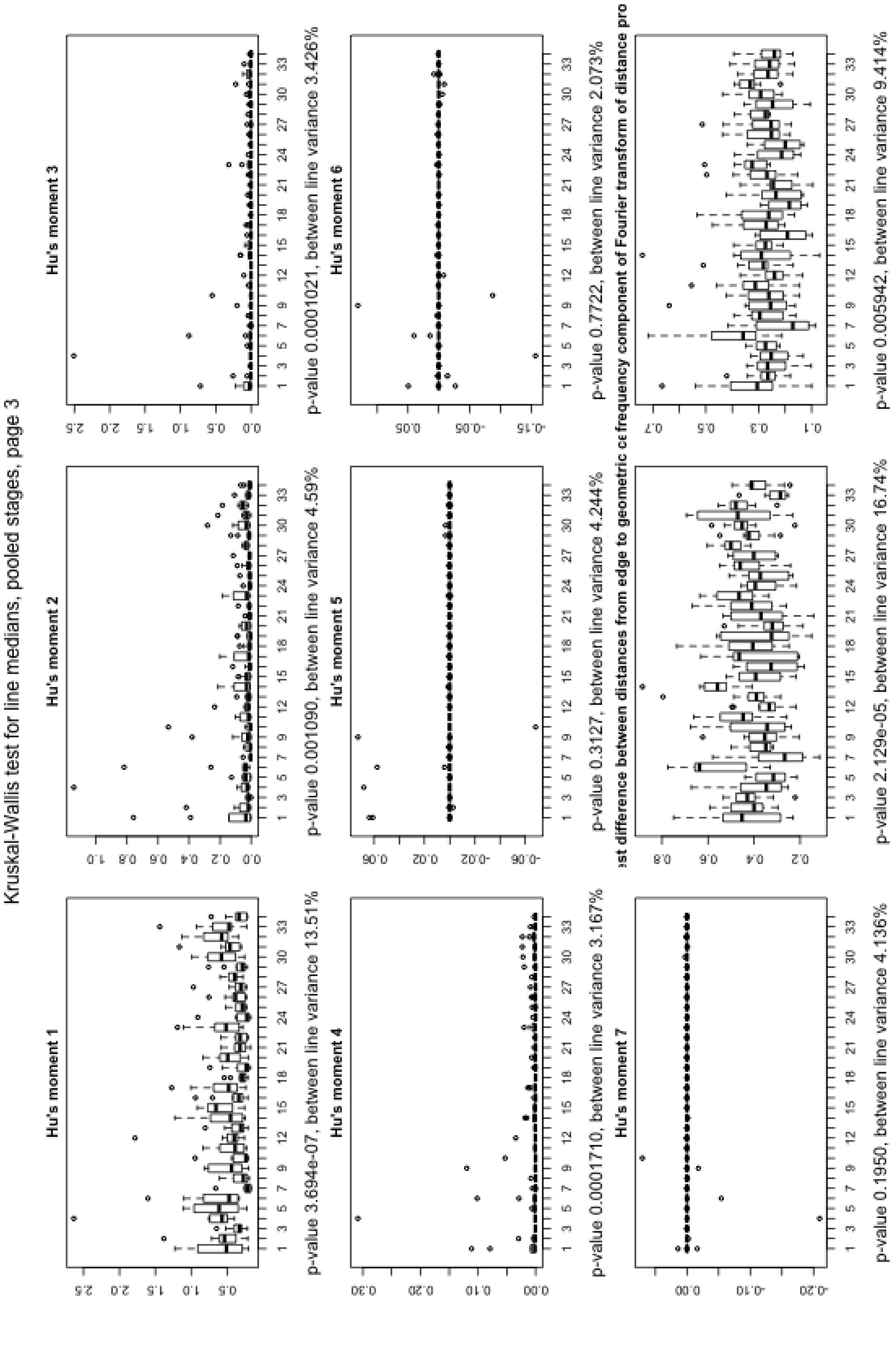
where n_0 is a normalization factor introduced due to the differences in the number of measurements for various lines and approximately reflecting degrees of freedom.

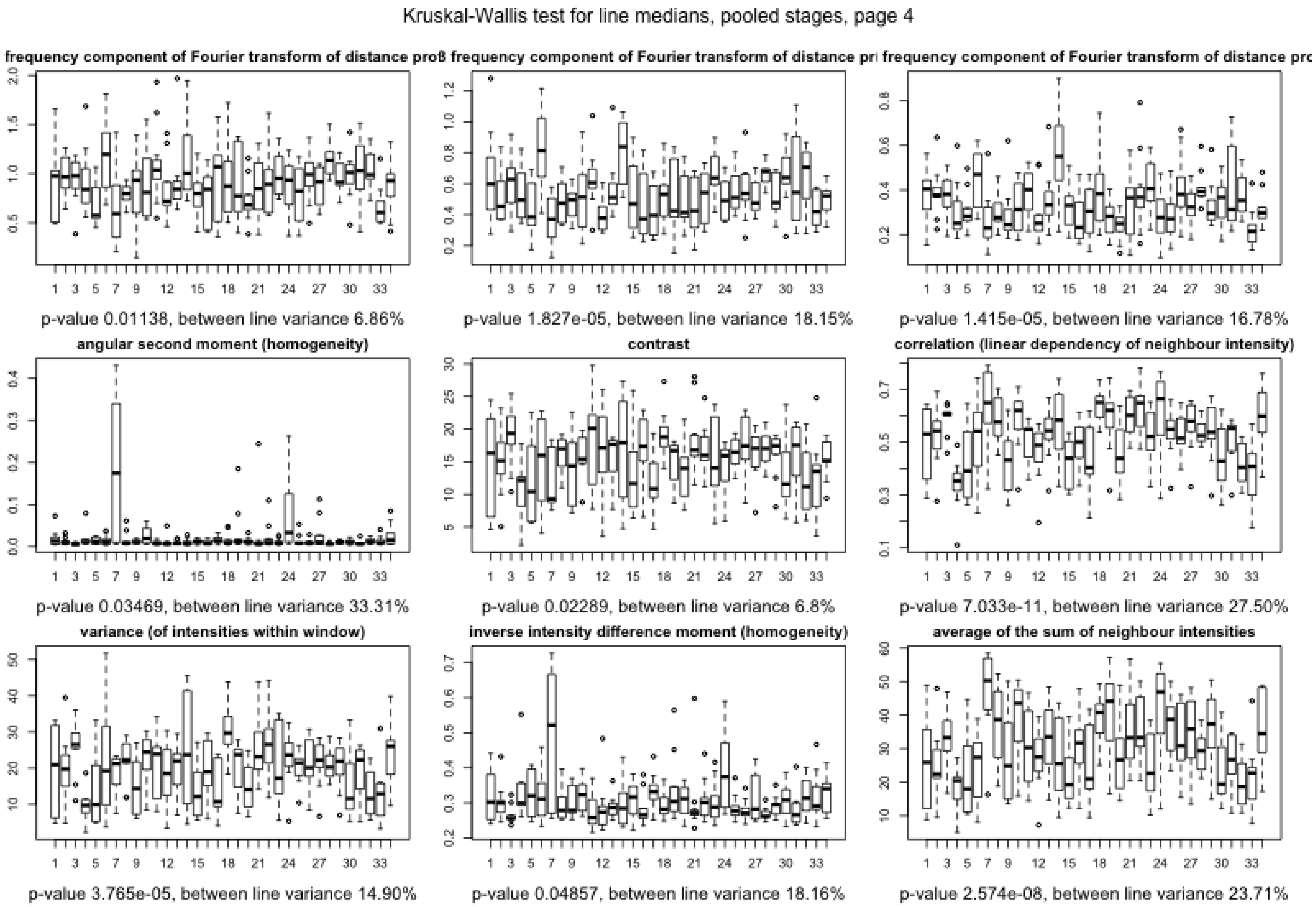
In Fig.2.7 the boxplots of traits split by line are shown, together with results of the Kruskal-Wallis test and the estimated between line variance. Note that p-values are not corrected for the number of traits tested (92 in total). Therefore only traits having p-values smaller than 10^{-5} can be considered interesting from the association study point of view. There are a few points of caution to express before describing and interpreting results. First, it is not known whether there exists a genotype and a locus that can elucidate variation between lines that are different according to the Kruskal-Wallis test. Second, the sample size for each line is 10, while it is much bigger when lines are grouped by genotypes, like it is done when an association with a particular locus is tested. Third, the sample for each line includes all developmental stages, therefore in case of uneven stage ratio or different genetic factors affecting each stage test results might not be fully accurate. In Appendix (Fig.A.1 and Fig.A.2) I provided the results of the Kruskal-Wallis test applied to each stage separately. Separate analysis has a tendency to show larger p-values due to the removal of confounding developmental factors.

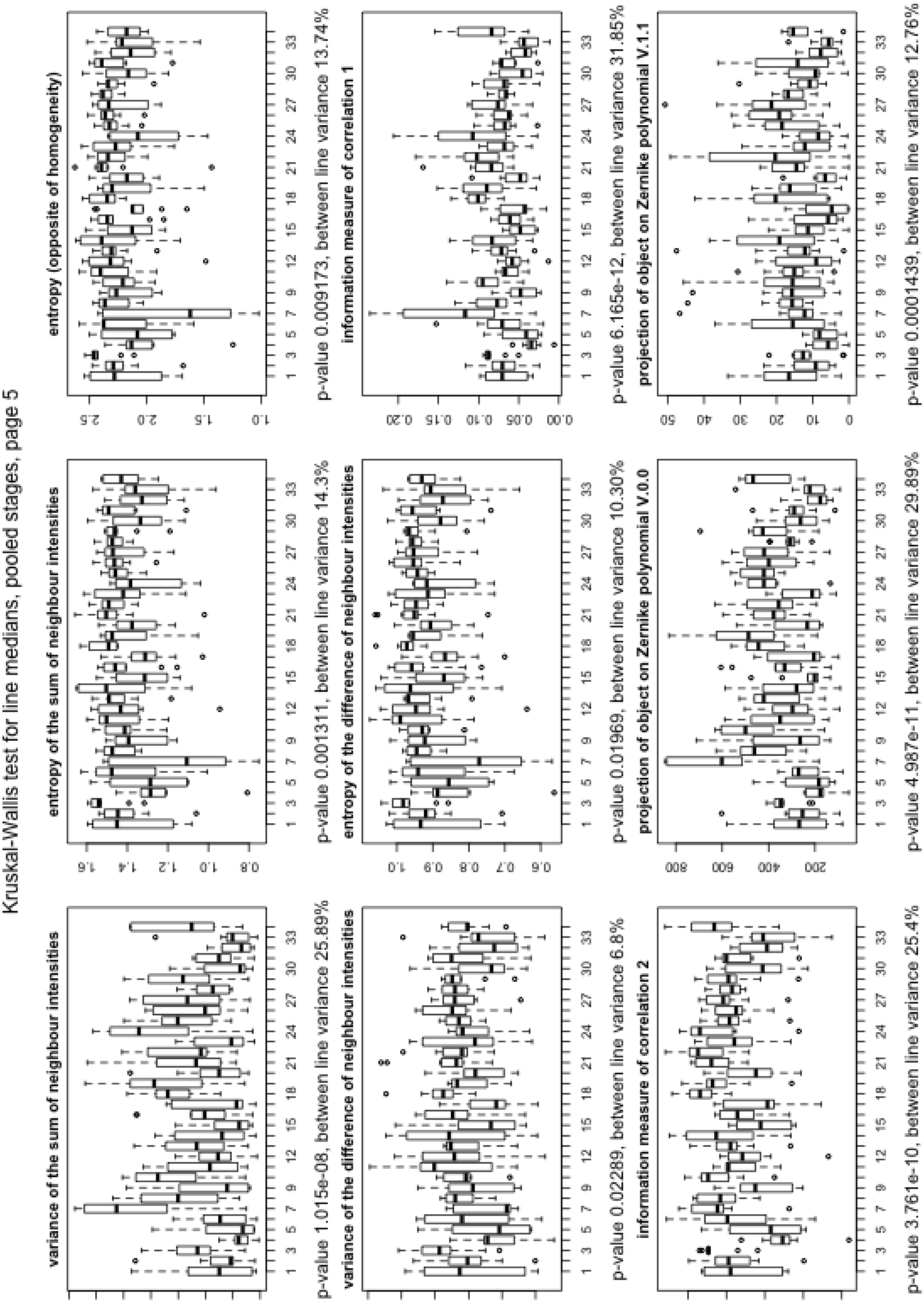
There are several interesting trends of trait variation within and between lines. There are a few purely geometric measures that seem to be interesting, like compactness, largest difference between distances from edge to geometric center

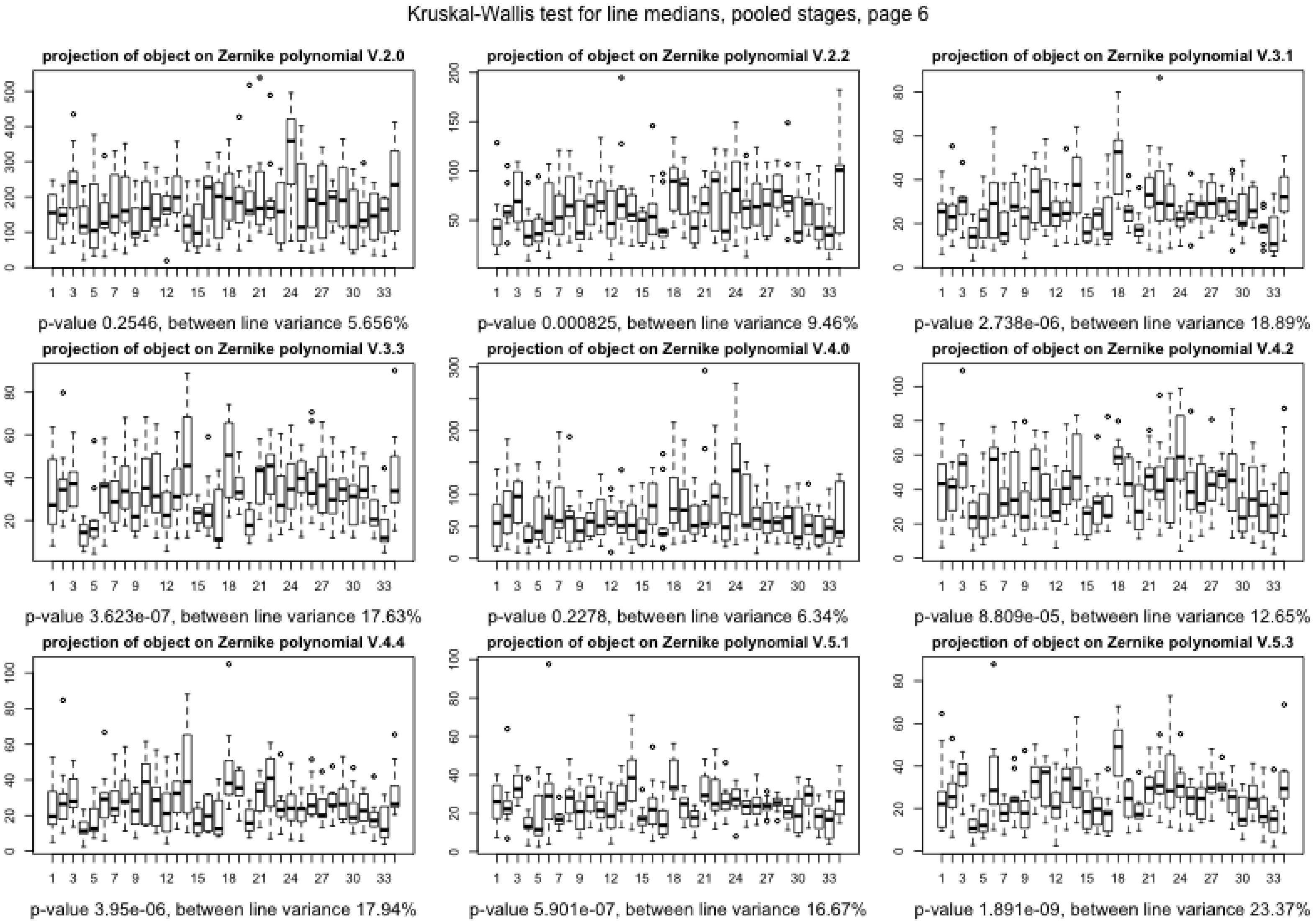

Kruskal-Wallis test for line medians, pooled stages, page 1
area
p-value 0.002042, between line variance 10.48%
1000 2000 3000 4000 5000
perimeter
p-value 0.007001, between line variance 7.353%
120 160 200 240
mean distance from center to perimeter
p-value 0.002608, between line variance 9.792%
20 25 30 35 40
standard deviation of distance from center to perimeter
p-value 0.001723, between line variance 9.535%
1 2 3 4 5 6 7 8
effective radius (radius of a circle with the same area)
p-value 0.002786, between line variance 9.865%
20 25 30 35 40
acircularity (deviation from a circle with eqivalent area)
p-value 0.0002766, between line variance 12.64%
0.05 0.10 0.15 0.20
compactness
p-value 4.078e-06, between line variance 20.69%
0.90 0.95 1.00 1.05
major axis of binary mask
p-value 0.006458, between line variance 6.444%
100 200 300 400 500 600
minor axis of binary mask
p-value 0.001834, between line variance 11.61%
100 150 200 250 300 350
1 3 5 7 9 12 15 18 21 24 27 30 33

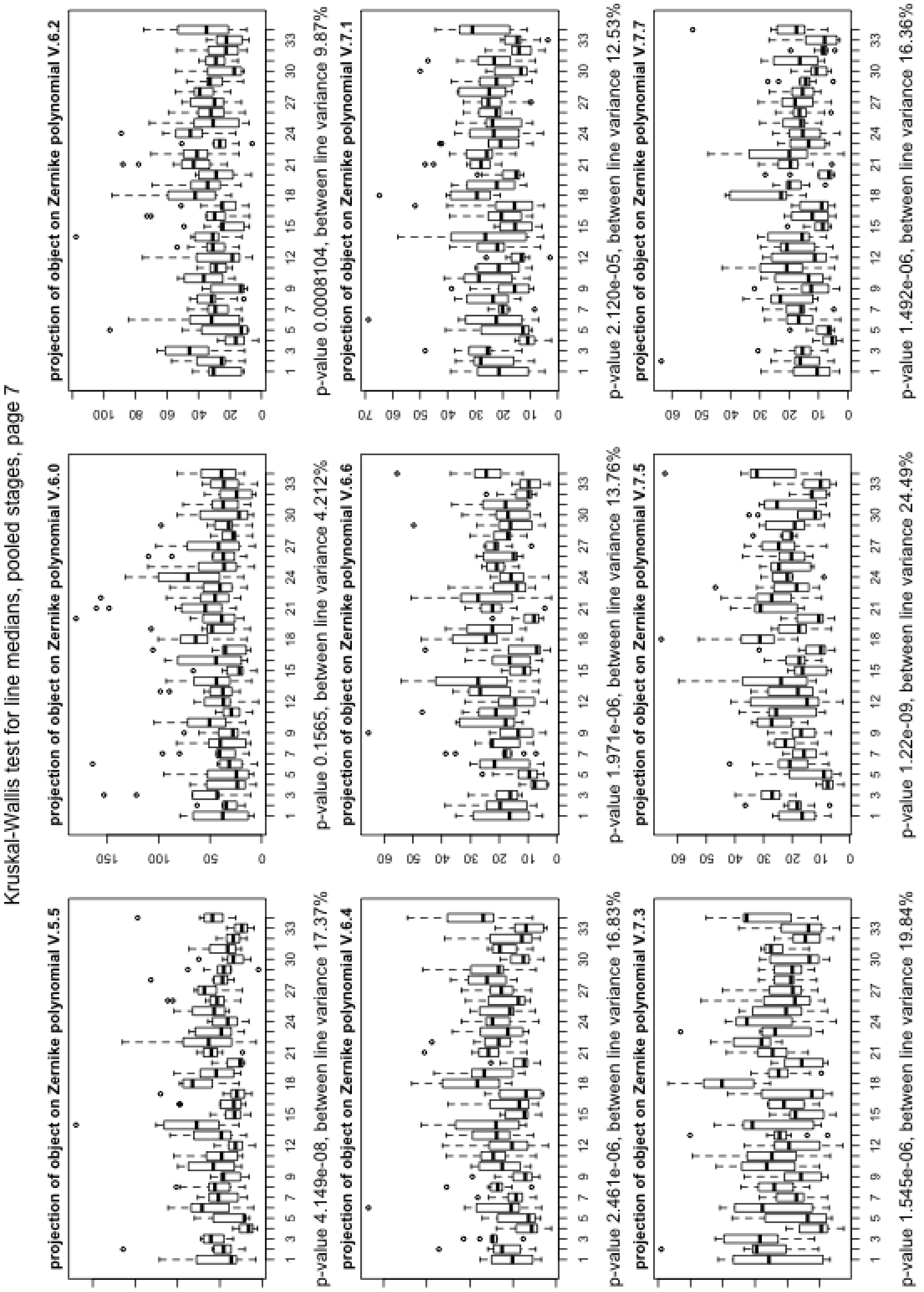

Kruskal-Wallis test for line medians, pooled stages, page 2
eccentricity of binary mask
0.3 0.4 0.5 0.6 0.7 0.8
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.03127, between line variance 5.623%
Hu's first invariant moment of binary mask
0.16 0.17 0.18 0.19 0.20 0.21
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.0002646, between line variance 11.74%
Hu's second invariant moment of binary mask
0.000 0.005 0.010 0.015
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.02301, between line variance 6.242%
total signal intensity
1000 2000 3000 4000
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 9.073e-13, between line variance 31.47%
x coordinate of centroid (normalized by area)
200 400 600 800
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.03994, between line variance 8.617%
y coordinate of centroid (normalized by area)
300 500 700
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.08604, between line variance 3.095%
major axis (largest eigenvalue of the image covariance)
100 200 300 400 500
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.006135, between line variance 6.606%
minor axis (smallest eigenvalue of the image covariance)
100 150 200 250 300
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.003542, between line variance 10.61%
object eccentricity (elongation)
0.3 0.4 0.5 0.6 0.7 0.8
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.02055, between line variance 6.816%

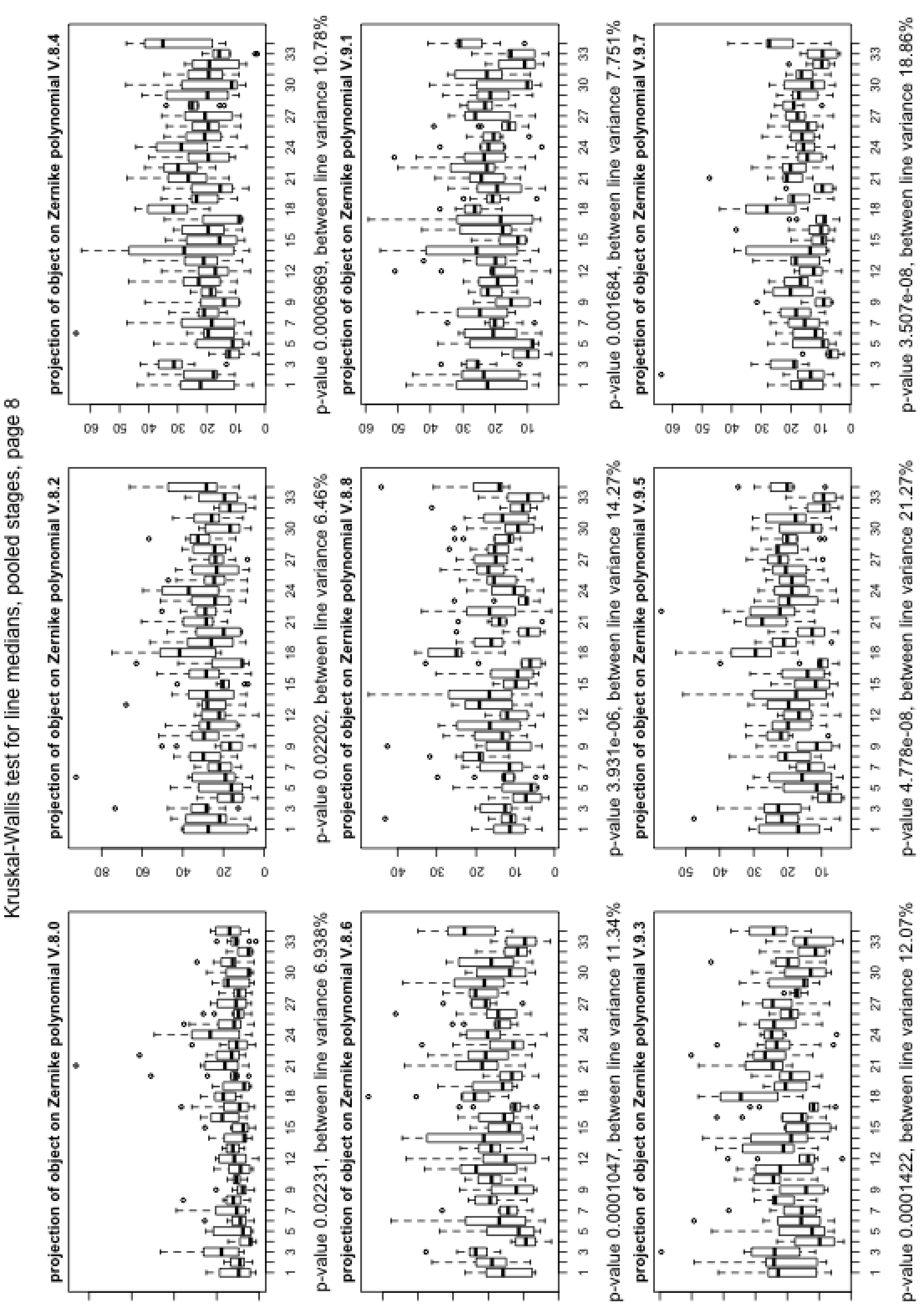

Kruskal-Wallis test for line medians, pooled stages, page 3
Hu's moment 1
p-value 3.694e-07, between line variance 13.51%
Hu's moment 2
p-value 0.001090, between line variance 4.59%
Hu's moment 3
p-value 0.0001021, between line variance 3.426%
Hu's moment 4
p-value 0.0001710, between line variance 3.167%
Hu's moment 5
p-value 0.3127, between line variance 4.244%
Hu's moment 6
p-value 0.7722, between line variance 2.073%
Hu's moment 7
p-value 0.1950, between line variance 4.136%
est difference between distances from edge to geometric ce
p-value 2.129e-05, between line variance 16.74%
frequency component of Fourier transform of distance pro
p-value 0.005942, between line variance 9.414%
1 3 5 7 9 12 15 18 21 24 27 30 33
0.5 1.0 1.5 2.0 2.5
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.5 1.0 1.5 2.0 2.5
0.00 0.10 0.20 0.30
-0.06 -0.02 0.02 0.06
-0.15 -0.05 0.05
-0.20 -0.10 0.00
0.2 0.4 0.6 0.8
0.1 0.3 0.5 0.7

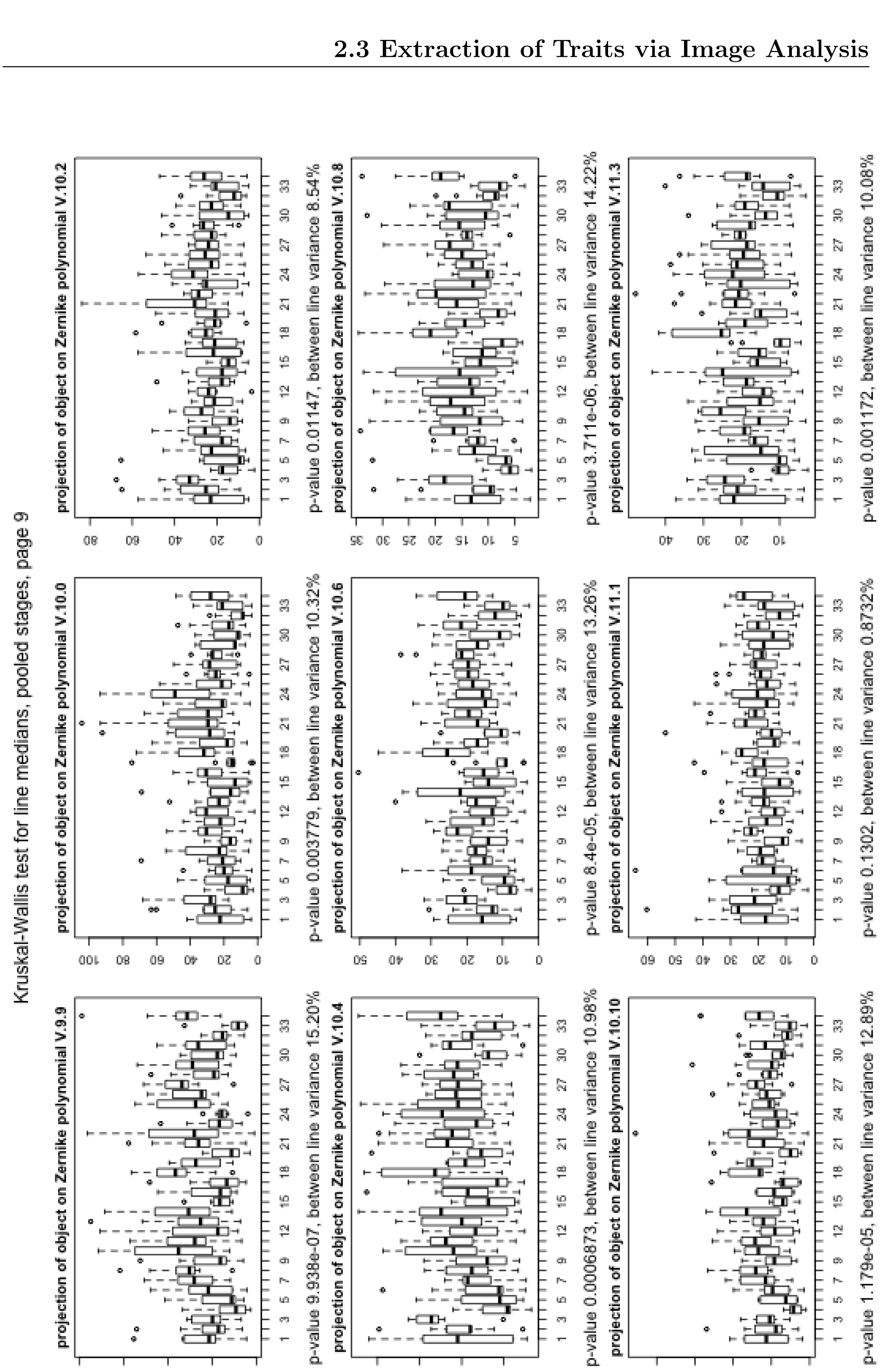

Kruskal-Wallis test for line medians, pooled stages, page 4
frequency component of Fourier transform of distance proß
0.5
1.0
1.5
2.0
p-value 0.01138, between line variance 6.86%
frequency component of Fourier transform of distance pr
0.2
0.4
0.6
0.8
1.0
1.2
p-value 1.827e-05, between line variance 18.15%
frequency component of Fourier transform of distance prc
0.2
0.4
0.6
0.8
p-value 1.415e-05, between line variance 16.78%
angular second moment (homogeneity)
0.0
0.1
0.2
0.3
0.4
p-value 0.03469, between line variance 33.31%
contrast
5
10
15
20
25
30
p-value 0.02289, between line variance 6.8%
correlation (linear dependency of neighbour intensity)
0.1
0.3
0.5
0.7
p-value 7.033e-11, between line variance 27.50%
variance (of intensities within window)
10
20
30
40
50
p-value 3.765e-05, between line variance 14.90%
inverse intensity difference moment (homogeneity)
0.2
0.3
0.4
0.5
0.6
0.7
p-value 0.04857, between line variance 18.16%
average of the sum of neighbour intensities
10
20
30
40
50
60
p-value 2.574e-08, between line variance 23.71%
1 3 5 7 9 12 15 18 21 24 27 30 33

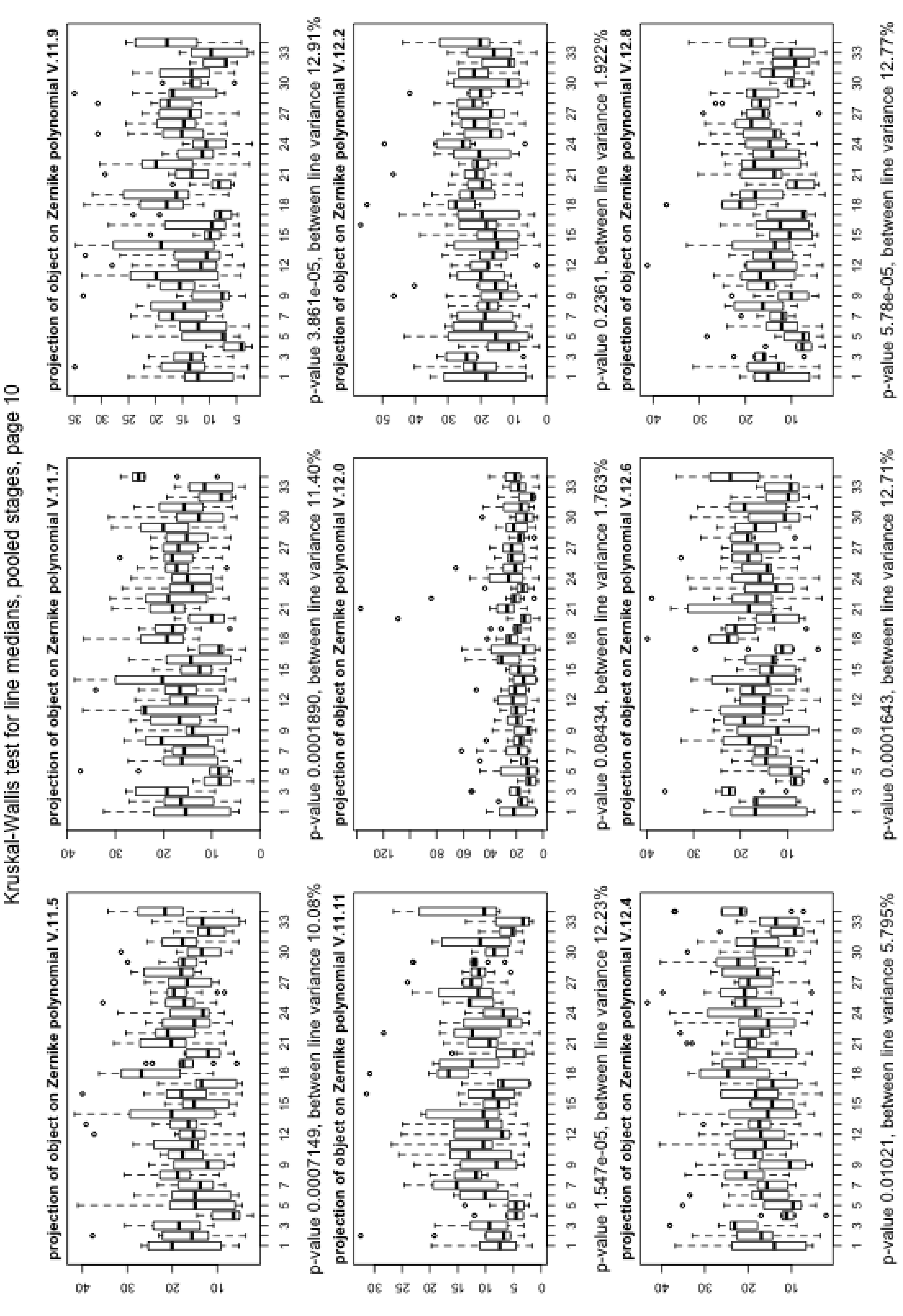

Kruskal-Wallis test for line medians, pooled stages, page 5
variance of the sum of neighbour intensities
p-value 1.015e-08, between line variance 25.89%
entropy of the sum of neighbour intensities
p-value 0.001311, between line variance 14.3%
entropy (opposite of homogeneity)
p-value 0.009173, between line variance 13.74%
variance of the difference of neighbour intensities
p-value 0.02289, between line variance 6.8%
entropy of the difference of neighbour intensities
p-value 0.01969, between line variance 10.30%
information measure of correlation 1
p-value 6.165e-12, between line variance 31.85%
information measure of correlation 2
p-value 3.761e-10, between line variance 25.4%
projection of object on Zernike polynomial V.0.0
p-value 4.987e-11, between line variance 29.89%
projection of object on Zernike polynomial V.1.1
p-value 0.0001439, between line variance 12.76%
1 3 5 7 9 12 15 18 21 24 27 30 33


Kruskal-Wallis test for line medians, pooled stages, page 6
projection of object on Zernike polynomial V.2.0
0 100 200 300 400 500
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.2546, between line variance 5.656%
projection of object on Zernike polynomial V.2.2
50 100 150 200
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.000825, between line variance 9.46%
projection of object on Zernike polynomial V.3.1
0 20 40 60 80
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 2.738e-06, between line variance 18.89%
projection of object on Zernike polynomial V.3.3
20 40 60 80
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 3.623e-07, between line variance 17.63%
projection of object on Zernike polynomial V.4.0
0 50 100 200 300
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.2278, between line variance 6.34%
projection of object on Zernike polynomial V.4.2
0 20 40 60 80 100
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 8.809e-05, between line variance 12.65%
projection of object on Zernike polynomial V.4.4
0 20 40 60 80 100
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 3.95e-06, between line variance 17.94%
projection of object on Zernike polynomial V.5.1
0 20 40 60 80 100
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 5.901e-07, between line variance 16.67%
projection of object on Zernike polynomial V.5.3
0 20 40 60 80
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 1.891e-09, between line variance 23.37%


Kruskal-Wallis test for line medians, pooled stages, page 7
projection of object on Zernike polynomial V.5.5
p-value 4.149e-08, between line variance 17.37%
projection of object on Zernike polynomial V.6.0
p-value 0.1565, between line variance 4.212%
projection of object on Zernike polynomial V.6.2
p-value 0.0008104, between line variance 9.87%
projection of object on Zernike polynomial V.6.4
p-value 2.461e-06, between line variance 16.83%
projection of object on Zernike polynomial V.6.6
p-value 1.971e-06, between line variance 13.76%
projection of object on Zernike polynomial V.7.1
p-value 2.120e-05, between line variance 12.53%
projection of object on Zernike polynomial V.7.3
p-value 1.545e-06, between line variance 19.84%
projection of object on Zernike polynomial V.7.5
p-value 1.22e-09, between line variance 24.49%
projection of object on Zernike polynomial V.7.7
p-value 1.492e-06, between line variance 16.36%
1 3 5 7 9 12 15 18 21 24 27 30 33


Kruskal-Wallis test for line medians, pooled stages, page 8
projection of object on Zernike polynomial V.8.0
0
50
100
150
200
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.02231, between line variance 6.938%
projection of object on Zernike polynomial V.8.2
0
20
40
60
80
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.02202, between line variance 6.46%
projection of object on Zernike polynomial V.8.4
0
10
20
30
40
50
60
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.0006969, between line variance 10.78%
projection of object on Zernike polynomial V.8.6
0
10
20
30
40
50
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.0001047, between line variance 11.34%
projection of object on Zernike polynomial V.8.8
0
10
20
30
40
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 3.931e-06, between line variance 14.27%
projection of object on Zernike polynomial V.9.1
10
20
30
40
50
60
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.001684, between line variance 7.751%
projection of object on Zernike polynomial V.9.3
0
10
20
30
40
50
60
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.0001422, between line variance 12.07%
projection of object on Zernike polynomial V.9.5
10
20
30
40
50
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 4.778e-08, between line variance 21.27%
projection of object on Zernike polynomial V.9.7
0
10
20
30
40
50
60
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 3.507e-08, between line variance 18.86%


Kruskal-Wallis test for line medians, pooled stages, page 9
projection of object on Zernike polynomial V.9.9
p-value 9.938e-07, between line variance 15.20%
projection of object on Zernike polynomial V.10.0
p-value 0.003779, between line variance 10.32%
projection of object on Zernike polynomial V.10.2
p-value 0.01147, between line variance 8.54%
projection of object on Zernike polynomial V.10.4
p-value 0.0006873, between line variance 10.98%
projection of object on Zernike polynomial V.10.6
p-value 8.4e-05, between line variance 13.26%
projection of object on Zernike polynomial V.10.8
p-value 3.711e-06, between line variance 14.22%
projection of object on Zernike polynomial V.10.10
p-value 1.179e-05, between line variance 12.89%
projection of object on Zernike polynomial V.11.1
p-value 0.1302, between line variance 0.8732%
projection of object on Zernike polynomial V.11.3
p-value 0.001172, between line variance 10.08%
1 3 5 7 9 12 15 18 21 24 27 30 33


Kruskal-Wallis test for line medians, pooled stages, page 10
projection of object on Zernike polynomial V.11.5
p-value 0.0007149, between line variance 10.08%
projection of object on Zernike polynomial V.11.7
p-value 0.0001890, between line variance 11.40%
projection of object on Zernike polynomial V.11.9
p-value 3.861e-05, between line variance 12.91%
projection of object on Zernike polynomial V.11.11
p-value 1.547e-05, between line variance 12.23%
projection of object on Zernike polynomial V.12.0
p-value 0.08434, between line variance 1.763%
projection of object on Zernike polynomial V.12.2
p-value 0.2361, between line variance 1.922%
projection of object on Zernike polynomial V.12.4
p-value 0.01021, between line variance 5.795%
projection of object on Zernike polynomial V.12.6
p-value 0.0001643, between line variance 12.71%
projection of object on Zernike polynomial V.12.8
p-value 5.78e-05, between line variance 12.77%
1 3 5 7 9 12 15 18 21 24 27 30 33

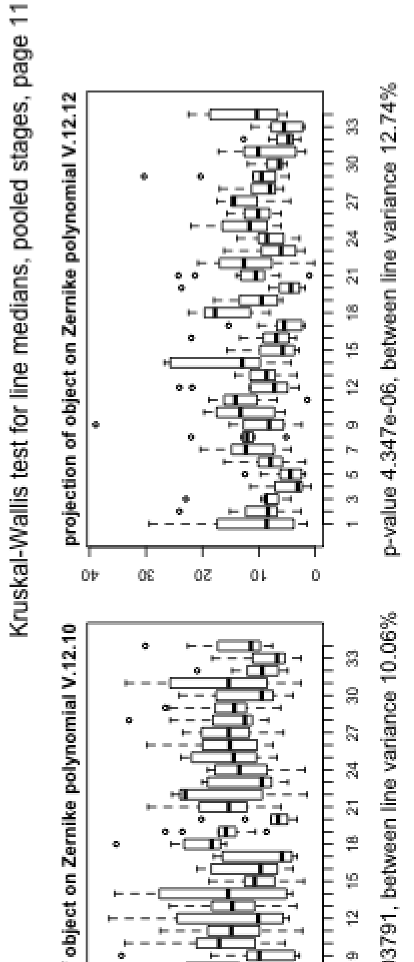


Figure 2.7: Traits characterizing nurse cell nuclei. Numbers on x axis correspond to *Drosophila melanogaster* lines, while y axis denotes trait value. Kruskal-Wallis test p-value and between line variance is indicated below each trait panel.

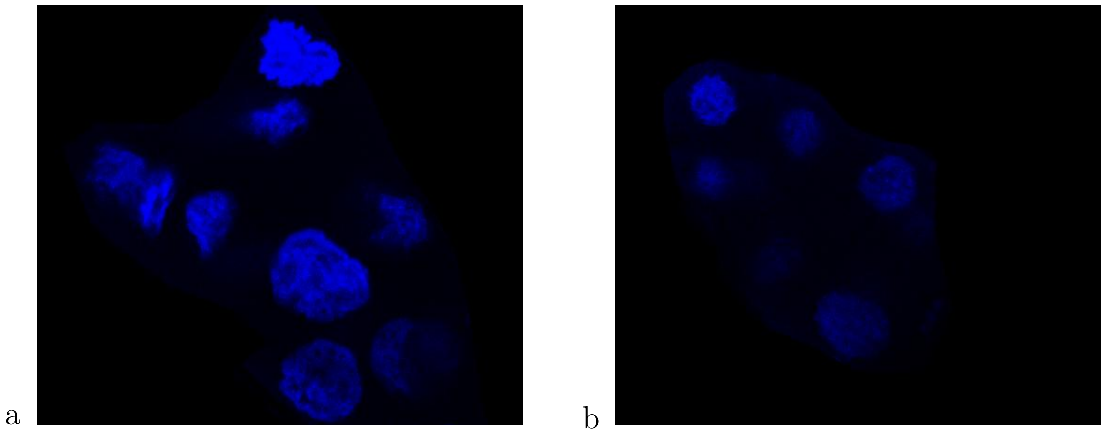


Figure 2.8: Extreme phenotypes for cell nuclei compactness: a. low compactness (line RAL-358), b. high compactness (line RAL-362).

and Hu's first invariant moment of object shape. Also two frequency components of distance profile show quite significant variation among lines. It seems that nurse cell nuclei have a tendency to differ in terms of their shape, especially in their degree of resembling a circular shape. Even smaller p-values are obtained for some of the moments: total signal intensity and four first Hu's moments are among the most variable traits between lines. Many pronounced differences between lines were discovered with texture descriptors: half of Haralick features show small p-values. Some Zernike moments are interesting as well, such as V.3.3., V.5.3. and V.5.5. It is interesting to note that Zernike polynomials of lower order give more understanding of line variability than those of highest orders. Polynomials of degree 3, 5, 6, 7 are the most represented among significant traits, while only a few moments of order 9 and 12 can be classified as interesting. The same tendency can be observed for Hu's invariants as lower order moments show smaller p-values. This could be explained by the importance of a plain signal intensity which is more related to lower order polynomials. Alternatively, more global characteristics represented by lower order polynomials have greater power to discern between lines.

In most cases the estimate of between line variance is consistent with the Kruskal-Wallis test, i.e., the smaller is the p-value, the bigger is the variance component attributable to the difference between lines. This relationship is well illustrated by the first page of Fig.2.7. Nevertheless, there are a few cases where this rule does not hold; for example, Hu's moments exhibit inconsistent trends for the relationship of between line variance and various magnitudes of p-values (see the third page of Fig.2.7). It seems that these exceptions can be fully attributed to the violation of the normality assumption.

There are a few traits with between line variance over 20%. This is sufficiently large proportion for a trait to be a good candidate for finding associations. Among them are Zernike moments V.9.5, V.7.5, V.5.3 and V.0.0, both information measures of correlation for signal intensities, average and variance of the sum of the neighbour signal intensities, correlation of the neighbour intensity and homogeneity, total signal intensity and compactness.

I looked at some of the significant traits in more detail. To understand what kind of difference between lines is captured by the Kruskal-Wallis test I studied samples of extreme phenotype values. For this purpose measurements corresponding to the 0.025 and 0.975 quantile were taken as examples of extreme phenotypes. Here by measurement I mean the average measurement for cell nuclei within one egg chamber. Fixing a quantile instead of taking the largest or smallest measurement helps to avoid outliers that are not representative for a trait.

Compactness is one of the geometric shape descriptors that show large between-line variation. In Fig.2.8 examples of cell nuclei with large and small compactness are collected. Cell nuclei having small compactness have a tendency to deviate from the circular shape and are rather large. Compact cell nuclei have almost perfect circular shape and are of modest size.

I depicted extreme phenotypes for variance of signal intensities within window as an example of a significant trait describing cell texture (Fig.2.9). The texture of cell nuclei with high signal variance has more pronounced granularity and lots of small details. Also the total signal intensity for those cell nuclei tends to be larger, because more variance can potentially be observed for bigger intensity range. This trait captures local properties of image, but has limitations in identification of global aspects of cell nuclei.

One of the most interesting moment descriptors was given by Zernike moment V.3.3. Examples in Fig.2.10 indicate that Zernike polynomials capture charac-

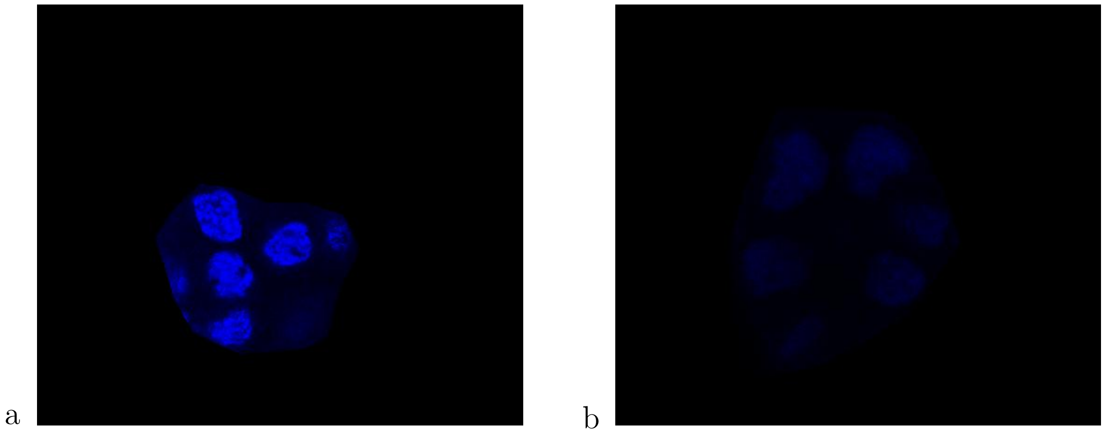


Figure 2.9: Extreme phenotypes for variance of signal intensities within window: a. high variance (line RAL-304), b. low variance (line RAL-208).

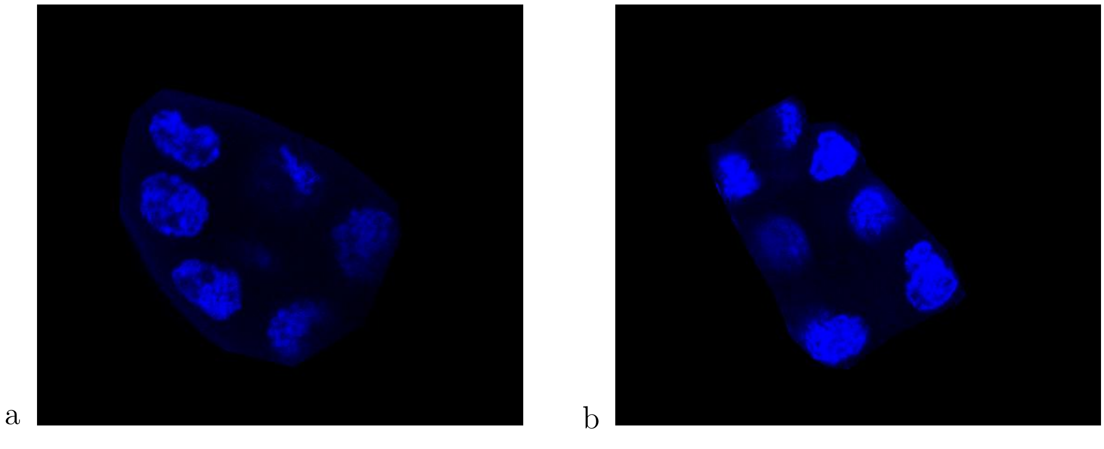


Figure 2.10: Extreme phenotypes for Zernike moment V.3.3.: a. high moment values (line RAL-427), b. low moment values (line RAL-714.

teristics of cell nuclei that are quite different from texture traits. Both extreme phenotype samples for Zernike moment V.3.3. show a lot of small detail and granularity. At the same time the distribution of DNA within those nuclei is visibly different. In the first case it is skewed towards one end, while in the second case the localization of DNA is very even. The size of nuclei in both images is quite similar, as is the shape. Most probably differences between these egg chambers would elude geometric characterization. Thus, the main role of Zernike moments is to identify various global patterns of DNA distribution.

Chapter 3

Models for Association Study in *Drosophila melanogaster*

3.1 A Simple Model and its Limitations

The aim of this section is to introduce models used in association discovery for *Drosophila melanogaster*. First, simple non-parametric tests for associations are proposed. Then more sophisticated modeling techniques that are capable of dealing with nested levels of phenotypic measurements are explained.

As the *Drosophila melanogaster* strains used in the association study are homozygous for most of the loci and the few regions where this does not hold are excluded, it is sensible to propose an additive model to test for genotype/phenotype associations:

$$y_i = \mu + \beta I(x_i = AA) + \varepsilon_i \tag{3.1}$$

where y_i is the value of a phenotype for an individual i, x_i is the genotype of locus under study for an individual i and ε_i is random variation. $I(\cdot)$ stands for an indicator function, which in the current case codes allele combination aa as 0, but AA as 1. Then β is the magnitude of the allele effect and μ is the phenotype mean for individuals with alleles aa.

At this moment a clarification of what y_i denotes is necessary. There is a choice between using a single phenotypic measurement for each line and between using a single measurement for each oocyte image (10 measurements per line). When studying nurse cells there is even a third option - to get a single measurement from each cell nucleus (there are 5 or 6 cells per image). The errors ε_i have to be independently distributed under any model chosen. For example, if there is any internal correlation present among the measurements coming from a single line, it is not possible to ignore it and just fit equation 3.1. Such a situation is imaginable in the case of at least a few image measurements, because each line represents a different sample that is analysed separately on a microscope. For example, manual focusing on each sample slide can subtly affect some of the measurements. If the time for experiments was not limited and samples taken from a single line incorporated more oocytes from more flies, it would be beneficial for further analysis to have 3 microscope slides per line to separate technical variation from the variation within fly line. However, when performing a test for genotype effect, measurements from several lines are pooled according to genotype. This should allow the estimation of the internal correlation among the measurements coming from a single slide well enough, although this correlation cannot be separated from variation associated with line.

There are two solutions for the level of detail of phenotypic measurements. The first is to aggregate lower levels of measurements and then use a simple model to test for the genotypic effect using one averaged measurement per line. The second is to use some kind of nested model, where we are estimating internal correlations at lower levels and then using this information in testing genotype effect. For the first scenario the Mann-Whitney test is proposed, while for the second scenario a mixed model nested Anova is applied. Both approaches were tried out in the association study of *Drosophila melanogaster*.

3.2 Testing the Significance of Genotype

The aim of building a model is to test the alternative hypothesis that the locus on the genome has a significant impact on phenotype as opposed to the null hypothesis that the locus does not influence phenotype values. Various asymptotic distributions are used to obtain p-values from the test statistic for each loci. P-values from single locus analysis can be converted into whole genome significance level either by application of Bonferroni correction or by treating them as metrics of association in permutation tests.

First, a non-parametric Mann-Whitney test is considered. The purpose of this test is to check whether phenotype medians differ for various genotype classes. This type of test seems to be appealing, because the corresponding rank statistics are distribution free, i.e. no assumption about the distribution of errors ε_i in equation 3.1 has to be made. It is important in the current association study, because it is difficult to assume some distribution for 92 rather different phenotypes (see Kruskal-Wallis test pictures in Fig.2.7). However, this approach requires that the impact of other, non-genomic, cofactors on phenotype are somehow eliminated before performing this test. This limitation is caused by the fact that the Mann-Whitney test, as well as a more general class of simple linear rank statistics, cannot capture cofactors in the model without them being part of hypothesis being tested. It means that it is possible to test the combined effect of a specific allele and specific value of cofactor, but it is not feasible to construct a test for the effect of allele alone.

I will define the Mann-Whitney statistic more precisely. The statistic is based on the ranks assigned to phenotype values, where rank indicates how large the phenotype value is relative to other measurements. The rank is defined from the equation $X_i = X_{N(R_{Ni})}$, where $X_{N(1)} \leq X_{N(2)} \leq \cdots \leq X_{N(N)}$ is the order statistic of a set of real-valued observations $X_1, \ldots, X_N$ (van der Vaart, 1998). Assuming that the sample corresponding to the first allele has size m and the second allele sample has size n and R_N is the rank vector of a pooled sample of size $N = m+n$, the Mann-Whitney statistic is defined as C from the following equation:

$$C = \sum_{i=m+1}^{N} R_{Ni} - \frac{n(n+1)}{2} \tag{3.2}$$

This definition holds apart from cases when $C < m \cdot n - C$. In such a situation the Mann-Whitney statistic is defined as $m \cdot n - C$. The value of the statistic will be further denoted with U.

According to the theorem about rank tests, the Mann-Whitney statistic exhibits certain asymptotic behaviour, i.e. the variable

$$u = \frac{U - \dfrac{mn}{2}}{\sqrt{\dfrac{mn(N+1)}{12}}} \tag{3.3}$$

is distributed as $N(0,1)$ as $N \to \infty$ (van der Vaart, 1998). (Certain regularity conditions have to be satisfied as well, such as the requirement that $m/n \to c$ when $N \to \infty$, where c is some constant.) Here N is nearly 40 and the largest of m and n is guaranteed to be at least 17 and almost always is greater than 20 (see minor allele frequencies for *Drosophila melanogaster* SNPs in Fig.3.2b). Thus, according to recommendations in literature, it is safe to use the normal approximation instead of calculating the exact probability by combinatorial techniques (Sokal and Rohlf, 1995). Therefore p-values for tests of the significance of genotype effect were obtained from this approximation.

Next, a mixed model nested Anova is considered for association detection. I already implicitly defined various levels of measurements as the level of genotype, level of a single line, level of image and level of a single nurse cell. The chosen Anova model is of mixed type, because the genotype effect is fixed, lines are nested within genotypes and images and cells are random factors. Then the phenotypic variance is modeled thus:

$$y_{ijkh} = \mu + G_i + L_{ij} + C_{ijk} + \varepsilon_{ijkh} \tag{3.4}$$

where G_i is the variance in phenotype attributable to genotype i, L_{ij} denotes the within-line variance for line j belonging to genotype i and C_{ijk} is the variance of cells within an image k, which is nested in line j and genotype i. The error term is denoted by ε_{ijkh}, which according to the model assumptions should be normally distributed. As the three levels of measurements are considered simultaneously, the total number of measurements is almost 2000. Therefore the assumption of error normality seems to be plausible due to the central limit theorem.

The test of significance for a single term in the model (3.4) is based on checking whether the sum of squares estimating the variance associated with the term is considerably larger than the sum of squares evaluating the variance of another term. Typically only sums of squares for two adjacent terms in the model would be compared. For example, to test that the difference between nurse cells coming from different egg chambers is larger than the difference among cells within the same egg chamber, one would compare the variance of C_{ijk} with the error variance ε_{ijkh}. As the aim is to test the significance of the genetic component, my interest is focused on terms G_i and L_{ij}. The variance associated with G_i can be estimated in the following way:

$$MS_{genotype} = \frac{l \cdot m \cdot c \sum_i (\overline{Y}_i - \overline{Y})^2}{g-1} \tag{3.5}$$

where $\overline{Y}_i$ is the average of measurements corresponding to a genotype i and $\overline{Y}$ is the average of all measurements. As for theremaining constants, l denotes the number of lines within the genotype class, m is the number of images (egg chambers) per line, c is the number of cells per egg chamber and g is the number of different genotypes. Variance associated with L_{ij} can be estimated in a similar way:

$$MS_{line} = \frac{m \cdot c \sum_i \sum_j (\overline{Y}_{ij} - \overline{Y}_i)^2}{g \cdot (l-1)} \tag{3.6}$$

where $\overline{Y}_{ij}$ is the average of phenotypic measurements for line ij and $\overline{Y}_i$ is the average of measurements for genotype i. Then the statistic measuring the significance of genotypic effect can be expressed as:

$$\frac{MS_{genotype}}{MS_{line}} = \frac{g \cdot l \cdot (l-1) \sum_i (\overline{Y}_i - \overline{Y})^2}{(g-1) \sum_i \sum_j (\overline{Y}_{ij} - \overline{Y}_i)^2} \tag{3.7}$$

This statistic is distributed as the F distribution with degrees of freedom $g-1$ and $g(l-1)$.

The accuracy of this test relies upon the assumption that the experimental design is fully balanced, i.e., there is a single number l that corresponds to the number of lines embedded within each genotype and also m and c are not varying over lines and images. However, the phenotypic measurements have some inherent unbalancedness that cannot be corrected by choosing a different experimental design. First, the number of lines that correspond to each genotype is highly variable for different SNPs. Second, the number of cells (or rather their visibility) within an image varies between egg chambers and we do not want to trim these measurements to some constant, because that would mean having samples from only a few locations within egg chambers, which might lead to bias. Only the number of images per line is constant (10), although there are factors potentially disturbing this balance as well. In the case of unbalanced design, the Sattherthwaite approximation can be used to make the p-values more accurate (Sokal and Rohlf, 1995). The applicability of this method is subject to a few conditions. For example, if the test of interest concerns the variance ratio $\frac{MS_{genotype}}{MS_{line}}$, it is necessary to check that $df_{genotype} < 100$ and $df_{genotype} < 2df_{line}$. Both conditions are met within the current study, because $df_{genotype} = g - 1 = 1$. Thus, after testing for one more inequality, which is not described here because of its complexity, it is feasible to use this approximation instead of the less precise test defined in the (3.7).

3.3 Models with Cofactors

Now models suitable for testing genotypic effect at various phenotype granularities are specified. However, there are a few cofactors having impact on phenotype alongside genetic factors. Here I am proposing several models that deal with non-genetic parameters as well as with genetic.

As can be seen from the description of experimental procedures, samples collected represent different developmental stages of oocyte, namely stages 9 and 10. There are 10 samples for each *Drosophila melanogaster* line, where the mixture of stages is completely random. As there is no biological reason for some of the stages being more represented than others, the median of stage ratios should be 1 for lines pooled by the allele of some SNP. However, the variation of stage ratios is of bigger interest, because uneven mixture of developmental stages can affect the applicability of models and choice of cofactors. If the variance was small, we could assume that measurements are distributed in the same way within all lines. Then it is possible to omit developmental stage from the analysis, at least if it is expected that the genotype effect has the same direction and magnitude for both stages. If the genotype effects were contrasting in both stages, statistical power would be lost by analysing all measurements by a model not containing developmental stage. If the variance of stage ratios was small it would also be possible to take average over 10 samples from a line and then use a non-parametric Mann-Whitney test for analysis. The plot of the logarithm of stage ratios in Fig.3.1 shows that they are normally distributed with mean 0. Most of the data points fall within the limits of $(-0.5, 0.5)$. It means that most stage ratios are roughly between $(1/3, 3)$, which follows from transformation of $\log_{10}$ used in the plot. It means that there are four options for modeling developmental stage:

- non-parametric Mann-Whitney test applied to stages 9 and 10 separately;
- Anova model in (3.4) applied to stages 9 and 10 separately;
- Mann-Whitney test on pooled measurements from both stages, where weighted average is used for pooling so that both stages are equally represented;
- Anova model, including a stage cofactor, for the full dataset.

The first three options of association study models do not require much further explanation. Two of them rely on splitting phenotypes into two parts and applying the models of the previous section to these parts separately.

The fourth scenario would require testing the model with a new cofactor S_{ijk} denoting the variation within a particular stage of cells in image k:

$$y_{ijkh} = \mu + G_i + S_{ijk} + G_i \times S_{ijk} + L_{ij} + L_{ij} \times S_{ijk} + C_{ijk} + \varepsilon_{ijkh} \tag{3.8}$$

Notice that several interaction terms are included in the model as well. It seems quite natural to assume that genotype effect might be of a different nature in various stages. For example, localization of certain proteins takes place only in one of the stages, besides, cell growth and degradation happens in several waves, affected by different factors in each. Also the internal correlation within measurements coming from a single line can manifest itself differently for various developmental stages. For example, egg chamber images from stage 9 and stage 10 can be very different in terms of brightness.

In practice the last model was not used in the association study due to several factors:

- model fitting becomes much more complicated with two way Anova and unbalanced nested measurement levels;

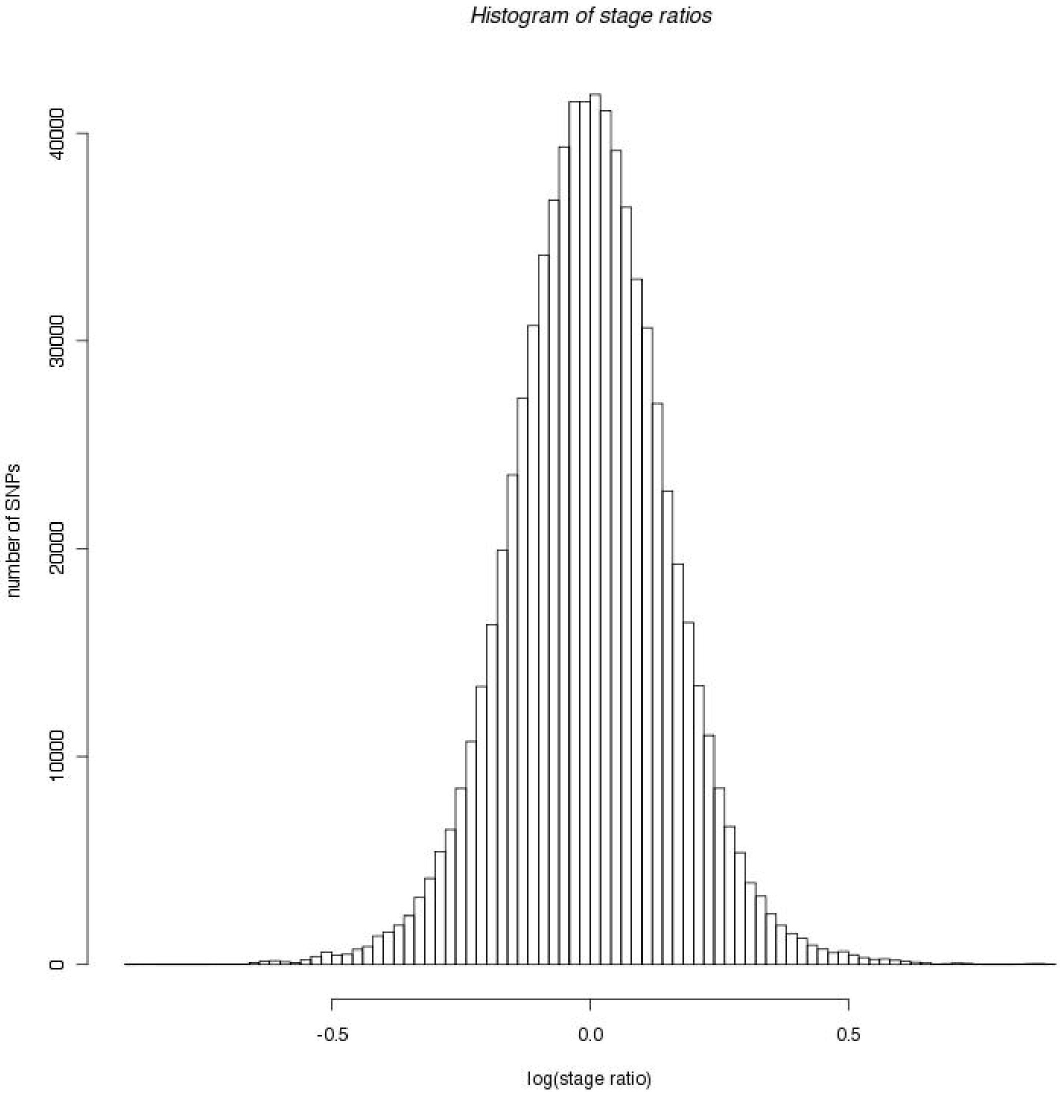


Figure 3.1: Histogram of developmental stage ratios for all SNPs, where ratio is defined as $n_{aa,9} \cdot n_{AA,10}/(n_{aa,10} \cdot n_{AA,9})$ and $n_{aa,9}$ is the number of measurements with allele aa and stage 9.

- the gain in the number of used measurements is only twofold if compared to separate analysis of both stages, while the number of parameters doubles as well.

Apart from developmental stage, there are no obvious biological conditions that would differ for various samples. It is easy to imagine some more technical factors, like staining, mounting etc., affecting the result of the association test. However, there is a reason to hope that those elements are captured in within line and within image variation or properly averaged. As it will be seen later from various quantile-quantile plots showing asymptotic behaviour of test statistics, this statement seems to hold.

3.4 Data Quality Control for *Drosophila melanogaster* SNPs

The genotypes of *Drosophila melanogaster* strains were taken from the Drosophila Population Genomics Project (DPGP) collection (Drosophila population genomics project, 2010). SNPs from release 1.0 were used in the current study together with the data quality filters initially applied to release 1.0 genomes.

In addition to the DPGP quality filters, control of two further aspects of data quality was undertaken, namely missing data rate and allele frequency. In both cases the main concern is to have enough measurements for both genotypes to be able to discern effects of alleles on phenotypes.

There are two main sources of missing data. Genome regions of a line that show identity by descent with some other line are omitted from the analysis of former line (converted into missing alleles). Also regions that show a significant amount of residual heterozygosity are excluded from analysis. Apart from longer gaps introduced by these eliminations, there is a smaller portion of alleles missing at random, mostly due to the errors in genotyping. The distribution of missing data rates among SNPs can be seen in Fig.3.2a. It turns out that there are no loci with no missing values. This can be explained by long stretches of identity by descent regions and heterozygous regions. However, most of the SNPs still have fewer than 5 missing allele values, corresponding to a missing data rate of 12%

or less. The choice of a threshold controlling the maximum missing data rate is a tradeoff between the number of SNPs available for analysis and the statistical power of the association test for a single SNP. After plotting the potential losses of SNPs under certain thresholds (see Fig.3.3), I decided to allow the missing data rate to be 10% or lower. This means that only SNPs with 3 or less missing alleles are included in the analysis.

Another important parameter that should be controlled is minor allele frequency (MAF). The traditional approach pursued in many association studies is to exclude SNPs with MAF $< 5\%$, as the rarity of the allele forbids reliable assessment of its association p-value. I chose to follow the same path, especially because the total number of fly lines is only 34. It means that SNPs with a single instance of the minor allele are discarded, but those having at least 2 lines with the minor allele are retained. The distribution of minor allele counts over various loci is shown in Fig.3.2b. Most of the loci have 0 or 1 minor allele instances; nevertheless after application of the 5% threshold there is still a sufficiently large number of SNPs retained, as seen in Fig.3.3.

One might ask what happens when both types of data quality control are combined. If SNPs having a large number of missing alleles also had a low MAF and vice versa, data quality control would not lead to significantly greater SNP losses than those described previously. After plotting the number of lost SNPs in two dimensions (missing rate versus MAF, Fig.3.3), it is possible to see that with the combination of 5% MAF and 10% missing thresholds 1,279,042 SNPs are retained for further analysis.

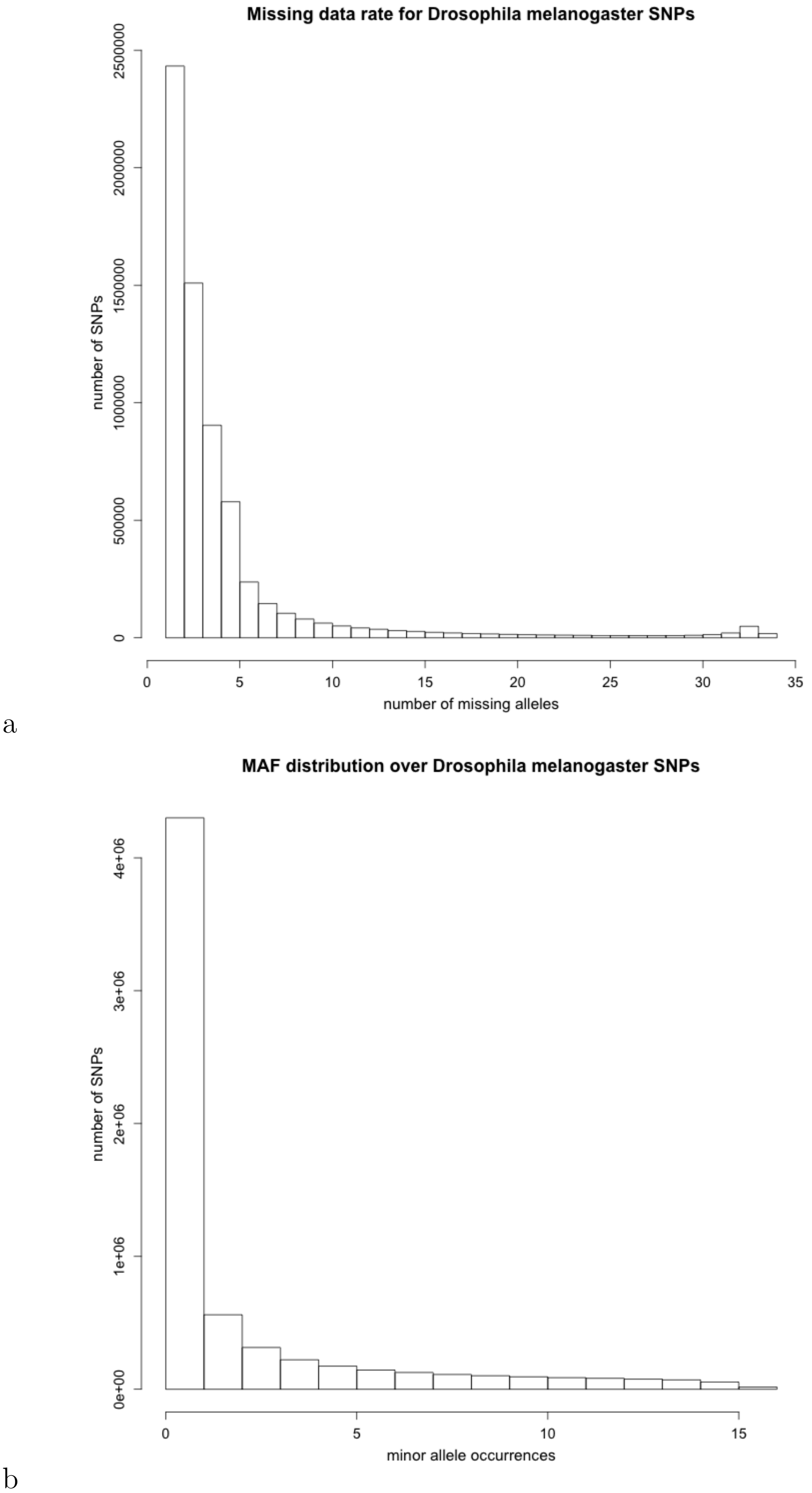


Figure 3.2: Histograms of MAF and missing data rate distributions over *Drosophila melanogaster* SNPs: a. missing alleles, b. minor allele occurrences.

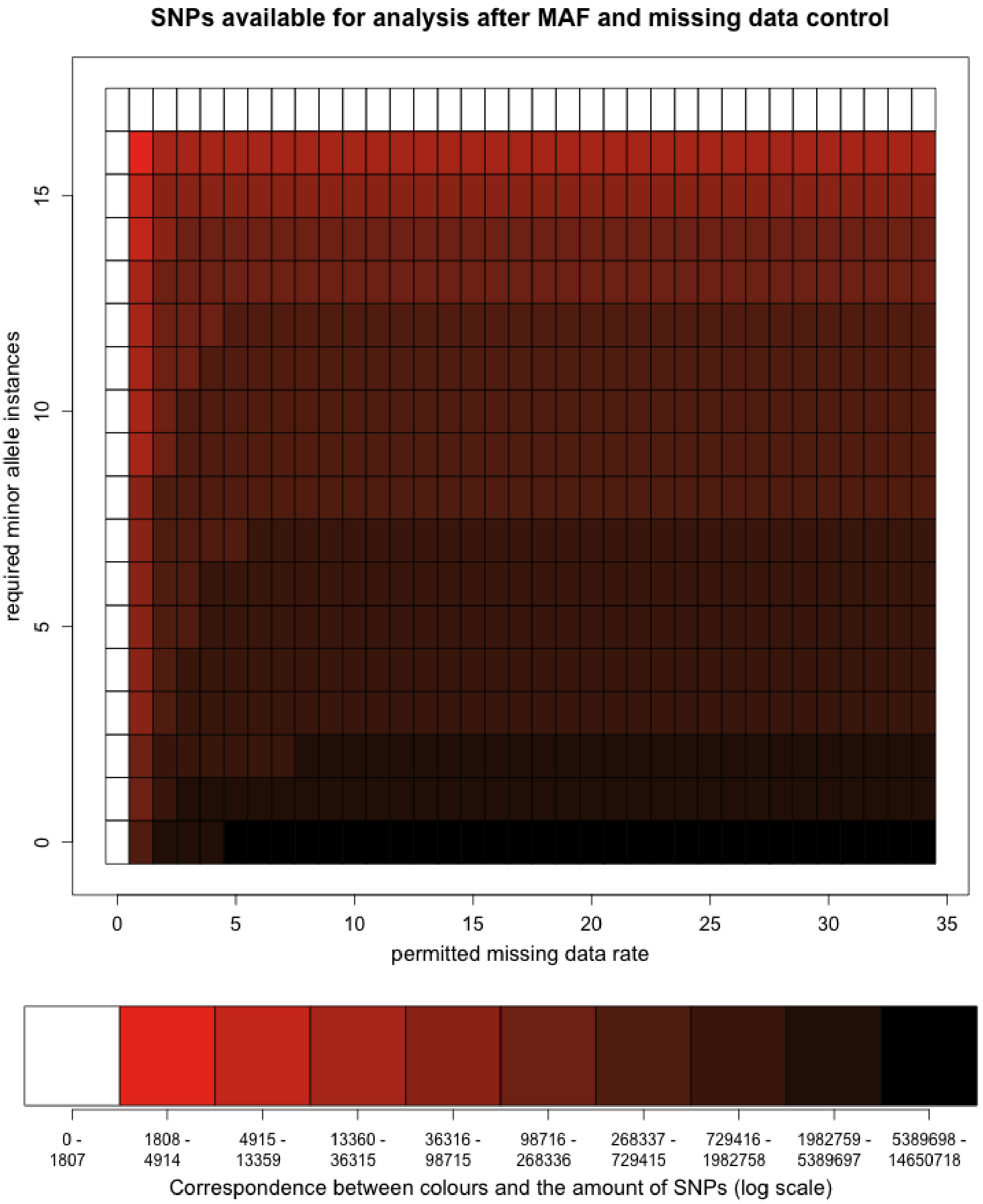


Figure 3.3: Impact of various thresholds set upon maximum missing data rate and minimum MAF. The darker the square the more SNPs are left for further analysis after thresholding (log scale used). White squares denote absence of SNPs for a particular threshold combination.

Chapter 4

Association Study for Oogenesis Markers in *Drosophila melanogaster*

4.1 Association Study Results

Here I am seeking to identify specific variants showing a significant association with 92 traits described in the Chapter 2. Two approaches were used to elucidate associations. The first approach was based on non-parametric tests, particularly the Mann-Whitney test. The second approach exploited nested Anova models and relied on asymptotic behaviour of test statistics corresponding to that of normal distribution. P-values for SNPs were analysed and significance thresholds for claiming an association were established by permutation tests. Crucial part of the analysis consisted of dissecting the locations of associated SNPs and searching for candidate genes.

4.1.1 Non-Parametric Approach to Association Discovery

First, I used the Mann-Whitney test to assess the strength of associations between genotype and phenotype. Before performing the test, the phenotypic measurements of cell nuclei were pooled within each image and then the image-wide measurements were pooled within each line. The measurements for various developmental stages were weighted in such a way that all stages have an equal impact on the aggregated measurement of each line. This way, the effect of the genotype can be assessed by testing whether the trait values of lines having a specific allele are significantly different. P-values for the Mann-Whitney test were calculated from the normal distribution that approximates the test statistic.

The minimal p-values obtained for each trait across all the SNPs tested are shown in Fig.4.1. It can be seen that all traits have their smallest p-value in the range from 10^{-4} to 10^{-6}. The largest values i.e., the weakest associations, are attributable to simpler geometrical cell nuclei characteristics, such as the area, perimeter, effective radius and the mean distance from center to perimeter. In contrast, more detailed descriptors of distance from center to perimeter, such as frequency components of Fourier transform of distance profile, have considerably better p-values than other geometric traits. Some of the traits that are related to, and probably correlated with the distance profile also have relatively small p-values. Examples of those include major and minor axis, eccentricity and the first two Hu's invariants of the geometric shape.

Cell texture descriptors exhibit smaller p-values than purely geometric traits. However, there are no distinct peaks indicating associations as p-values for Haralick features do not differ considerably from trait to trait.

The smallest p-values among all traits studied are obtained from Zernike moments with degree 8 or higher. There is a lot of variation among Zernike moment associations, with a marked increase in the degree of association towards higher order traits.

It is interesting to note that the total signal intensity yields a very weak association. This can be interpreted in favour of the hypothesis that technical artifacts do not significantly bias the analysis, because staining intensity and the focus of the microscope do not seem to produce false associations.

If I were to apply the Bonferroni correction to the p-values of non-parametric tests to claim an association, the p-values had to be as low as 7.282283×10^{-10} due to adjustment for 92 different traits and 746,302 different SNPs with type I error equal to 0.05. None of the traits gives SNPs with associations close to that threshold. To address this issue permutation tests were performed for 5

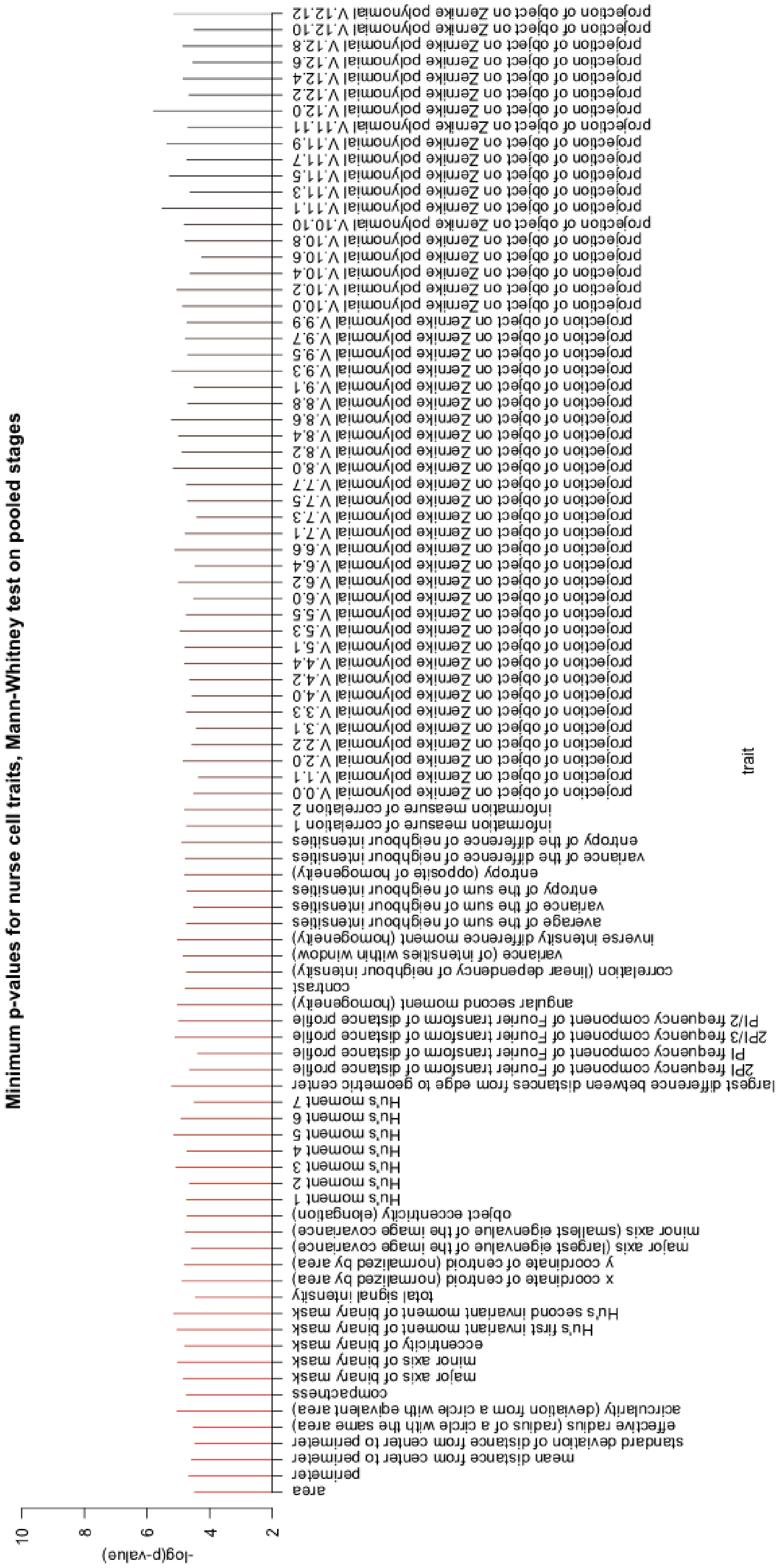


Figure 4.1: Minimum p-values for phenotypes, Mann-Whitney test

traits with relatively small p-values. However, none of the tests could confirm significant associations (results not shown).

To enquire into the reasons for the lack of associations, the quantiles of the Mann-Whitney tests for *Drosophila melanogaster* SNPs were compared to the theoretical quantiles of normal distribution. The quantile-quantile plots of 10 traits randomly taken from the set of texture descriptors, Hu's moments and Zernike moments are shown in Fig.4.2. There is an almost perfect compliance between the Mann-Whitney test statistic and the normal distribution quantiles. The closeness of the theoretical quantiles to those obtained by association tests is emphasized by the slopes of the lines going through first and third (25% and 75%) quartiles of the data. As the aspect of these lines does not deviate from 45 degrees, it is possible to conclude that normal distribution correctly approximates the behaviour of the Mann-Whitney tests performed. Therefore the results described above are reliable.

The tails of the empirical distribution of Mann-Whitney test statistic have a tendency to depart from normality. In all cases the direction of departure is towards a less significant Mann-Whitney test result. It means that the extreme ends of the test statistics yield larger p-values than would be expected from the corresponding normal quantiles. Significant associations would carry tails in the opposite direction, towards p-values that are smaller than those obtained from the quantiles of normal distribution. A likely explanation is the lack of power to discern associations due to the insufficient number of lines (34) and pooled measurements for each line. This is confirmed by a consistent nature of deviations from normality for different traits. Also it seems impossible to distinguish subtler differences within smaller p-values in the tails of the empirical distribution. This is implied by the minimal p-values for various traits as well as the tiny variance of minimal p-values across the traits.

Despite the shortcomings of association discovery using the Mann-Whitney test, this approach can still be useful as an unbiased way to assess the significance. For example, the Mann-Whitney test does not require the assumption of normality of phenotypic measurements. Therefore, clues obtained from Mann-Whitney p-values can be useful to validate results of more complicated statistical tests whose performance is dependent on satisfying more conditions.

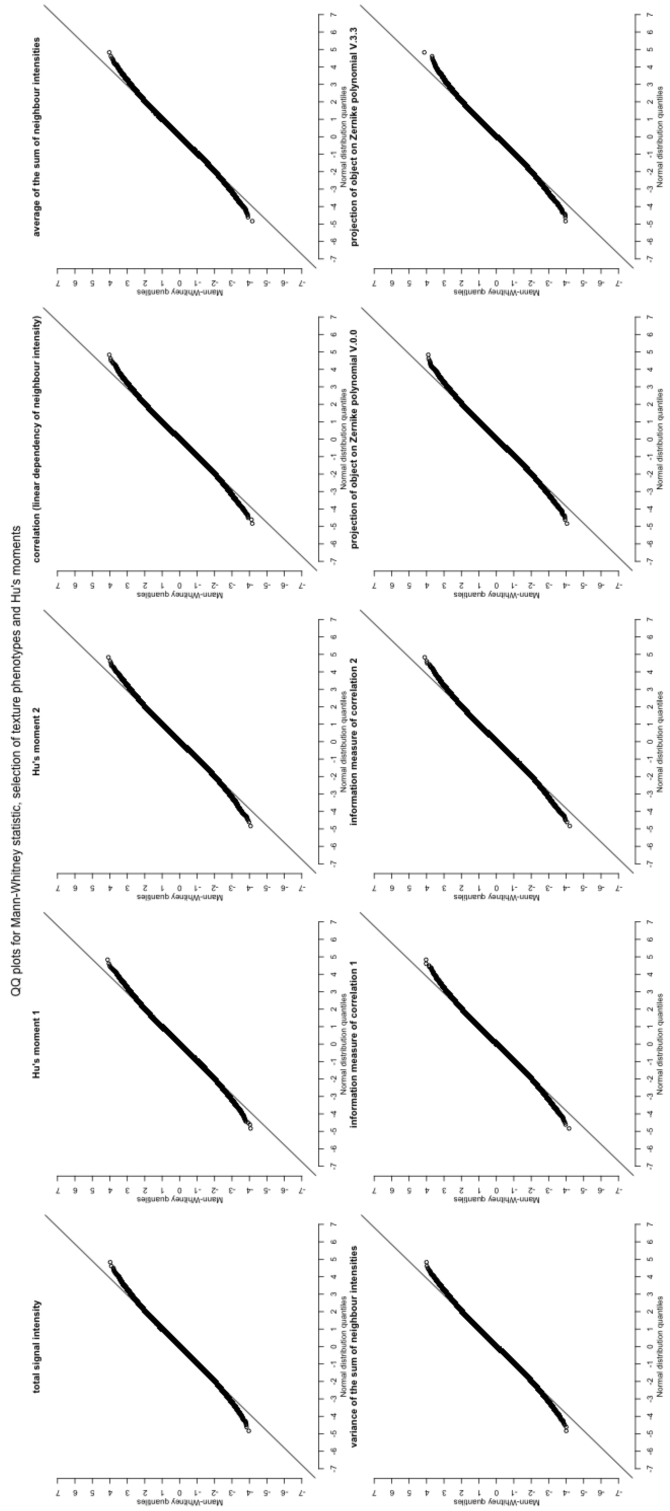


Figure 4.2: Quantile-quantile plots for Mann-Whitney statistic, selected texture traits, Hu's moments and Zernike moments.

4.1.2 A Parametric Approach to Association Discovery

As a more stringent approach than Mann-Whitney test, I performed tests for genotype effect based on nested Anova models. This approach does not require averaging of measurements within each line or image. On the other hand, since the Anova model is parametric, it relies on other assumptions, such as asymptotic normality of trait distribution.

Each developmental stage was analysed separately. The hypothesis of significant variance between different genotypes was tested. P-values were obtained from the test statistic using Sattherthwaite approximation when relevant conditions were satisfied. For the remaining SNPs the degree of association was assessed by a conventional Anova test for the ratio of sum of squares, in which variance among genotypes is compared to variance among lines.

Fig.4.3 and Fig.4.4 show the minimum p-values for various traits where the minimum is taken from all SNPs for a single trait. The p-values for stage 9 are presented in Fig.4.3 and stage 10 p-values are reported in Fig.4.4. The comparison of results between stages shows that p-values tend to be larger for stage 9 traits. Since the number of measurements is roughly equal for both stages, this is likely to reflect larger number of associations for stage 9 rather than systematic differences in the statistical power between analyses of these data sets. If we compare the peak landscapes to the Mann-Whitney test results, the Anova test p-values are generally higher. At the same time the lowest valleys correspond to p-values slightly lower than 10^{-4} for both tests. It seems that by using nested models better resolution for the smallest p-values is obtained, while larger p-values remain similar on both settings.

How do the results for the two developmental stages relate to each other? Notably, some of the most prominent trait peaks coincide for both stages. For example, the cluster of Hu's moments (from Hu's moment 3 to Hu's moment 7) exhibits the smallest p-values for both developmental stages. Also, the angular second moment (homogeneity) has a high p-value peak in both stages. Another common trend for the two stages is relatively low p-values for Zernike moments in comparison with other traits. However, in each stage there are several Zernike

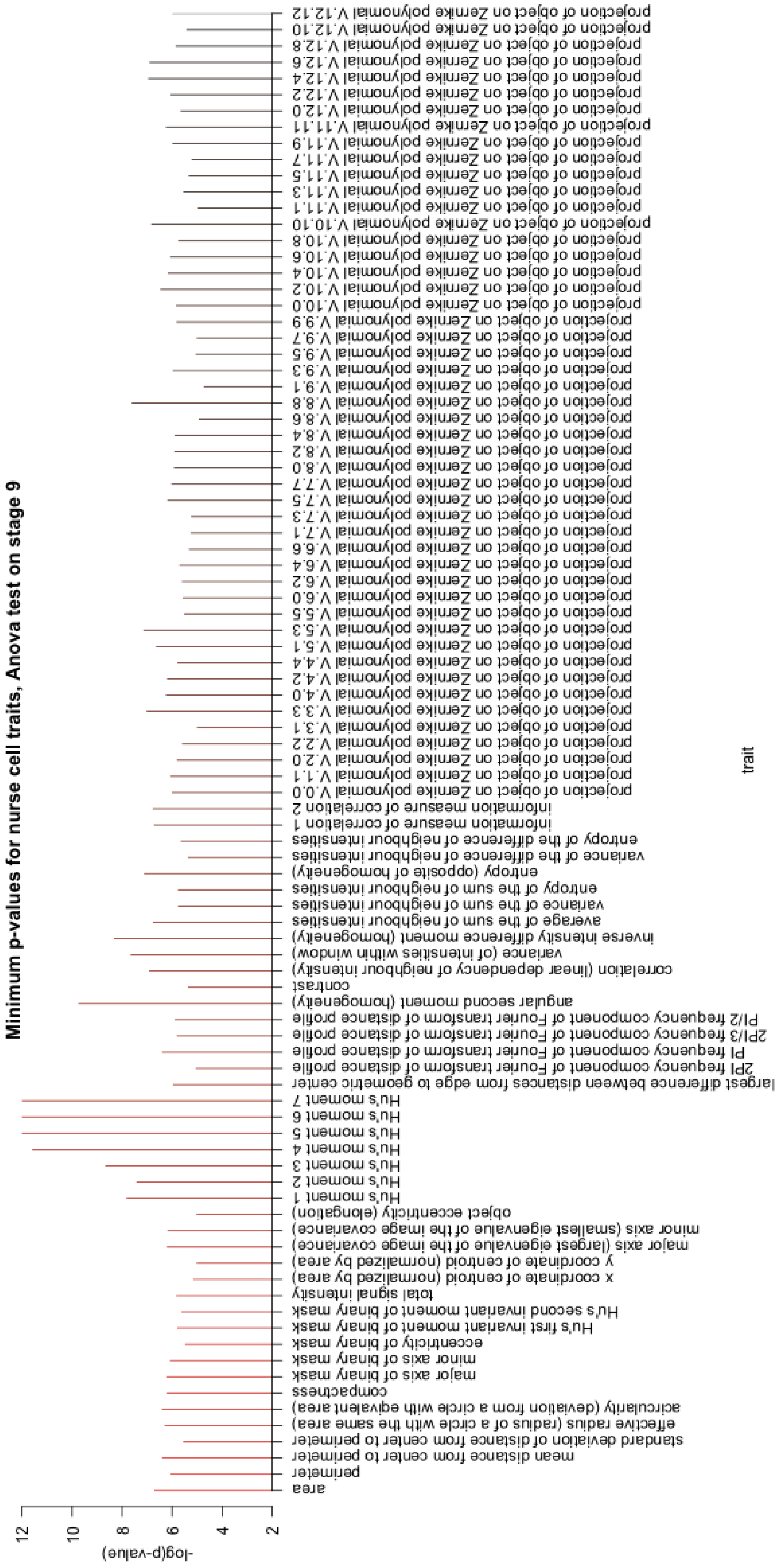


Figure 4.3: Minimum p-values for traits, Anova test, stage 9

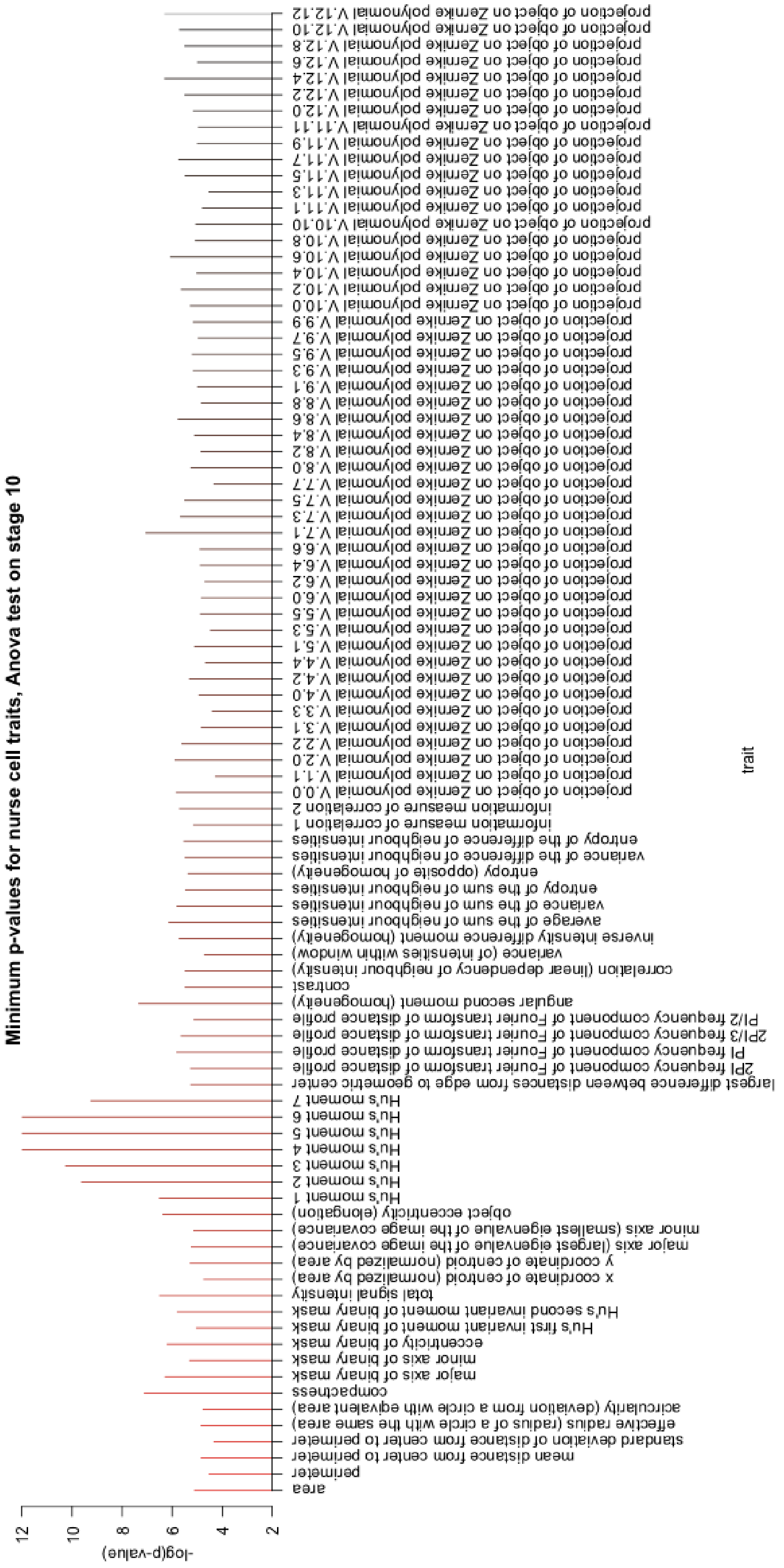


Figure 4.4: Minimum p-values for traits, Anova test, stage 10

moment peaks with small p-values that cannot be found in the other developmental stage.

Apart from a cluster of Hu's moments, other significant traits of stage 9 include a few Zernike moments (such as V.3.3, V.5.3, V.8.8, V.10.10, V.12.4, V.12.6), several Haralick features (such as correlation, variance, inverse intensity difference or homogeneity, average sum of neighbour intensities, entropy, information measure of correlation 1 and 2). Simple geometric traits, among them area, effective radius, acircularity and major and minor axis, yield p-values of medium significance. In general the most important p-values come from texture descriptors and some moment measures (Hu's moments).

Stage 10 has several important p-value peaks representing higher order Zernike moments as well, although they are not perfectly coinciding with those of stage 9 (V.7.1, V.8.6, V.10.2, V.10.6., V.11.7, V.12.4, V.12.12). Haralick features of stage 10 do not reach a degree of significance comparable to stage 9. There is a considerable variation across p-values of geometric features at stage 10. Some geometric traits are more significant in stage 10 than in stage 9 and rather prominent in the p-value landscape of stage 10 as well. These include compactness, major axis, eccentricity and Hu's second invariant of shape. The remaining geometrical descriptors of cell nuclei are among traits yielding the weakest associations in stage 10 (see area and perimeter, for example).

The p-values for a number of traits obtained by Anova model are consistent with the results of the Mann-Whitney test. For example, the same trend for higher order Zernike polynomials to be more significant than lower order ones can be observed in both tests. The geometrical traits (apart from several instances in stage 10) have relatively lower significance in both tests than other traits. However, it is difficult to make direct and detailed comparisons between results of different tests, because the variation among peak height is much more pronounced with more sophisticated models.

Nominally significant SNPs from Anova test after Bonferroni correction would be those with p-values lower than 3.64114×10^{-10}. For stage 9 there are 5 traits that match this criteria: Hu's moments 4 to 7 and angular second moment or homogeneity. For stage 10 very similar set of traits is obtained, i.e., Hu's moments 2 to 6. I referred to the results of Kruskal-Wallis test for these traits in Fig.2.7.

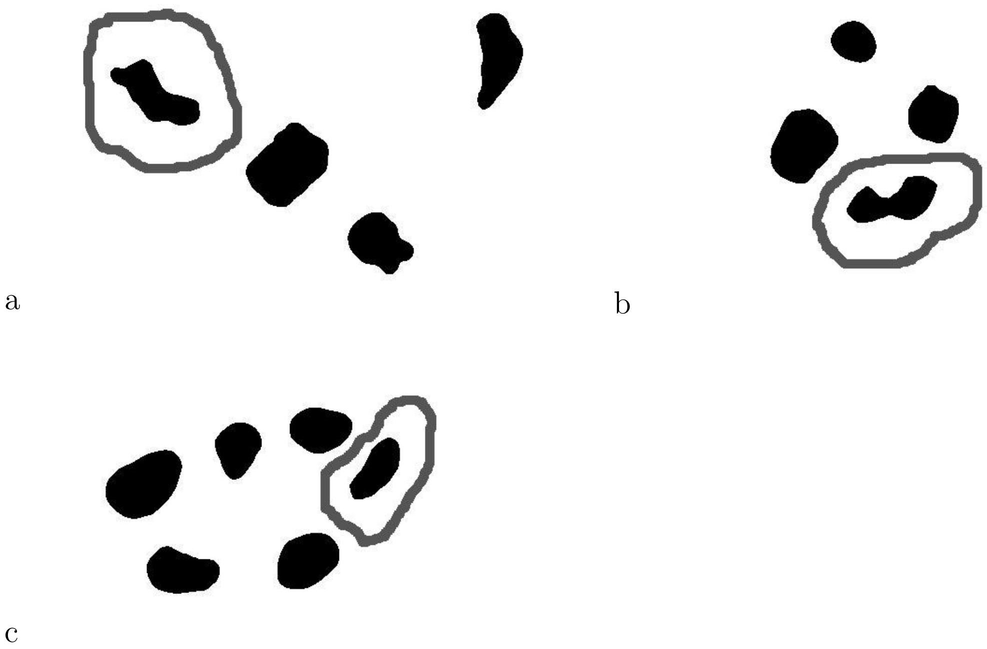


Figure 4.5: Masks of cell nuclei yielding extreme phenotypes for Hu's moments. The cell nucleus with unusual Hu's moment value is circled with gray.

It turns out that this set of traits has extremely insignificant Kruskal-Wallis test p-values in comparison to other traits. If we look at the trait distribution for different lines (see boxplots in Fig.2.7), it is possible to see that the variation of a trait within a given line is quite uneven in these cases. Besides, there is a relatively large number of far outliers. At the same time the line medians are rather similar. These obervations strongly suggest a violation of the distribution normality assumptions on which Anova test is based. First, the within-line variances are unequal, second, there are outliers that shift line means heavily, while leaving medians intact and third, Kruskal-Wallis test p-values indicate that even the weaker statement that at least two lines are different does not hold. Thus, a conclusion can be made that the most prominent association peaks are false positives.

To understand where the outliers for Hu's moments come from, I looked into the images and cell nuclei yielding extreme phenotypes for these traits. It turned out that all outliers for all Hu's moments are produced by 9 images only, where the unusual numerical value of a trait is associated with a single cell in each image. This particular cell affects the average phenotype of an image (egg chamber) to a great extent and thus a considerable deviation from the trait median for a fly line can be seen in Fig.2.7. The cell nuclei with extreme phenotypes have some characteristics in common: the difference between minor and major axis is large (typically major axis is three or four times larger than minor axis) and the compactness of the nucleus is extremely low. In several cases the nucleus' shape is not convex, but is concave instead. Three examples of such cells can be seen in Fig.4.5. It is rather difficult to detect this type of outliers during preliminary analysis, because their main distinction is the elongated and concave nucleus' shape, however, there are many elongated nuclei that do not have extreme phenotypes (the same can be said about many concave nuclei). Particularly interesting example can be seen in Fig.4.5, part c.: the nucleus having extreme phenotype is not very different from other nuclei within image. As the number of cells with outlying trait values is small, they will be excluded from the subsequent analysis incorporating more fly lines (see conclusions in Chapter 7).

I will now consider a set of traits with the p-values slightly lower than the nominal threshold of significance. It is possible to identify a few peaks for both stages with the minimum p-values between 10^{-8} and 3.64114×10^{-10}. For stage 10 these peaks correspond to compactness, Hu's moment 7, angular second moment and Zernike moment V.7.1. For stage 9 there are relatively significant p-values for Hu's moment 1, Hu's moment 2, Hu's moment 3, variance of intensities, inverse intensity difference moment, entropy of intensities and Zernike moments V.5.3 and V.8.8. These traits yield several candidate SNPs that have to be subjected to further biological and statistical validation. However, even in this set of traits there are some obvious false positives that can be spotted by the Kruskal-Wallis test. Such cases are Hu's moment 7 and angular second moment for stage 10 and Hu's moment 2, Hu's moment 3, the entropy and the inverse intensity difference moment for stage 9 (see Fig.2.7).

A map of candidate SNPs selected from both stages is depicted in Fig.4.6. In total 10 different association regions can be distinguished. The most pronounced associations are located on the X chromosome (the regions around 13.7Mb and 17.4Mb). The remaining candidate SNPs can be spotted mostly on both arms of chromosome 2, apart from two SNPs on chromosome 3R. The SNPs on chromosome 3R are the least significant in terms of p-values.

Table 4.1: Locations of candidate SNPs

Locus	P-value	Stage	Trait
13674734, X	1.5589×10^{-8}	9	Hu's moment 1
17433852, X	2.2917×10^{-8}	9	variance (of intensities within window)
12426632, 2L	2.5346×10^{-8}	9	Zernike moment V.8.8
4158962, 2R	7.3990×10^{-8}	9	Zernike moment V.5.3
6005161, 2L	7.6986×10^{-8}	10	compactness
6005175, 2L	7.6986×10^{-8}	10	compactness
25274820, 3R	8.8455×10^{-8}	10	Zernike moment V.7.1
12436406, 2L	2.1308×10^{-7}	9	Zernike moment V.8.8
12437435, 2L	2.1308×10^{-7}	9	Zernike moment V.8.8
12436407, 2L	2.1308×10^{-7}	9	Zernike moment V.8.8
12436397, 2L	2.1308×10^{-7}	9	Zernike moment V.8.8
12437449, 2L	2.1308×10^{-7}	9	Zernike moment V.8.8
5995489, 2L	2.3448×10^{-7}	10	compactness
9113181, 2L	4.5683×10^{-7}	9	Hu's moment 1
7448890, 2R	5.6820×10^{-7}	9	Hu's moment 1
7391809, 2R	7.8688×10^{-7}	10	compactness
17787083, 3R	7.9820×10^{-7}	10	compactness
12433348, 2L	8.1643×10^{-7}	9	Zernike moment V.8.8
12436591, 2L	8.1643×10^{-7}	9	Zernike moment V.8.8
12436567, 2L	8.1643×10^{-7}	9	Zernike moment V.8.8
12436872, 2L	8.1643×10^{-7}	9	Zernike moment V.8.8
20228295, 2L	9.8760×10^{-7}	9	Hu's moment 1

A more detailed analysis of association regions in table 4.1 shows that stage 9 traits contribute the majority of the candidate SNPs. In addition, p-values are stronger for stage 9 associations. All traits apart from compactness, which is a purely geometric feature, and the variance of signal intensities belong to various classes of moments. For instance, Hu's moment 1 is the rotational invariant of the overall intensity distribution. In this situation it might be related to the

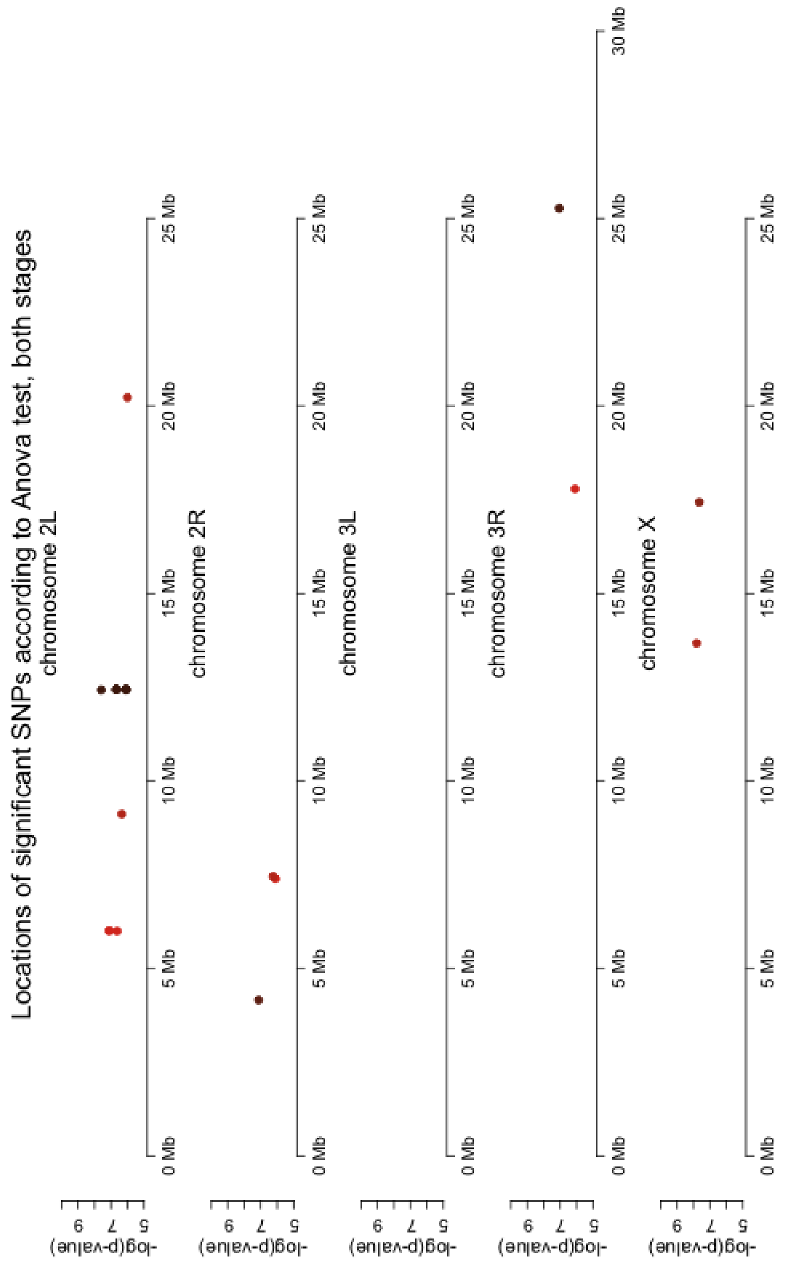


Figure 4.6: Map of candidate SNPs, pooled stages

distribution of DNA within nucleus. The compactness describes distribution of DNA in terms of nucleus' surface and volume relationship. The three Zernike moments possibly describe three different patterns of DNA distribution, while the variance of intensities is a local measure of consistency of the texture. It seems that all of the associated traits are more or less related to the distribution of the DNA within the nucleus, while the measures of shape, such as area, perimeter and other global characteristics, do not show significant associations. This is hardly surprising, if we consider that all of the associations are stage specific.

The statistical validation of candidate SNPs was carried out by permutation tests. Minimum p-values from 99 permutations for each trait were obtained. The minimum was taken over all SNPs for a single permutation and a single trait and the magnitude of the real p-values relative to the permuted ones was determined. If only one trait was tested, any SNPs with p-values smaller than 95% of permuted p-values could be claimed significant. However, as 6 traits are validated, more stringent significance thresholds have to be adopted due to multiple testing.

The permutations themselves were constructed by swapping the line labels while leaving image and cell labels within line intact. This approach effectively shuffles genotype labels for traits while preserving the internal correlation that is present within nuclei from one oocyte (image) and images from one sample (line). This approach was pursued, because ignoring the nested structure of measurements can lead to systematically larger p-values on the permutation side compared to the real p-values. The same set of line label permutations was used for all traits, because the correlation among traits has to be preserved.

The results of permutation tests in Fig.4.7 show that 5 out of 6 traits have at least one significant SNP. Only Hu's moment did not yield any strong associations. This result is consistent with relatively large p-value of the Kruskal-Wallis test for Hu's moment (yet another proof of the predictive power of non-parametric methods for association discovery). Other traits, apart from Hu's moment, have one SNP each that falls among top 5 permutations. The strongest association is exhibited by Zernike moment V.8.8., where the real trait gives the best p-value among 100 minimum p-values. Some more candidate SNPs apart from the best SNP for a trait were tested for compactness and Zernike moment V.8.8. The

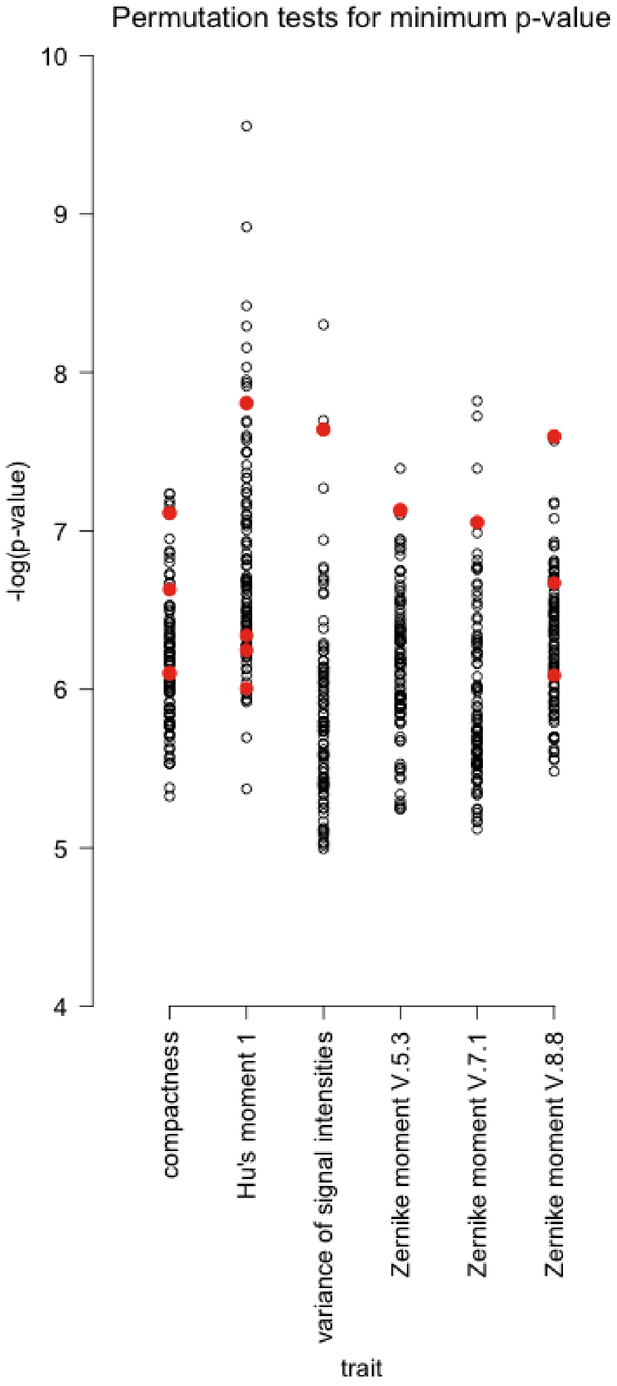


Figure 4.7: Permutation tests for selected traits, pooled stages. The association p-values for real traits are marked in red, while minimal p-values from permuted data sets are black.

second best SNP for compactness is located among the top 10 results of permutation tests. However, in the light of the multiple traits tested, this does not seem to be an important association result. The remaining candidates for compactness and Zernike moment V.8.8. are clearly not significant. SNPs validated by permutation tests are summarized in Table 4.2.

Table 4.2: Locations of validated SNPs

Locus	P-value	Stage	Trait
17433852, X	2.2917×10^{-8}	9	variance of signal intensities
12426632, 2L	2.5346×10^{-8}	9	Zernike moment V.8.8
4158962, 2R	7.3990×10^{-8}	9	Zernike moment V.5.3
6005161, 2L	7.6986×10^{-8}	10	compactness
6005175, 2L	7.6986×10^{-8}	10	compactness
25274820, 3R	8.8455×10^{-8}	10	Zernike moment V.7.1

The locations of validated associations are shown in Fig.4.8. There are five distinct association regions in total, because two SNPs affecting nuclei compactness are located nearby and are in perfect linkage disequilibrium. Three of five association regions are located on chromosome 2. These regions are related to nuclei compactness and Zernike moments. The strongest association comes from SNP on chromosome X affecting texture measured by the variance of signal intensities. A single association on chromosome 3R is related to Zernike moment V.7.1.

To assess the magnitude of allele effect upon a trait for all validated SNPs I plotted the trait distributions for lines split by genotypes (see Fig.4.9). Only the traits showing significant association at a given developmental stage are included in Fig.4.9. Measurements for nuclei coming from the same image (egg chamber) are averaged. Aggregated measurements for all *Drosophila melanogaster* lines are depicted in Fig.4.9. The difference between trait medians between genotype groups should reflect the genotype effect. In all cases medians are considerably different, thus confirming that association p-values were not due to outliers or other deviations from normality. However, note that variances are not equal for all lines. This is somewhat expected, because the maximum number of measurements

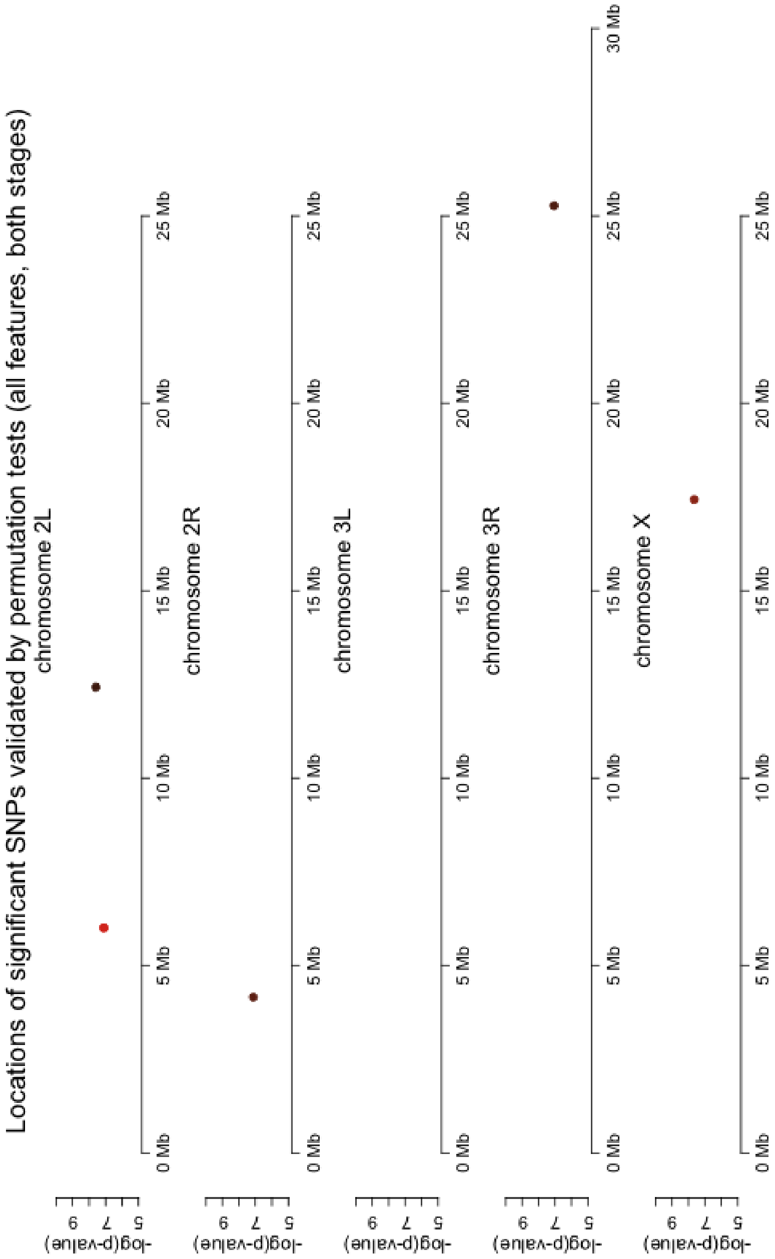


Figure 4.8: Map of validated SNPs, pooled stages

for a single line at a given developmental stage is less than 10 after pooling all cells per image. Also the experimental design is not fully balanced, therefore the number of images per each line can vary slightly. This problem is partially alleviated by combining lines corresponding to the same genotype as can be seen in Fig.4.10. The figure shows that the variation of the trait is relatively similar for different genotypes. However, care should be taken in interpreting the latter figure, because internal correlation is likely to be present within measurements coming from the same line.

4.2 Discussion

Let us look at the validated association regions in a broader context. First, it is interesting to ask which candidate genes are located in the vicinity of discovered associations. Table 4.3 summarizes the information on genes in the vicinity of significant SNPs. It turns out that the most significant association (variance of signal intensities) yields a SNP (chromosome X, 17.4Mb) inside the intronic region of CG12432 gene with unknown molecular function according to the information from modENCODE (modENCODE; The modENCODE Consortium et al., 2010). A narrower subregion of this intron corresponds to a binding site, which includes the associated locus as well.

The next strongest association corresponding to Zernike moment V.8.8. maps within elf-RA gene. The associated locus hits a protein binding site incorporated within an intron. Another gene belonging to the elf-RA family and located close to the SNP associated with Zernike moment V.8.8. is *arrest*. It is situated between base 12,205,843 and 12,312,795 on chromosome 2L, while the SNP itself is located at 12,426,632. *arrest* is coding for a transcription factor involved in oogenesis, negative regulation of *oskar* mRNA translation, germ cell development and negative regulation of translation among other functions mentioned in FlyBase (FlyBase; Tweedie et al., 2008). Several alleles of this gene affect phenotypes of female germline cyst, egg and germarium (FlyBase; Tweedie et al., 2008).

The region associated with Zernike moment V.5.3. is located within a gene desert. However, well studied developmental genes localise in a relative proximity to it (*spaw*, *hubl*, *whip*, *cola*). Fig.4.11 shows precise positions of these genes and

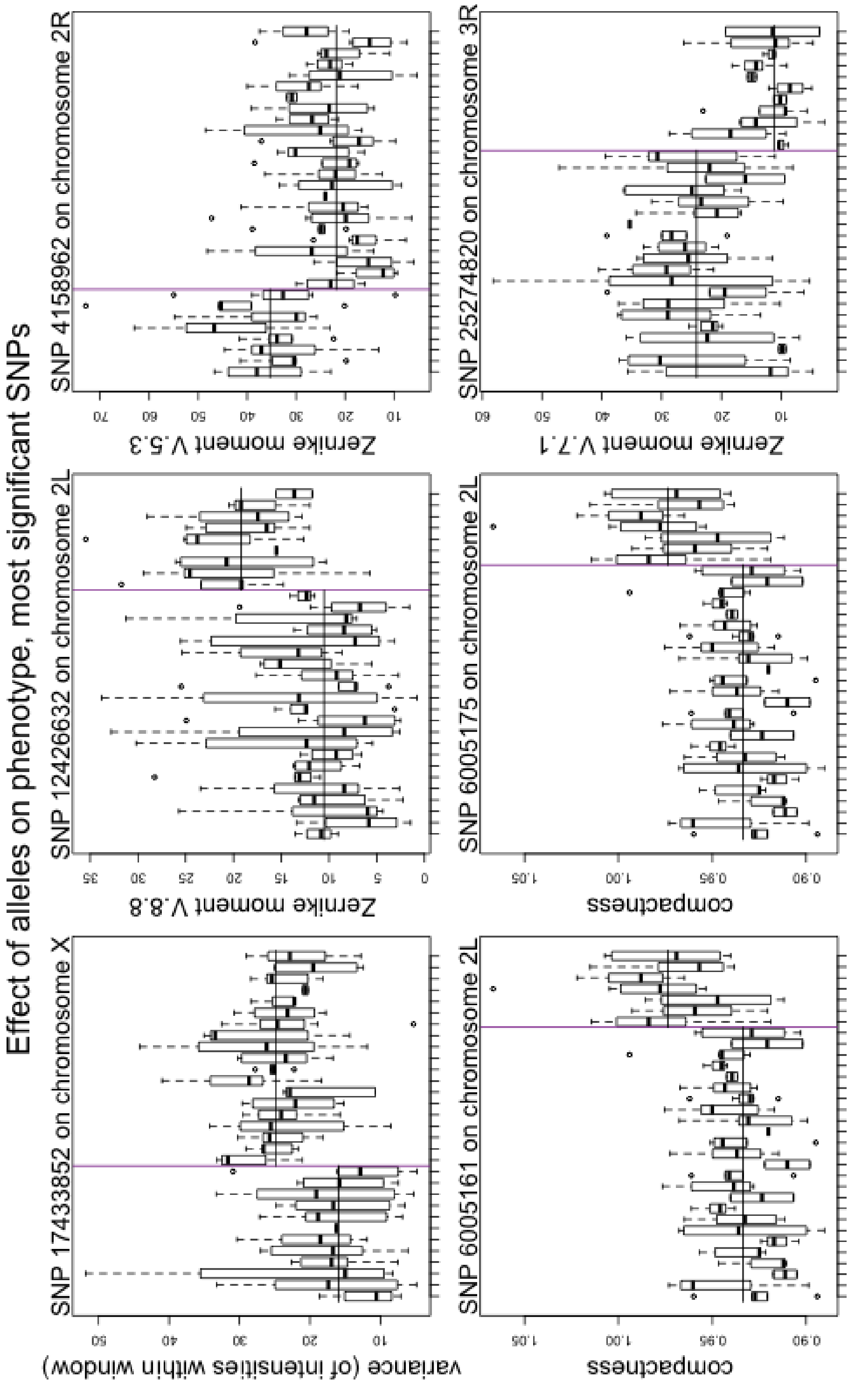


Figure 4.9: Trait separation due to the genotypes of the most significantly associated SNPs. Vertical line indicates the split between traits that correspond to genotype *aa* and genotype *AA*. Two horizontal lines denote trait medians in both genotype classes.

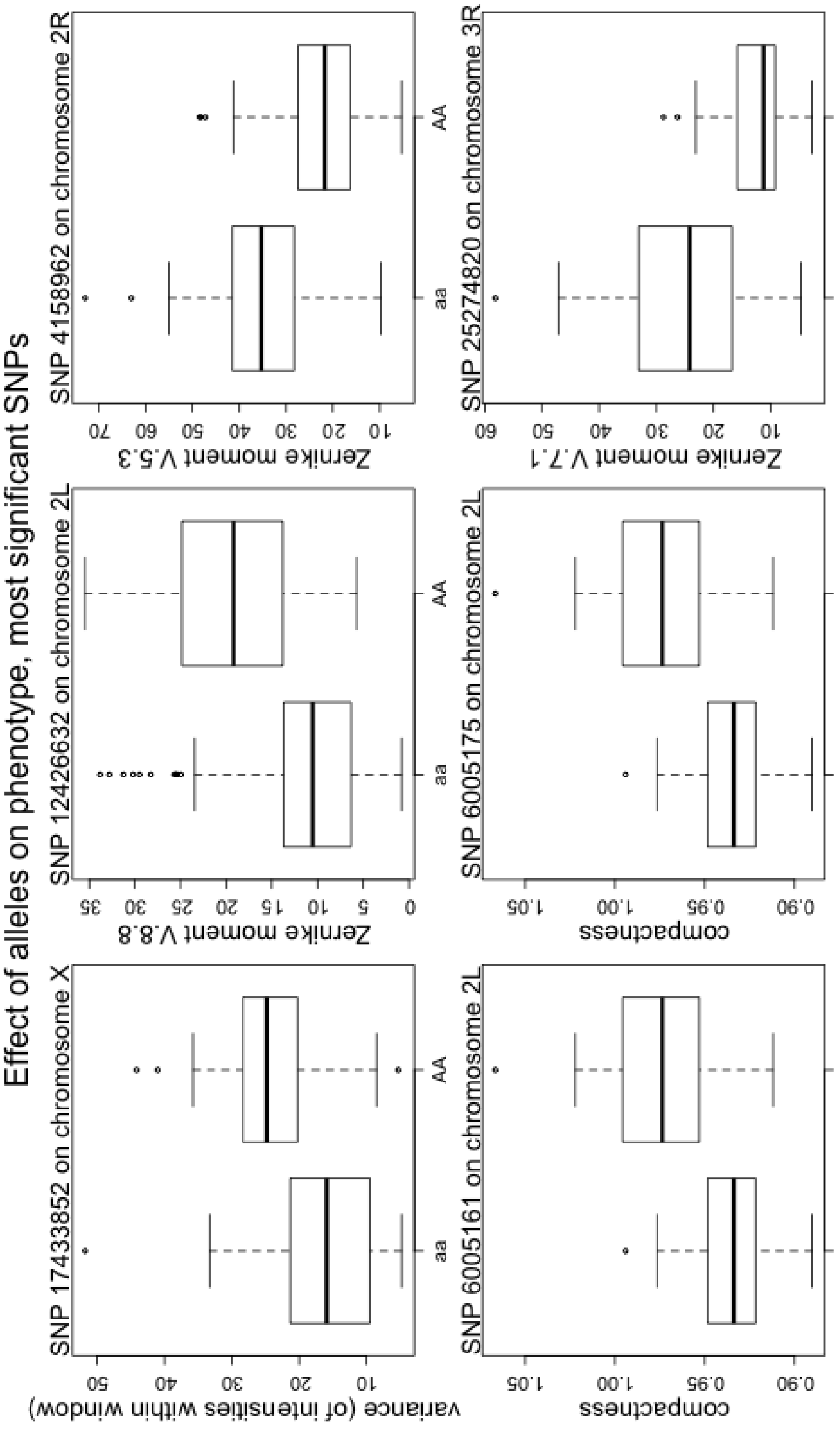


Figure 4.10: Trait separation by genotype for validated SNPs.

also their distance to the associated variant. This locus is interesting from a gene regulatory point of view, because it maps to a protein binding site. The regulatory activity has been detected in all embryonic stages of development within yellow cinnabar brown speck strain of fruit flies. Also experiments with other types of embryonic cell lines confirm the localization of this binding site. This site is activated in dorsal mesothoracic disc tissue (larval stage) and some adult tissues as well (modENCODE; The modENCODE Consortium et al., 2010).

The best association with cell nuclei compactness is found in the intronic region (also a binding site) of the Gal-RA gene coding β-galactosidase which is crucial for carbohydrate metabolism. It is interesting to note that β-galactosidase is used as a senescence marker, because it is expressed in lysosomes.

SNP yielded by Zernike moment V.7.1 is located within the intron of the Ptp99A gene that codes for a thyrosine phosphatase involved in cell-cell interactions. Tyrosine phosphorylation plays an important role in the regulation of various developmental processes, especially during oogenesis and early embryogenesis in *Drosophila melanogaster* (Fitzpatrick et al., 1995). Ptp99A RNA is present in germ cells of the germarium and also in nurse cells throughout oogenesis. The Ptp99A transcript is one of the very few phosphatase transcripts that can be detected in follicle cells. Ptp99A transcript first emerges in follicle cells during stage 9 and then reaches large quantity at stage 10. This type of localization indicates a significant role of Ptp99A in cell-cell signaling. It is interesting to note that homozygous null mutation for Ptp99A produces viable and fertile flies (Hamilton et al., 1995). The discovered association within the intron of this gene might be related to the regulation of the expression level that in turn causes subtle phenotypic changes, but does not lead to severe mutations.

Even candidate SNPs that did not pass the permutation tests hit genes relevant for early developmental processes. For example, the second best association for cell nuclei compactness falling within top 10% of permutation p-values is located in the lid-RA gene coding for a histone transferase (SNP on chromosome 2L, 6.0Mb).

Both of the least significant SNPs for the compactness yield interesting locations on genome as well. The SNP on chromosome 2R, 7.4Mb is situated in the intron of *inv* gene, which is a well known developmental gene. According to

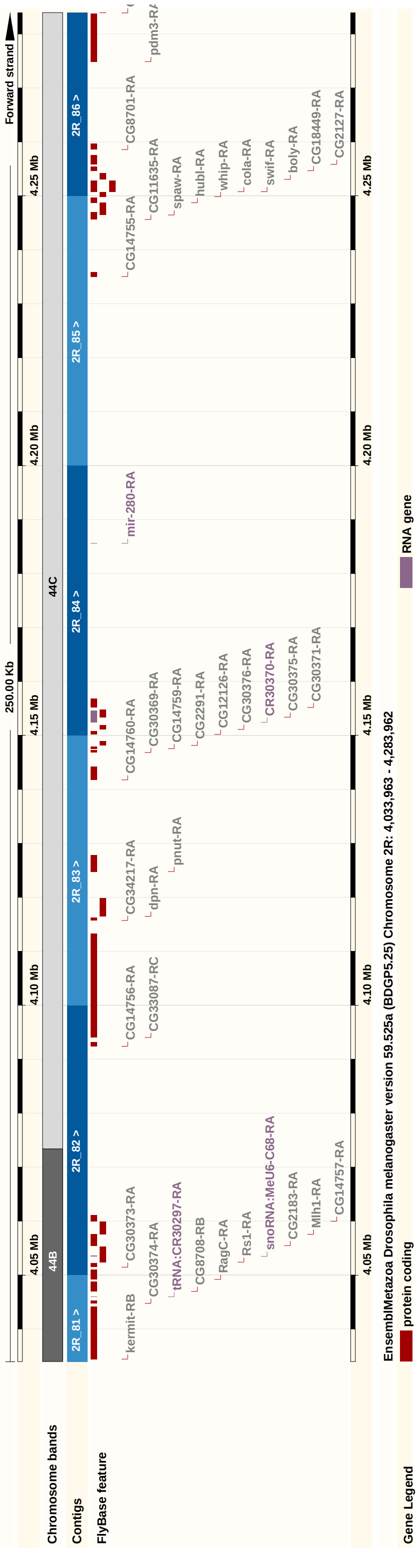


Figure 4.11: Screenshot from EnsEMBL database showing the location of SNP associated with Zernike moment V.5.3.

the FlyBase (FlyBase; Tweedie et al., 2008) this gene has a functional role in neuroblast fate determination and compartment pattern formation. Various *inv* alleles can influence the outcome of segmentation, in particular mesothoracic and thoracic segment, and organ system subdivision (FlyBase; Tweedie et al., 2008). This gene is also related to the embryonic development, for example, to the development of embryonic nervous system. The SNP on chromosome 3R, 17.8Mb is located within the gene Eip93F coding for a transcription factor involved in apoptosis. Other functional roles of this gene involve cellular response to hypoxia, autophagic cell death and autophagy (FlyBase; Tweedie et al., 2008). It is interesting to note that associations with compactness were discovered for stage 10 measurements instead of stage 9, because nurse cells start to degrade in late stage 10. This aspect helps to understand why a gene coding transcription factor necessary for apoptosis might be relevant. A follow-up study with an increased statistical power to detect associations will be necessary to verify these locations.

Weaker and less significant associations for Zernike moment V.8.8. are clustered on chromosome 2L, around 12.4Mb. All of them map to the region of the strongest association with the same trait, suggesting that these variants are connected with elf-RA as well.

Table 4.3: Candidate genes for validated SNPs

Locus	Trait	Candidate genes	SNP type
17433852, X	variance of signal intensities	CG12432	downstream, intronic
12426632, 2L	Zernike moment V.8.8	elf-RA	both downstream and upstream
4158962, 2R	Zernike moment V.5.3	gene desert, *spaw*, *hubl*, *whip*, *cola* in proximity	both downstream and upstream
6005161, 2L	compactness	Gal-RA	downstream, intronic
6005175, 2L	compactness	Gal-RA	downstream, intronic
25274820, 3R	Zernike moment V.7.1	Ptp99A	intronic

More associations on chromosome X were expected, because the sex linked character of many QTLs in *Drosophila melanogaster* is well known.

Most of the validated association loci are located within introns. Meanwhile none of the validated SNPs belong to a coding region. One of the explanations for this phenomenon might be that the true causative variant is not located within intron, but in the nearby coding region instead. However, this scenario is unlikely, because the linkage disequilibrium for *Drosophila melanogaster* is rather short and sequence data is used in association study instead of markers. Another explanation might be related to the fact that there are a lot of regulatory sequences and enhancers located in introns. Also alternative splicing is frequently determined by introns. Thus the variants discovered might participate in regulating the level of gene expression. This hypothesis is confirmed by the information from modENCODE (modENCODE; The modENCODE Consortium et al., 2010) indicating that most of the associated SNPs belong to a binding site and Table 4.3, where most of the SNPs are marked as upstream or downstream. In such a case phenotypic differences associated with various mutations at these loci should be subtler. The collected microscope images indicate that traits vary rather smoothly and continuously consistent with the nature of the associated variants.

The overall impression from association studies for developmental traits of *Drosophila melanogaster* is that metrics of cell nuclei that are more robust towards noise yield more significant associations. For example, various types of moments are well represented among candidate SNPs. Also the final list of validated SNPs contains mostly moments. Apparently, the robustness of Zernike moments against small shape perturbations has been instrumental in capturing essential characteristics of cells (Murphy et al., 2000; Xia et al., 2007).

A major limitation of the association study presented here is the insufficient statistical power due to only 40 fly lines being resequenced at first. Recently an additional 100 resequenced lines have become available, therefore I will be extending the study to more individuals (see Conclusion). The amount of lines precludes us from using robust non-parametric tests for association discovery, because they cannot exploit all of the nested levels of measurements. This is rather unfortunate, because in general non-parametric tests have proved to yield very accurate predictions, even when dealing with skewed and unbalanced trait distributions. For example, the results of Kruskal-Wallis test were crucial for eliminating false positive results. However, even when using more sophisticated models, such as nested Anova tests, we could wish for more power to discern the smallest p-values. This problem was partially solved by using permutation tests to establish thresholds for association significance, but an increased sample size would still be beneficial.

Chapter 5

Models and Model Selection for Association Study in Human

5.1 Separate Versus Joint Estimation of Allele Effects

A disease status can be modeled in several different ways, for example, it can be expressed via equation of a generalized linear model or captured in a decision tree. Both types of disease models can be subjected to some sort of model selection methods, which are usually necessary to use, if the joint estimation of allele effects of several loci is desired. Here I shall explore opportunities of using generalized linear models together with the stepwise model selection and decision trees together with the probabilistic model selection in association discovery.

One of the most typical approaches for modeling factors contributing to disease exploits the generalized linear model:

$$g(y_i) = \mu + \beta_1(2 \cdot I(x_i = AA) + I(x_i = Aa)) + \beta_2(1 - I(x_i = aa)) + \varepsilon_i \quad (5.1)$$

where y_i is the value of a phenotype for an individual i, x_i is the genotype of the locus under study for an individual i and ε_i is random variation. $I(\cdot)$ stands for an indicator function yielding 1 if the condition in the brackets is satisfied (in the current case 1 is returned, if individual i has that particular genotype). The formula $2 \cdot I(x_i = AA) + I(x_i = Aa)$ models the additive effect of an allele A, while

$1 - I(x_i = aa)$ models the dominance effect. Then β_1 is the magnitude of the additive effect, β_2 is the size of the dominance effect and μ is the phenotype mean for individuals with allele combination aa. As the trait is discrete, the phenotype value is transformed via a monotone link function $g(\cdot)$ yielding a probability of individual having a disease. The logit function is used as the link in the current study. Sometimes only the additive effect of alleles is modeled, i.e. β_2 is set to 0; then the whole model is called the additive model. When both additive and dominance effects are included in the model it is called the genotypic model.

The model proposed in (5.1) is suitable for testing the effect of each SNP upon phenotype separately from effects of other SNPs. As discussed in (Jansen, 1993) this is not ideal way to assess the effect of particular allele upon phenotype, because some SNPs can borrow the effect from linked loci and appear more significant than they actually are when a single locus model is tested. If the genotype effects of several SNPs are modeled jointly, the estimation of their total impact is more correct. Besides I would like to distinguish between situations when a SNP is genuinely associated with a phenotype and when it is just in linkage disequilibrium with a causative variant. Both the task of joint effect estimation and the task of identification of a true causative variant within the association region can be solved by various model selection methods, where the set of associated loci is gradually incremented and sometimes decremented until arrival at the final estimation of genotype effects upon phenotype. The result of performing such a model selection would look like this:

$$g(y_i) = \mu + \sum_{j=1}^{p} (\beta_{j1}(2 \cdot I(x_{ij} = AA) + I(x_{ij} = Aa)) + \beta_{j2}(1 - I(x_{ij} = aa))) + \varepsilon_i \quad (5.2)$$

where x_{ij} stands for the allele of loci j and individual i. The most complicated task here is selecting the set of interesting loci or indices j.

An alternative approach to modeling genetic components of disease can be based on decision trees. A tree describing the impact of a single locus upon phenotype is depicted in Fig.5.1. Each node of the tree corresponds to a classification outcome, which is either case or control (”healthy” or ”disease” in the picture). The classification outcome is determined by the majority vote of training set measurements within each node. Each edge in the tree is labeled by a classification

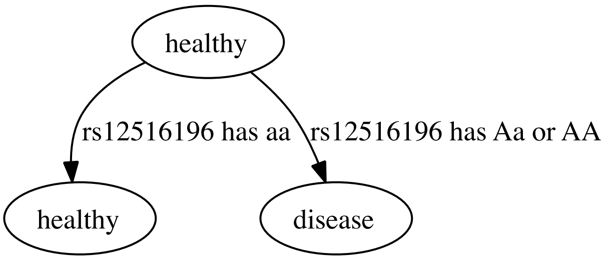


Figure 5.1: A decision tree of depth 1 incorporating a single locus.

condition. A condition is a single inequality that involves one predictor variable and one constant. A finite set of equalities with constants is also admissible as a classification condition, at least if the predictor variable is discrete and can assume a finite number of values, because it is easy to reorder constants in such a way that the condition can be expressed via a single inequality. In case/control studies the predictor variable corresponds to a locus and the constant is a particular combination of alleles. Then the role of the inequality is to separate one allele combination from others.

Joint effects of alleles can be modeled by using bigger and more elaborate trees, like shown by the example tree in Fig.5.2. I will call a depth of tree the maximum number of loci that participate in classification of a single measurement (individual). Thus, the tree is Fig.5.2 has depth 3, while the tree in Fig.5.1 has depth 1. If similar classification was made with generalized linear models, the number of coefficients fitted would correspond to the depth of a tree plus one (due to the intercept). However, the comparison of both modeling approaches is not entirely correct, because trees have many classification paths, while generalized linear models have only one fixed set of terms. Therefore trees are more flexible than generalized linear models when interactions between loci have to be captured and joint effects of alleles have to be estimated.

Trees do not have direct correspondence to the model in (5.1), because they cannot immediately express additive effect. However that is a problem for any

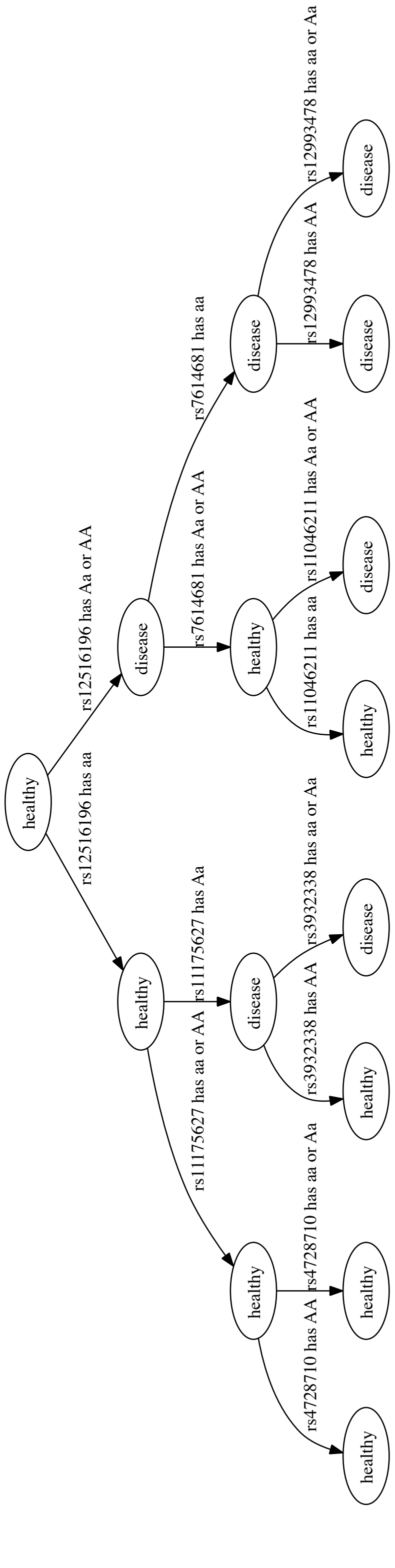


Figure 5.2: A decision tree of depth 3 incorporating multiple loci.

tree-like structure in statistical modeling and could not be solved by choosing a different tree structure. A major advantage of trees is that they do not require to assume any association model beforehand and thus are more flexible than regression models. In addition, trees are more intuitive for the interpretation, especially when many loci are included in the model. In contrast to the additive models trees can easily deal with the problem of incomplete penetrance usually met in association studies. Therefore using decision trees for modeling a disease alongside additive and genotypic models seems to be a natural choice. Additional advantage of trees is that they can deal with genetic heterogeneity and subpopulations more successfully than regression models similar to (5.2) (Lunetta et al., 2004).

5.2 Assessment of Significance of Allele Effects

The significance of the additive or dominance effect within model (5.1) can be tested by referring to asymptotic distribution of logarithm of likelihood ratios. A set of measurements is denoted by $X_1, \dots, X_n$ (in the current case these are genotypes of individuals) and the density from which they are taken is denoted by p_θ, where θ includes all of the parameters necessary to characterize the corresponding distribution. If a null hypothesis $H_0 : \theta \in \Theta_0$ is tested versus an alternative hypothesis that $H_1 : \theta \in \Theta_1$, we can use the following statistic based on the logarithm of likelihood ratio:

$$\Lambda_n = 2 \log \frac{\sup_{\theta \in \Theta} \Pi_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \Pi_{i=1}^n p_\theta(X_i)} \tag{5.3}$$

where $\Theta = \Theta_0 \cup \Theta_1$. This statistic is distributed as the chi-square distribution with degrees of freedom determined by the number of restricted parameters within θ under null hypothesis. For example, if a significance of the dominance effect has to be assessed, $H_0 : \beta_2 = 0$ is tested versus $H_1 : \beta_2 \neq 0$ by using the equation:

$$\Lambda_n = 2 \log \frac{\sup_{\beta_1,\beta_2} \Pi_{i=1}^n p_{\beta_1,\beta_2}(X_i)}{\sup_{\beta_1,\beta_2=0} \Pi_{i=1}^n p_{\beta_1,\beta_2}(X_i)} = -2 \sum_{i=1}^n (l_{\hat{\beta}_1,0}(X_i) - l_{\hat{\beta}_1,\hat{\beta}_2}(X_i)) \tag{5.4}$$

where l_θ denotes $\log p_\theta$ and $\hat{\beta}_1, \hat{\beta}_2$ refer to maximum likelihood estimators for β_1 and β_2. Then the likelihood ratio statistic is distributed as chi-square with one degree of freedom.

The simplest way how to measure the importance of a SNP in a classification tree is to count the occurrences of this locus within a tree. This method is too simple to give useful information about a SNP if a single small tree is considered, however, it becomes much more powerful when an ensemble of hundreds of trees is studied, because then each SNP gets much more chance to be included in a classification tree and the number of SNP occurrences exhibits more variation. Using the number of occurrences in evaluating the strength of the association has been proposed in (Valdar et al., 2009). The occurrence count of a locus within a set of models is considered to be indicative of association, where different models are built with distinct subsamples of the original data set using forward selection (Valdar et al., 2009). A set of trees is used instead of a set of models similar to (5.2) here.

Evaluation of trees and the significance of loci within them can be based on the Gini index (Friedman, 2001). First, I will define the probability measure that observations belonging to a certain node of a tree fall within outcome class k:

$$\hat{p}_k = \frac{1}{N} \sum_{x_i \in R} I(y_i = k) \tag{5.5}$$

Here N is the number of measurements in a node, R is the region defined by classification inequalities (decisions) leading to a node, y_i is the phenotype of an individual i and x_i is the allele vector for SNPs represented in a tree. Then the Gini index is defined for each node in a tree in the following way:

$$G = \sum_{k \neq k'} \hat{p}_k \hat{p}_{k'} \tag{5.6}$$

where indices k and k' span over all possible phenotype classes. The improvement in classification that is introduced by splitting a node into two can be expressed using Gini index:

$$\Delta G = G_0 - (p_L G_L + p_R G_R) \tag{5.7}$$

Here G_0 is the Gini index of a parent node to be split, while G_L and G_R are respective Gini indices of the left and right child nodes, and p_L and p_R describe the proportion of training set measurements that are assigned to each of the child nodes. The significance of a SNP can be measured by this metric. An alternative way to express the success of a classification induced by a particular split can be based on p-values of a chi-square test for the 2 × 2 table depicting the number of cases and controls belonging to each child node.

Alternative way to measure variable importance within random forests can be based on the permutation accuracy importance. The idea behind this method is to permute values of a predictor variable and then estimate the difference in prediction accuracy for the real and permuted predictor (Strobl et al., 2007).

All importance metrics discussed above can become biased when predictors differ considerably in terms of scale and the number of values they can assume (Strobl et al., 2007). For example, Gini index gives preference to variables with more categories (Boulesteix, 2006a,b; Kononenko, 1995). However, in the current study all predictors are rather uniform, therefore Gini index is a reliable measure of the significance of associations. It has been shown that Gini index is more successful in identification of important variables than the permutation importance when predictors are highly correlated (Nicodemus and Shugart, 2007). Permutation importance is more biased towards correlated variables than Gini index. To solve this problem a method exploiting conditional permutation scheme has been proposed (Strobl et al., 2008).

5.3 Stepwise Model Selection

The first method considered for building joint models is the stepwise model selection (Davison, 2003). This algorithm proceeds in a greedy manner by selecting a locus that improves a model the most in each step. This locus is added to a model and search is resumed. Sometimes removal of a non-significant term is also considered as a next step in model building. The process is continued until there are no more features that can significantly improve a model. The statistical significance of the model improvement is usually measured by Akaike's information criterion (AIC):

$$AIC = -2\sum_{i=1}^{n} l_{\hat{\beta}_1,\hat{\beta}_2}(X_i)) + 2p \tag{5.8}$$

where l is the logarithm of likelihood of a model in equation (5.2) and p is the number of parameters that are estimated within a model. By introducing the number of parameters in the model alongside log-likelihood, AIC penalizes large numbers of parameters. This statistic is suitable for making a decision whether to add or remove a term as well as for making the comparison between added terms or removed terms.

Stepwise methods are appropriate for dealing with dense SNP maps, because they do not try to describe the whole feature space and relationships between features, but instead build a sparse model from it. This is consistent with our expectations of having only a small fraction of SNPs associated with a phenotype. Also this method is suitable for discerning between closely linked variants, because once the best SNP within the region is added to the model other linked SNPs will be ignored in further feature selection rounds due to conditioning on the SNP added before.

5.4 Ensemble Learning and Probabilistic Model Selection

One of the disadvantages of stepwise model selection is that it is prone to overfitting. It can happen relatively soon in the model selection process, thus not all loci having similar magnitude of effect upon phenotype get a chance to enter a model unless their impacts upon a phenotype are orthogonal. There are two approaches to solve this problem. The first approach relies upon adding terms to a model only partially (typically it means rescaling of the magnitude of effect by a small factor ε before adding a term). The second approach is to make model selection probabilistic instead of deterministic so that all similarly important terms are equally likely to be added to a model. In this study preference is given to the second type of methods, because partial adding of terms can seriously suffer from missing alleles of SNPs even if the missing data rate is as low as 5% for each SNP. If a hundred of such SNPs are involved in a single model the proportion of individuals not having a missing allele for any SNP gets very small. It means that it is not possible to exclude missing data points, but instead a separate category for missing data has to be created (a practice poorly justified from theoretical point of view).

Within this study methods for probabilistic model selection based on subsampling of individuals are used. The idea is to combine small and moderately successful models into a bigger and more powerful phenotype prediction network. Each of the small models is built with a random subsample of measurements. Afterwards, models are combined via weight assignment to each of them. There are two problems to solve: selection of simple models from potentially huge feature space and putting together a big model, where each small learner has a specific weight and role. The algorithms that solve these tasks represent a type of ensemble learning strategy.

I will formulate the task of ensemble learning more precisely. The composite model explaining thr data will assume the form $F(x) = \beta_0 + \sum_{m=1}^{M} \beta_m f_m(x)$, where $f_m(x) = f(x; p_m)$ is a function of certain form calculated on data x and extra parameters $p = (p_1, p_2, \ldots)$. The task is to find a subset of functions $f_m(x)$ ("small models") that together will yield a good composite model, i.e., select $\{f_m(x)\}_1^M \subset \{f(x;p)\}_{p \in P}$ and calculate $(\beta_0, \beta_1, \ldots, \beta_M)$. The goodness of model fit is measured by some loss function $L(y, h(x))$, where $h(x)$ is the composite function ("big model") trying to explain the outcome y.

The general form of ensemble learning with individual subsampling is (Friedman and Popescu, 2003):

```
1. $F_0(x) = 0$
2. for $m = 1$ to $M$ do
   (a) $k_m = argmin_k \sum_{i \in S_m(\eta)} L(y_i, F_{m-1}(x_i) + f(x_i, k))$
   (b) update $F_m(x) = F_{m-1}(x) + \varepsilon \cdot f(x_i; k_m)$
3. return $\{f_{k_1}(x), f_{k_2}(x), \ldots, f_{k_M}(x)\}$
```

where $S_m(\eta)$ is a subsample of measurements of size $\eta \leq N$, but $0 \leq \varepsilon \leq 1$ is an updating step size controlling how much each of the selected small models should contribute to the total model.

A version of ensemble learning corresponding to the random forest algorithm is used throughout this study (Breiman, 2001; Liaw and Wiener, 2002). According to this type of learning strategy $f_m(x)$ is a decision tree and $\varepsilon = 0$. I use the random forest algorithm with $\eta = 0.632 \cdot N$ and perform sampling of individuals without replacement. This is practice is equivalent to using $\eta = N$ together with sampling with replacement. It has been shown that sampling without replacement does not induce any bias in variable selection for trees as opposed to sampling with replacement (Strobl et al., 2007). In addition subsampling without replacement is dependent on weaker assumptions for statistical inference (Politis et al., 1999). Therefore this approach is preferred in the current study. $\eta = 0.632 \cdot N$ is a typical choice of a subsample size, when sampling without replacement is used, and is recommended in many sources, although there are authors who suggest using $\eta = 0.5 \cdot N$ instead (Friedman and Hall, 1999). A specific problem connected with random forests is the sensitivity of their predictions to unbalanced subsamples of individuals, where the number of cases is not equal to the number of controls. If unbalancedness reaches a critical threshold, it becomes highly beneficial for the ensemble of trees to always cast a vote for the majority class of individuals within a subsample. One of the methods that can be used to solve this problem relies on down-sampling. According to this approach 0.632% of individuals from the minority class are selected and then the same amount of individuals is taken from the majority class (Chen et al., 1999). This method is used in the current study instead of alternative balancing strategies based on assigning weights to the votes of trees (Chen et al., 1999).

The algorithm for building a single tree introduces more variability in the learning process by SNP selection, because only a subset of SNPs is considered in each round of tree growing. This is beneficial for trees to exhibit larger variance within the ensemble. After growing individual trees they are combined by taking an average (in case of regression) or majority vote (in case of classification) of their predictions, like in bagging. A single tree is constructed in the following way according to the CART algorithm (Breiman et al., 1984):

until the minimum amount of individuals n_{min} that are classified to a node is reached

```
for each terminal node (leaf) of tree do
  i. select $k$ SNPs from the total number of $p$ SNPs
  ii. find the SNP from the selected set and split of its alleles that improves phenotype classification most (produces the smallest loss)
  iii. split the node into two
```

The recommended choice of k is $[\sqrt{p}]$ for classification (Hastie et al., 2008). This value of k is used within this study as well, because preliminary experiments have shown (results not depicted) that the optimal size of SNP subsample does not differ considerably from the recommended value. The minimum size of a node $n_{min} = 1$ for classification problems. Trees are grown up depth 5 within this study and results obtained from various depths are compared. The depth of trees has to be restricted somehow, because there are almost 5,000 individuals in total, therefore it will take many steps to arrive at nodes of size 1 (preliminary experiments have shown that there are hundreds of SNPs involved in a single fully grown tree). On the other hand, increasing n_{min} might not work either, because this practice is potentially discriminating for rare SNPs. The second argument in favour of limited depth of trees is the missing data rate. If tens of SNPs with 5% alleles missing are combined in a single classification path, very few individuals would be classified solely by non-missing alleles (tens of SNPs in a path is a realistic number for the amount of individuals to classify). The problem is somewhat smaller than in the case of gradient boosting, because each disease status is classified by a subset of SNPs in a tree, therefore it is not required that all alleles for all SNPs in a tree are known for each individual. Still there are many issues connected with missing alleles that are similar to the gradient boosting algorithm.

Various methods exist to determine when the model selection is finished and no significant model terms can be added to it. Using AIC together with the stepwise model selection procedure as a criterion for adding terms has already been discussed. For probabilistic model selection I rely upon cross validation to determine the stopping point (Hastie et al., 2008). However, instead of using the traditional scheme of cross validation based on splitting of individuals into N independent parts that are later used as test sets for estimation of prediction error, I use the approximation of cross validation based on out of bag samples. This approximation is suitable for random forests and produces test error that is very similar to cross validation error, while requiring much less computation time (one round of random forest building instead of ten). The idea is to estimate test error from predictions of phenotypes which are done with smaller ensembles of trees. The diminished ensemble consists only from trees that did not consider the individual to be predicted in the tree building process. This means that phenotype of individual i is predicted by a tree m, if $i \notin S_m(\eta)$ of the tree. A majority vote from trees in the diminished ensemble yields the final prediction for each individual. Once the test error obtained thus stabilizes and converges to a single point it is very likely that the ensemble is saturated.

5.5 Data Quality Control for Human SNPs

First, I applied all data quality filters that are used in the initial analysis of WTCCC data set (Wellcome Trust Case Control Consortium, 2007). One of the most important reasons for having nearly perfect agreement upon quality control methods is the ability to compare results with previous analysis. By carrying out the initial stages of data analysis in the same way it is possible to observe how the model selection algorithms affect the power of association discovery and how closely they can replicate results obtained by simple one locus models.

Thus, I applied the WTCCC list of excluded individuals, incorporating those suspected of different ancestry, those having high missing data rates over all SNPs, and those exhibiting identity or relatedness to other individuals under study. SNPs were omitted if their missing data rate in the context of the study of a particular disease was $> 5\%$. Also SNPs with MAF $< 1\%$ were excluded, but those with MAF $< 5\%$ were further tested to ensure that their missing data rate is $< 1\%$. A smaller number of SNPs were excluded due to being in noticeable Hardy-Weinberg disequilibrium or because of showing spurious associations for control individuals. In the previous study, SNP clusters yielding associations were inspected visually and poorly clustering SNPs removed. I used WTCCC list of SNPs with reported p-values, where poor clustering is tagged, to exclude this category of SNPs. After performing all data cleaning steps there were almost 5000 individuals and 327,000 SNPs left to study associations with T1D.

Previously no significant population stratification was found in the human case/control data (Wellcome Trust Case Control Consortium, 2007). I relied on these results and did not fit models with population related parameters. There are only 13 rather small chromosomal regions exhibiting some population structure, which were considered when interpreting results of model selection methods.

Chapter 6

Association Studies for Human Case/Control Data

6.1 Stepwise Model Selection and Associations with T1D

First stepwise model selection was tried for association discovery. Forward stepwise regression was performed until 20 loci were included in the model. Within each model selection step model likelihood and AIC was assessed. However, it was not possible to treat AIC as an estimator of test error due to the missing alleles of selected SNPs. Throughout model selection process individuals with missing alleles were omitted, thus the set of individuals considered in the next step was slightly diminished in comparison with the set of individuals in the previous step. It was not possible to omit individuals in advance either, because the exact subset of individuals used is dependent upon SNPs selected in each step. Therefore model likelihoods and AIC were subjected to permutation tests. Models for 99 phenotype permutations were obtained using the same stepwise regression algorithm. Improvement in AIC for models explaining permuted data and real data is depicted in Fig.6.1. It turned out that all 20 forward selection rounds yielded bigger AIC improvement for the real data than for any permutation. It means that adding of each locus to a real T1D model led to significant improvement in model likelihood. An additional thing to notice is that even permuted phenotypes give rather large improvement in AIC (over 50), despite the fact that these changes in likelihood are worse than those associated with the real case/control data. Most probably this kind of improvement is related to the problem of missing data outlined above. It also underlines the impossibility to assess models using asymptotic distribution of log-likelihood ratio.

Most of the 20 SNPs selected by stepwise regression are not significant according to the previous WTCCC study (see Table 6.1). Two out of five strong associations reported in WTCCC study are replicated, namely regions on chromosome 6 and 1, which are represented by rs9272723 and rs6679677. The most popular chromosome within the selected model is chromosome 1, which is contributing 4 distinct loci. Chromosome 6 is met three times. Some of the associations discovered seem to be very unusual, like loci on chromosomes 7 and 15 having low WTCCC p-values.

Overall percentage of successfully replicated associations seems to be rather low. Such a result is quite unexpected, because stepwise regression should give similar associations during the first steps and slightly enhanced signals from moderate associations during later rounds. It might be explained by the model selection criterion which is based on AIC. Apparently, ignoring of individuals with missing alleles is leading to subtle bias in the model selection. It seems that AIC is preferring models that are both good for explaining phenotype and also successful in getting rid of inconvenient cases and controls (the more, the better). Therefore the set of 20 loci selected might be only an approximation of the true picture. On the other hand, permutation test indicates that there are at least 20 significant loci, some of them linked, that can contribute towards explaining T1D phenotype. Thus, the full potential of stepwise model selection is still not achieved. One suggestion to solve the problem of biased likelihood estimations can be based on using likelihood ratios instead of pure likelihoods. Then the likelihood of the new model with added SNP can be compared to the likelihood of model selected in previous round, if individuals with missing alleles are excluded according to the largest model with extra SNP. Such a selection procedure based on likelihood ratios was implemented, however, a new round of permutation tests based on likelihood ratio comparison was not carried out due to serious requirements for computational time. At this point permutation test results in Fig.6.1

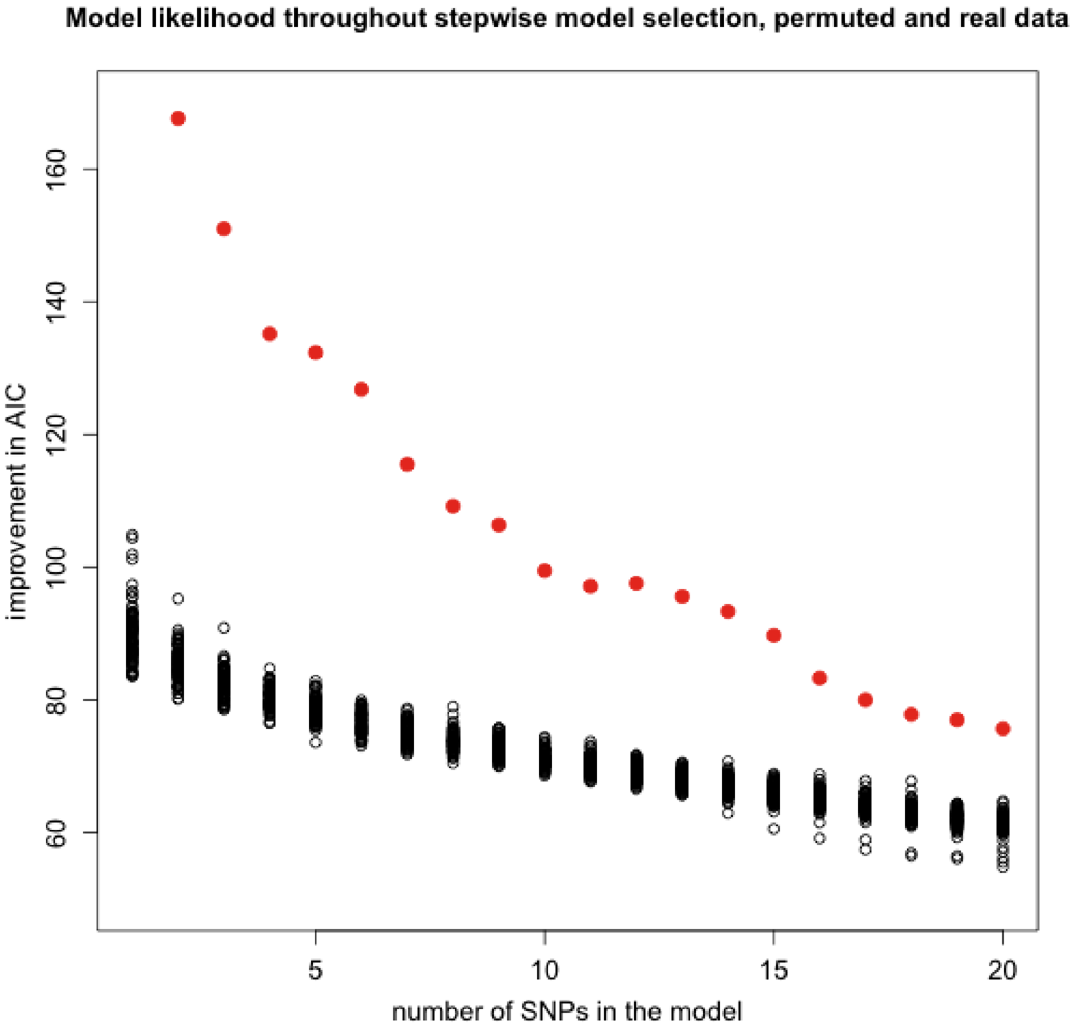


Figure 6.1: Permutation test assessing model likelihoods throughout the stepwise model selection. The red dots indicate change in AIC yielded by the real phenotype, while black dots correspond to model improvement for permuted phenotypes.

are considered as sufficient evidence for the existence of at least 20 loci that can significantly improve the association model.

Table 6.1: SNPs selected by stepwise regression, AIC criterion

SNP name	Location	ΔAIC	WTCCC p-value	SNP type
rs9272723	6, 32609427	684.01	3.32272e-128	intronic
rs41515647	1, 77517009	167.64	0.000187856	intronic
rs16849921	2, 214061022	151.01	0.00670717	intergenic
rs9366943	6, 37602406	135.16	0.00207957	3 prime UTR
rs972357	18, 29551854	132.34	0.901465	intergenic
rs7632456	3, 133575993	126.78	0.515676	intronic
rs3118595	9, 137427767	115.50	0.76196	within non coding gene
rs16829733	2, 134645600	109.19	0.750206	intergenic
rs41562	7, 104845094	106.33	0.0206315	intronic
rs6519313	22, 42552875	99.45	0.383243	upstream
rs7725644	5, 167276307	97.11	0.215327	intergenic
rs12413409	10, 104719096	97.55	0.0489497	intronic
rs2438083	6, 1277371	95.58	0.894945	intergenic
rs2056975	1, 22408527	93.32	0.0511778	intronic
rs6679677	1, 114303808	89.70	1.16646e-26	upstream
rs12592799	15, 89363596	83.27	0.418871	intronic
rs12578219	12, 79318928	79.97	0.00709807	intronic
rs10898312	11, 71280313	77.79	0.903	downstream
rs858305	7, 23242830	76.97	0.368365	upstream
rs7518159	1, 29621460	75.63	0.100213	intronic

The results of the modified stepwise model selection are depicted in the Table 6.2. The first six loci entering the model belong to the strongest association regions in previous WTCCC study. First, third and eighth SNPs represent MHC region. The second SNP hits well known T1D association region on chromosome 1 encompassing PTPN22 gene. Chromosome 12 is represented in the model with two loci, both of them successful replicates of associations on 12q13 and 12q24. The discovered region on chromosome 16 corresponds to the association locus on 16p13.

If the remaining loci in selected model apart from the strongest WTCCC associations are considered, it is easy to spot some more replications of the well known associations (see Fig.6.2). For example, the discovered SNP on chromosome 4 around 123Mb is located close to IL2-IL21 genes. The SNP on chromosome 18 at 12.7Mb is not far from the association region on 18p11 that is connected with T1D risk as well as with CD and RA. The region on chromosome 2 at 154Mb seems to be a novel association, which is far from the CTLA4 gene. Another novel association is located on chromosome 5 around 9Mb. The remaining model terms are quite unexpected and incorporate loci on chromosome 1 (212Mb), chromosome 7 (45Mb), chromosome 5 (86Mb) and chromosome 15 (32Mb) (Fig.6.2). Also two associations on chromosome 10 seem to be novel.

None of the chromosomes, apart from the chromosome 6, is represented in the selected model by more than two SNPs. It means that stepwise model selection can produce sufficient diversity within model and also avoid repeated hits of nearby loci or loci in strong linkage disequilibrium. All of the strongest associations of the previous study were successfully replicated this time. All p-values of currently selected SNPs within previous study were smaller than 0.00171814 (see Table 6.2). Thus, even the newly discovered associations exhibit some consistency with p-values of a simpler additive model. However, the significance of many loci within stepwise model building is greatly enhanced. Interestingly enough, the order in which loci are selected by the stepwise regression does not correspond to the decreasing significance within the previous study as seen from Fig.6.3. Although the first two SNPs are extremely significant according to the WTCCC analysis by the additive model, the significance of SNPs selected in further rounds fluctuates greatly. Only after round 10 a steady decrease in -log(reference p-values) can be observed in comparison with the first rounds.

Table 6.2: SNPs selected by stepwise regression, likelihood ratio criterion

SNP name	Location	Δlikel.	WTCCC	SNP type	cand. gene
rs74223445	6, 32604372	336.42	2.42e-134	intronic	HLA-DQA1

continued on the next page...

Table 6.2 – continued

SNP name	Location	Δlikel.	WTCCC	SNP type	cand. gene
rs6679677	1, 114303808	56.06	1.16e-26	upstream	RP11-426L16.1
rs2523691	6, 31420687	33.91	1.48e-10	within non coding gene	HCP5
rs17696736	12, 112486818	25.39	2.17e-15	intronic	C12orf30
rs11171739	12, 56470625	19.32	1.14e-11	upstream	OR9K1P
rs12924729	16, 11187783	15.38	2.21e-08	intronic	CLEC16A
rs17388568	4, 123329362	15.16	5.00e-07	intronic	ADAD1
rs9272723	6, 32609427	12.45	3.32e-128	intronic	HLA-DQA1
rs203884	6, 28077374	11.83	1.51e-14	intergenic	U2
rs2542151	18, 12779947	10.04	1.89e-06	intergenic	GNAL
rs1595719	2, 154095977	9.95	0.00171	intergenic	Y RNA
rs10807124	6, 33404064	10.23	2.74e-09	intronic	SYNGAP1
rs415024	5, 9392358	9.85	0.000114	intronic	SEMA5A
rs10863988	1, 212894474	9.92	0.000353	intergenic	AC096637.1
rs9366216	6, 170676326	9.85	0.000420	intronic	FAM120B
rs17230937	7, 45263634	9.55	0.00135	intergenic	CAMK2B
rs2544677	5, 86399262	9.34	8.23e-06	intergenic	NBPF22P
rs7097035	10, 21370420	9.09	0.000327	intronic	NEBL
rs2666236	10, 33418872	8.75	2.12e-05	intergenic	AL445071.1
rs12907720	15, 31897012	8.87	0.000298	intronic	OTUD7A

The effect of 20 selected loci upon T1D phenotype is depicted in Fig.6.4. The first two SNPs and rs9272723 have the most significant impact upon T1D. All selected SNPs show some additive effect upon phenotype, although magnitude of the effect varies greatly from very large (rs9272346) to quite negligible (rs17388568). In addition, the magnitude of effect does not always decrease in more advanced model selection rounds. Apparently, some of the SNPs are selected due to their interactions with other SNPs in the model or conditional effect. To explore the conditional effects of SNPs further, I depicted interactions of loci sequentially entering the model in Fig.6.5. Here each line corresponds to a different genotype of the locus that was selected in the previous round. The shape of line denotes the conditional effect on phenotype of the locus selected in the current round. Thus, parallel lines correspond to the additive summation of

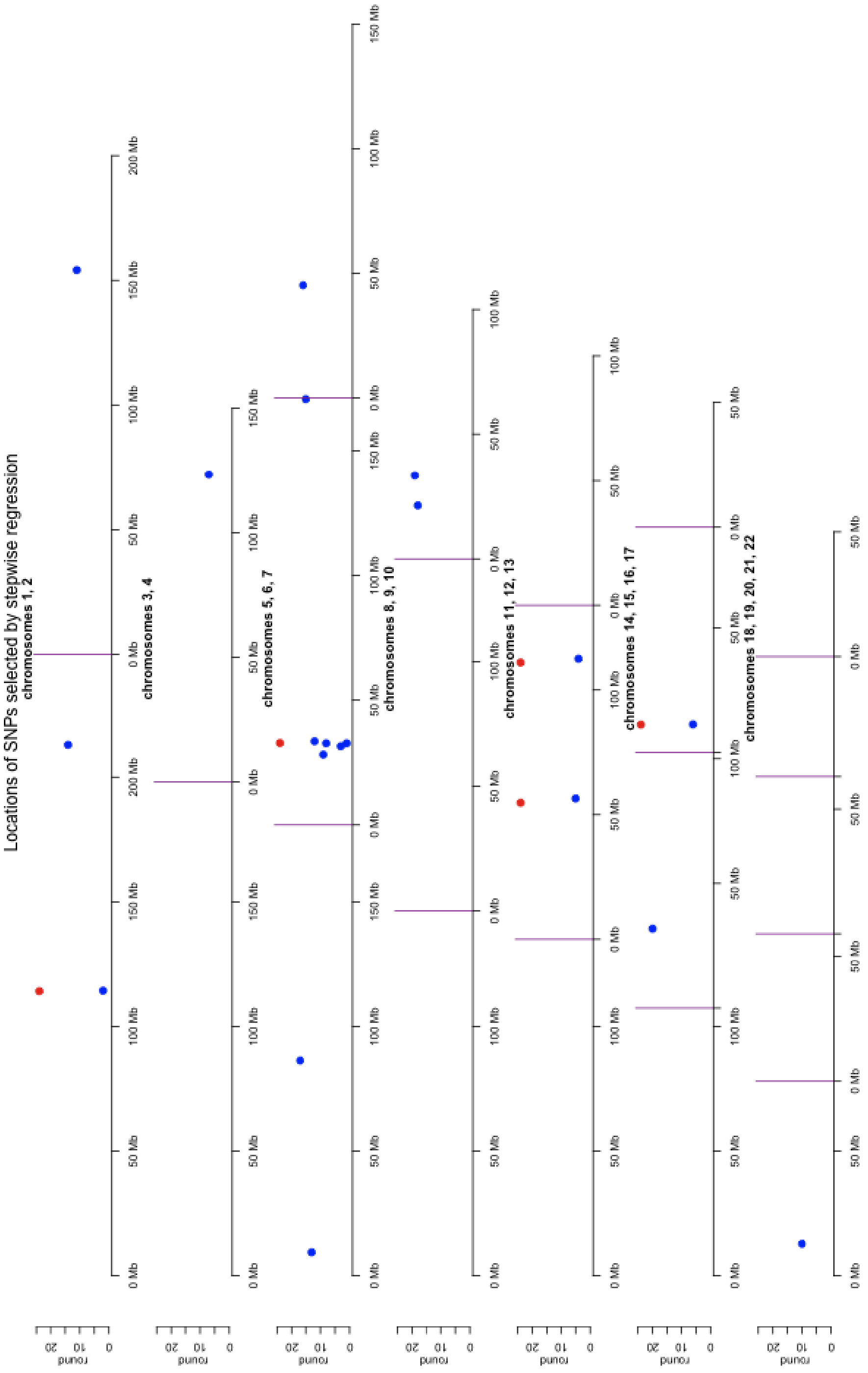


Figure 6.2: Locations of SNPs selected by the stepwise regression (marked by blue dots). Y axis corresponds to the number of round in which particular SNP has been selected. The red dots indicate SNPs with p-values smaller than 10^{-8} in WTCCC study.

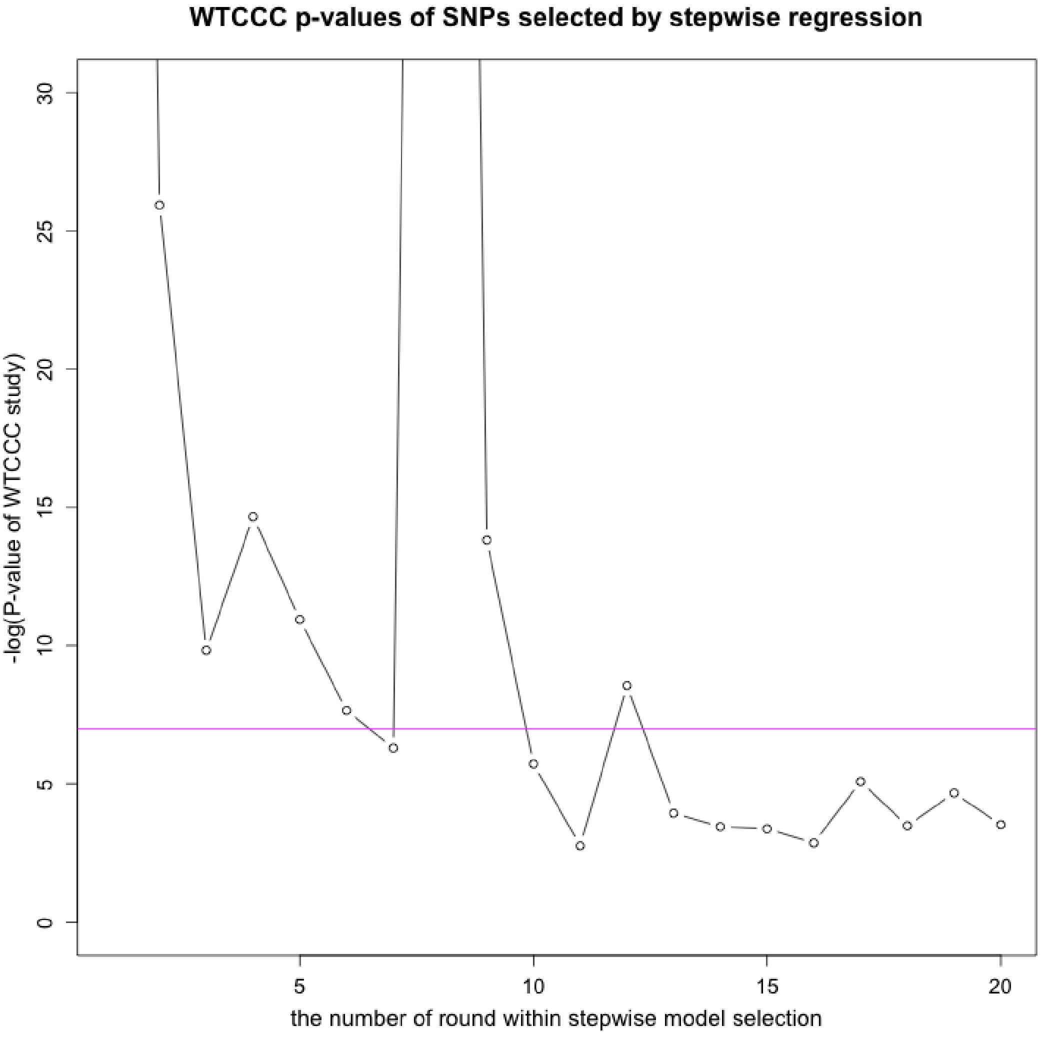


Figure 6.3: P-values yielded by the additive model within previous WTCCC study for the SNPs selected by the stepwise model selection by the number of selection round. The purple line separates p-values smaller than 10^{-8}.

the effects of two loci, while diverging lines correspond to some kind of interaction. A striking example of an epistatic effect would be lines crossing in which the proportions of cases to controls inverts for a particular genotype combination. The Fig.6.5 shows that in most cases SNP effects are summed over rounds. Departures from additivity are mostly connected with relatively rare alleles (see, for example, alterations of allele effects for rs2542151, rs9366216 and rs1595719). Thus, the conditional effect of a SNP is mostly enhanced via adding together several small SNP effects.

Similar study of conditional SNP effects was done for selected SNPs and five strongest WTCCC study associations. The results in Fig.6.6 show similar tendencies to Fig.6.5: effects are mostly additive with exceptions caused by rarer alleles. More detailed analysis of the linkage disequilibrium and closeness of location between selected SNPs and SNPs showing the strongest associations in previous study indicates that each of the strong associations has at least one locus in the selected model that is close to a previously discovered SNP. In addition, there is a very strong linkage disequilibrium between model SNPs and previous associations. It seems that stepwise model selection methods are highlighting other interesting variants within the linkage disequilibrium region of some well known associations. However, not much new information is gained about these regions, because in four cases out of six the linkage disequilibrium is perfect (see panels without any lines). The lack of connections between genotypes is denoting the lack of variance within genotypes after conditioning and clustering of all conditioned genotypes at a single point. None of the previously known strong associations enters the model twice or more frequently, apart from MHC region. This is a favourable property for discovering novel associations.

6.2 Association Discovery with Random Forest Method

WTCCC case/control data set containing nearly 2,000 cases and 3,000 controls and over 300,000 SNPs passing quality filters was analysed using random forests. Five different forests each containing 2,000 trees were built. The main difference

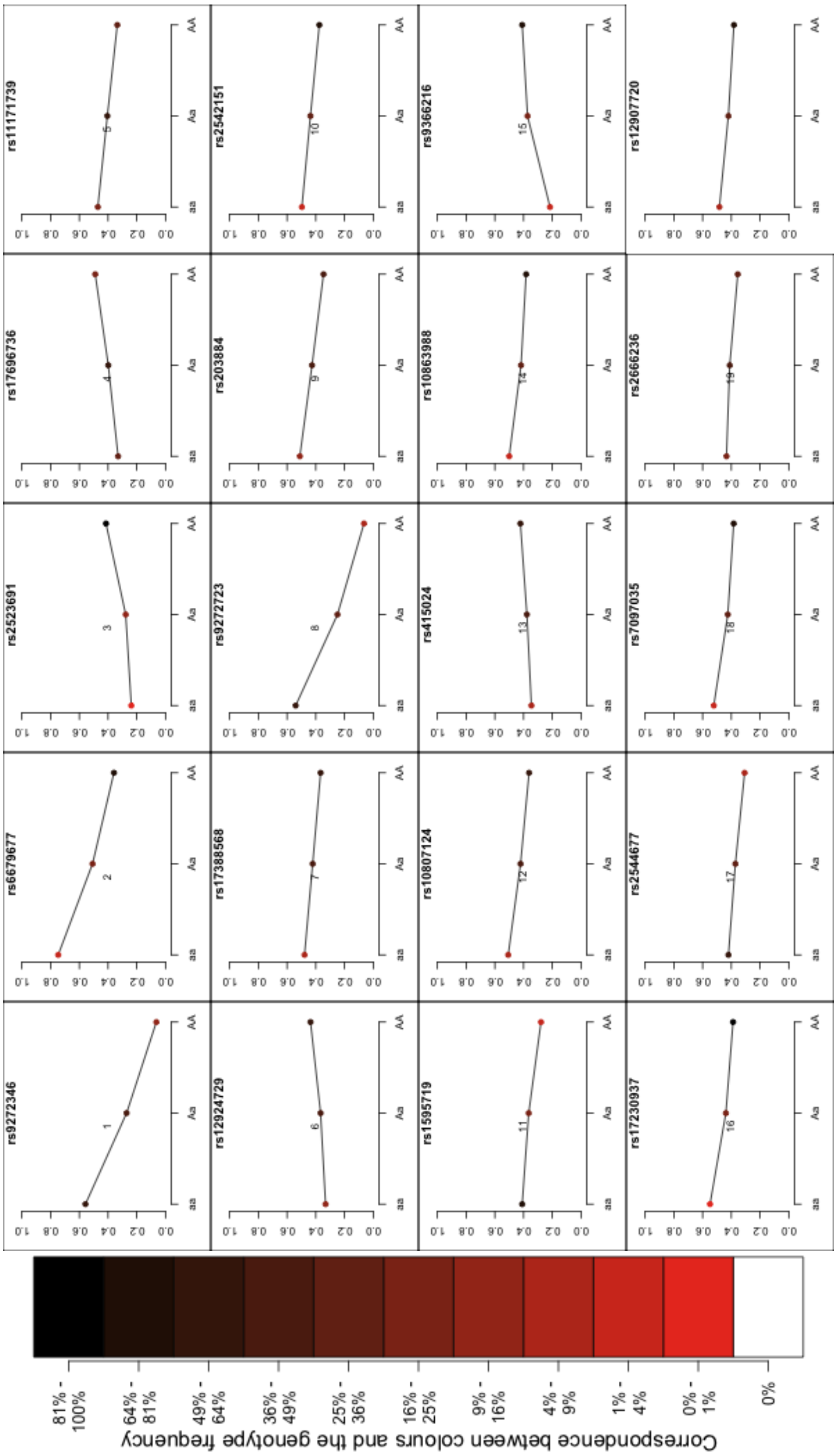


Figure 6.4: Effects of genotypes of selected SNPs upon the phenotype. Y axis denotes the percentage of cases among all individuals having a particular genotype. The colour of dots depicting different genotypes is set according to the genotype frequency (see the key for details). Small numbers in the centre of the plot indicate the sequence in which SNPs are entering the model.

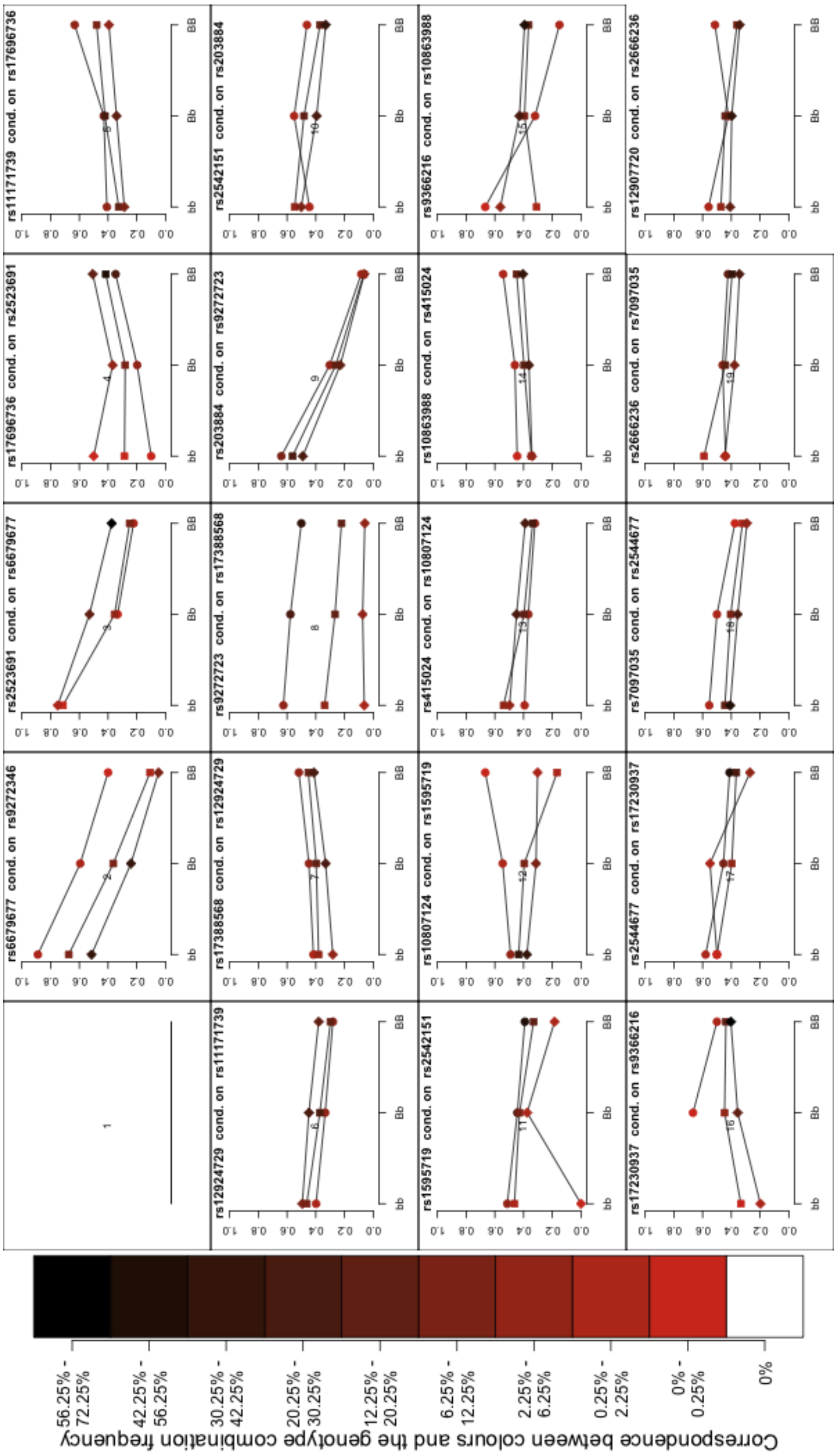


Figure 6.5: Effects of genotypes of selected SNPs, conditioned upon previous selected SNP. Each line corresponds to a different genotype of previous SNP. The genotype of the previous SNP is distinguished by point shape. The colour of dots indicates genotype frequency as before (see the key for details).

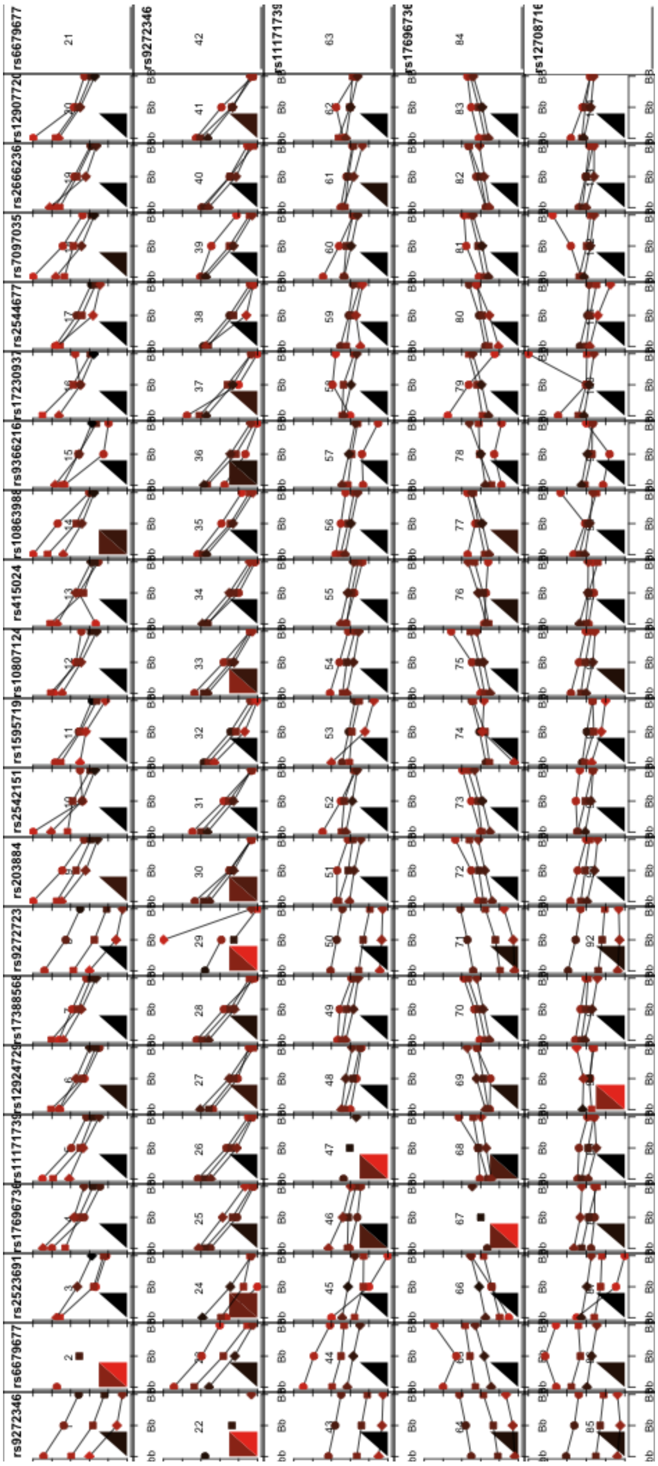


Figure 6.6: Effects of genotypes of previously discovered SNPs, conditioned upon SNPs selected by stepwise regression. The bottom triangle (the one with a side coinciding with the bottom line of the small plot) indicates the linkage disequilibrium, while the top triangle indicates the physical closeness of loci on the genome. If the top triangle is missing, SNPs are on different chromosomes. For the disambiguation of the colours see Fig.6.7.

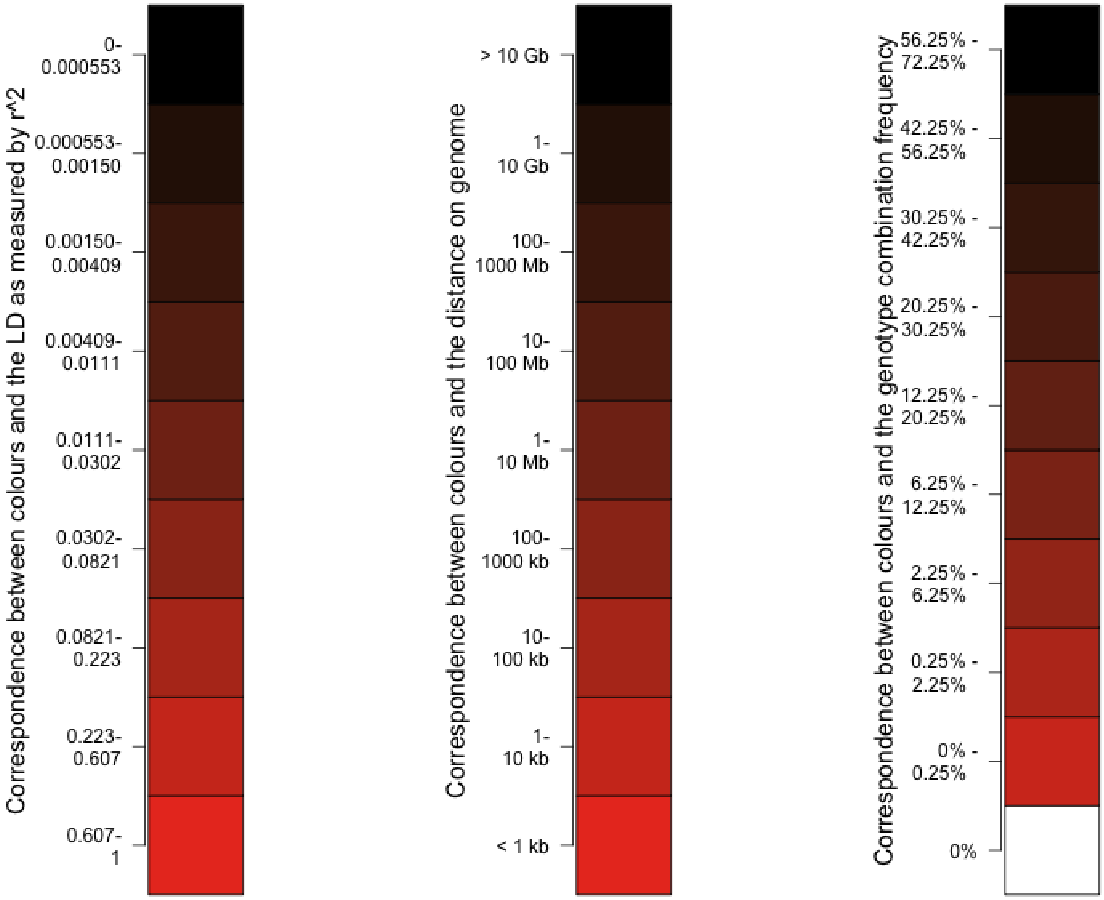


Figure 6.7: Colours used to depict the conditional effects of genotypes of previously discovered SNPs. Both linkage disequilibrium, absolute distance on genome and the genotype frequency are coded by distinct shades of red.

between them was the depth of trees used, which varied from 1 to 5. Thus, the maximum number of SNPs in a single tree is 31 SNP corresponding to depth 5.

The first task that has to be done when processing random forest output is determining the number of decision trees that is sufficient to explain the phenotype. To obtain this number cross validation based on out of bag samples was performed and test error estimated. Fig.6.8 depicts test errors obtained for different parameters.

As can be seen from the black lines in Fig.6.8, the best overall test error was obtained with trees of depth 2, where it approached 40%. Similar test error was obtained episodically with trees of depth 1, however later it stabilized at 45%. Ensembles of trees with depth 5 yield test error that is very similar to trees of depth 2. If the results are as similar as they are in the current experiment for trees of depth 2 and 5, usually preference would be given to smaller trees. However, predictions of trees of depth 5 are more consistent, because both cases and controls get comparable prediction error from ensemble (cases reach 40% and controls 45%). Such a consistency was not observed for ensembles of smaller trees, where prediction accuracy for cases and controls is diverging heavily. In addition, trees of depth 2 yield much larger test error for cases than controls, therefore this ensemble is not very useful or informative despite the small overall test error. The prediction error stabilizes around 800 trees within all ensembles. It means that the optimum number of trees within an ensemble is 800 or more. From the ensemble of trees of depth 5 1000 entities were selected for further analysis, because test error converges to a constant level only after 800 trees.

Now that the optimum size of ensemble is determined it is possible to assess the importance of particular SNPs in classification. The improvement in classification induced by a particular SNP can be measured by the change in Gini index, as described in previous chapter. If a comparison of SNP importance across all trees is required, the Gini impurity measure of a node has to be rescaled by the size of a node in terms of the number of measurements belonging to it. Otherwise large nodes are penalised for having to deal with large sample sizes which are harder to classify. Such a rescaling is also used when the goodness of a whole tree instead of a single split is assessed and Gini impurities are summed (Hastie et al., 2008). The distribution of node sizes for each depth of a split is depicted in Fig.6.9 The

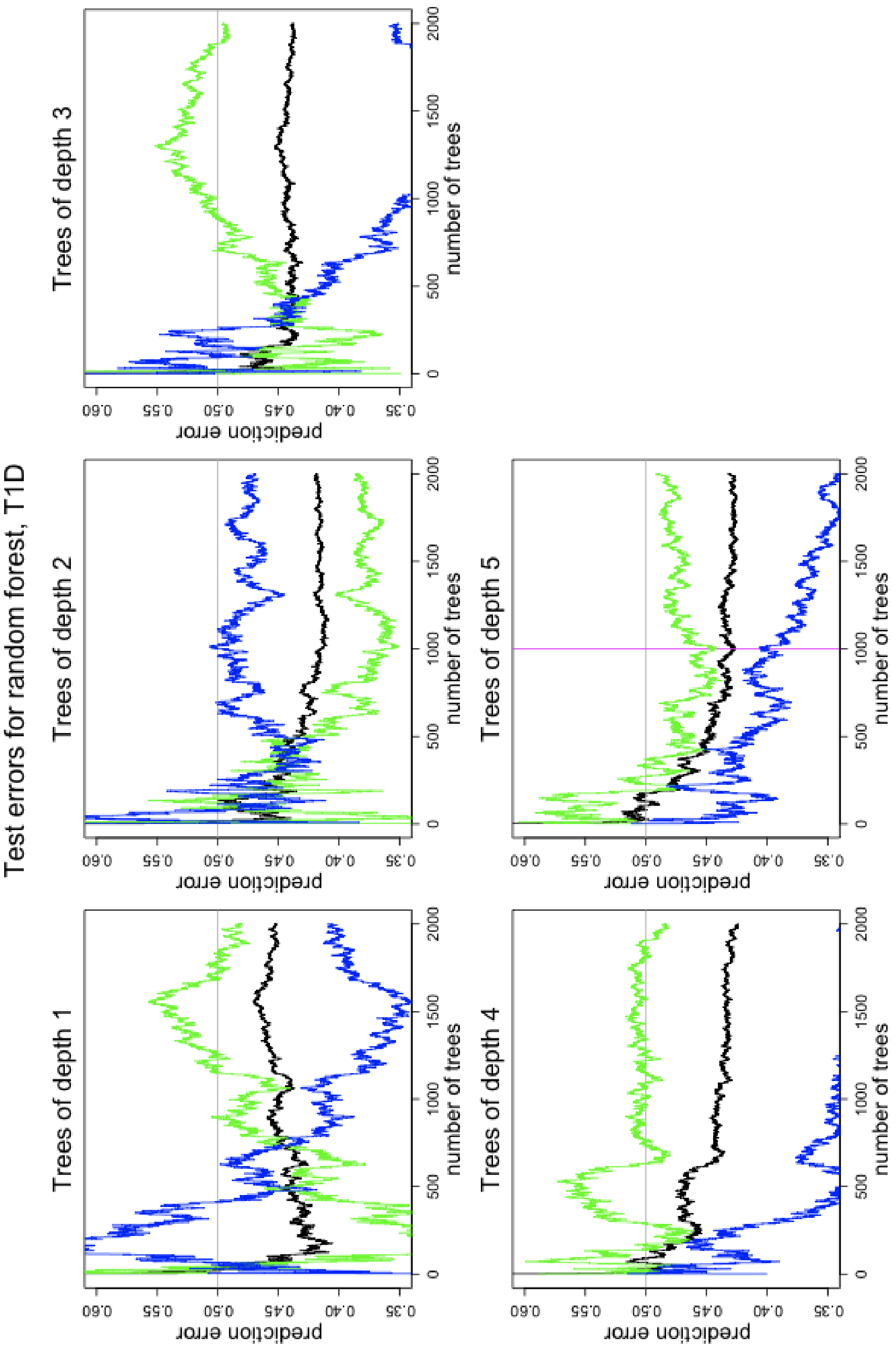


Figure 6.8: Test errors for various sizes of random forest and various sizes of decision trees. Black line corresponds to the overall test error, while the blue line denotes prediction error for cases and the green line depicts prediction error for controls. The gray line in the middle indicates 50% test error.

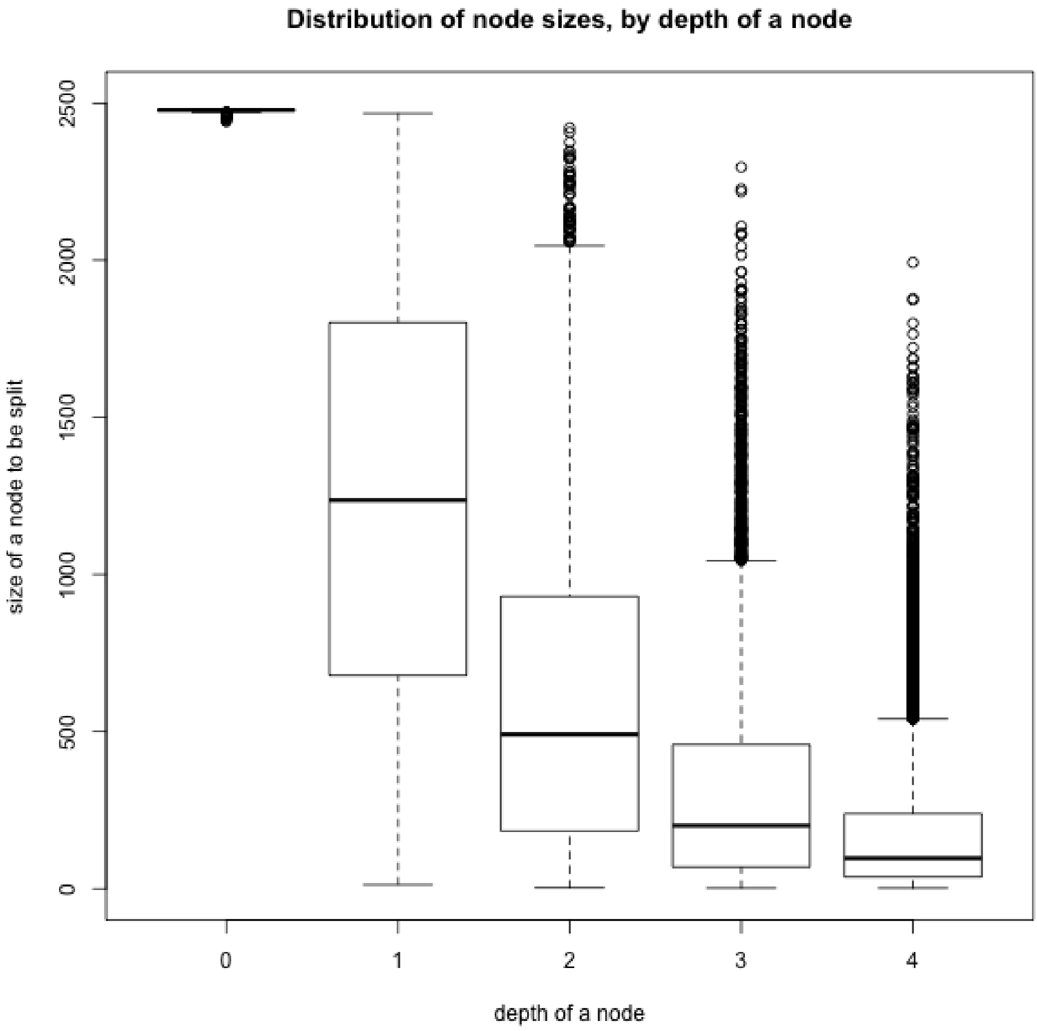


Figure 6.9: Boxplots of node sizes by the depth of a node within a tree.

figure indicates that trees within random forest are well balanced, because the mean size of a node decreases exponentially when the depth is increased. Also the variance of node size gradually decreases together with increase in depth, which is another proof of balancedness of trees. Nevertheless, there are more outliers for larger depths: a phenomenon that is probably unavoidable due to the large node count for these depths and accumulation of unbalancedness from previous levels.

Fig.6.10 shows SNPs selected by random forest algorithm, split by the depth of node in the tree, sorted by their Gini impurity. On the lower panels, the Gini score is plotted against the ranking of the SNPs by this measure. The plots for various depths only differ between nodes in the tails, with the nodes of depth 1 showing a thicker left hand tail, indicating that there are more ”important” SNPs selected in the first split. Right hand tails become longer at depth 5: apparently there are more relatively unimportant SNPs emerging at this depth.

The upper panel of Fig.6.10 depicts p-values of previous WTCCC study for SNPs selected by random forest algorithm. P-values yielded by additive model are shown. SNPs are ranked according to their importance within random forest, where importance itself can be read from the plots in the lower panel. Interestingly enough, not all highly ranked SNPs from random forest have small p-values in the previous study. It means that the model selection method used helps to discover SNPs that are good classifiers of disease status, but whose effect nevertheless cannot be adequately captured by additive model or, alternatively, whose unimportance in previous study is caused by a small subset of individuals that are probabilistically omitted in the sampling process to generate trees. Despite the noticeable diversity of p-values there is a strong correlation between the rank of SNP within random forest and p-value in previous study. This correspondence is rather pronounced for splits of depth 1, while it gradually deteriorates for deeper splits. Also the total number of SNPs with low p-values in WTCCC study is larger for splits at smaller depth. This tendency is rather pronounced despite the fact that the total number of SNPs involved in partitioning of nodes is exponentially increasing with the depth of split. It seems that the effects of SNPs emerging at deeper levels of trees are highly conditional upon the effects of SNPs at previous levels. Therefore many SNPs brought to forefront later in the tree

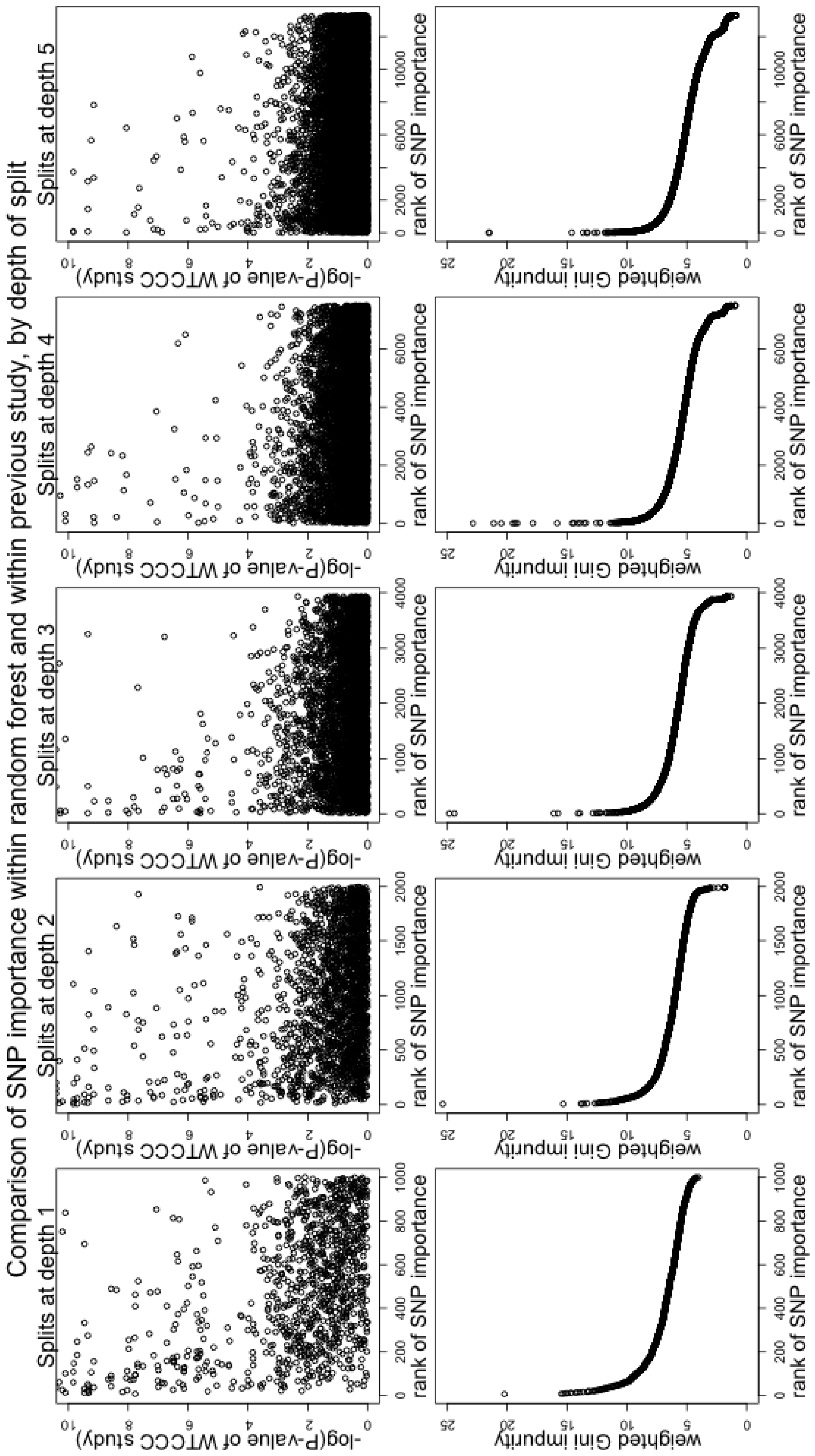


Figure 6.10: SNP importance as measured by weighted improvement in Gini impurity index and comparison with p-values of additive model from WTCCC study. SNPs are ranked by the Gini index both in upper and lower panel.

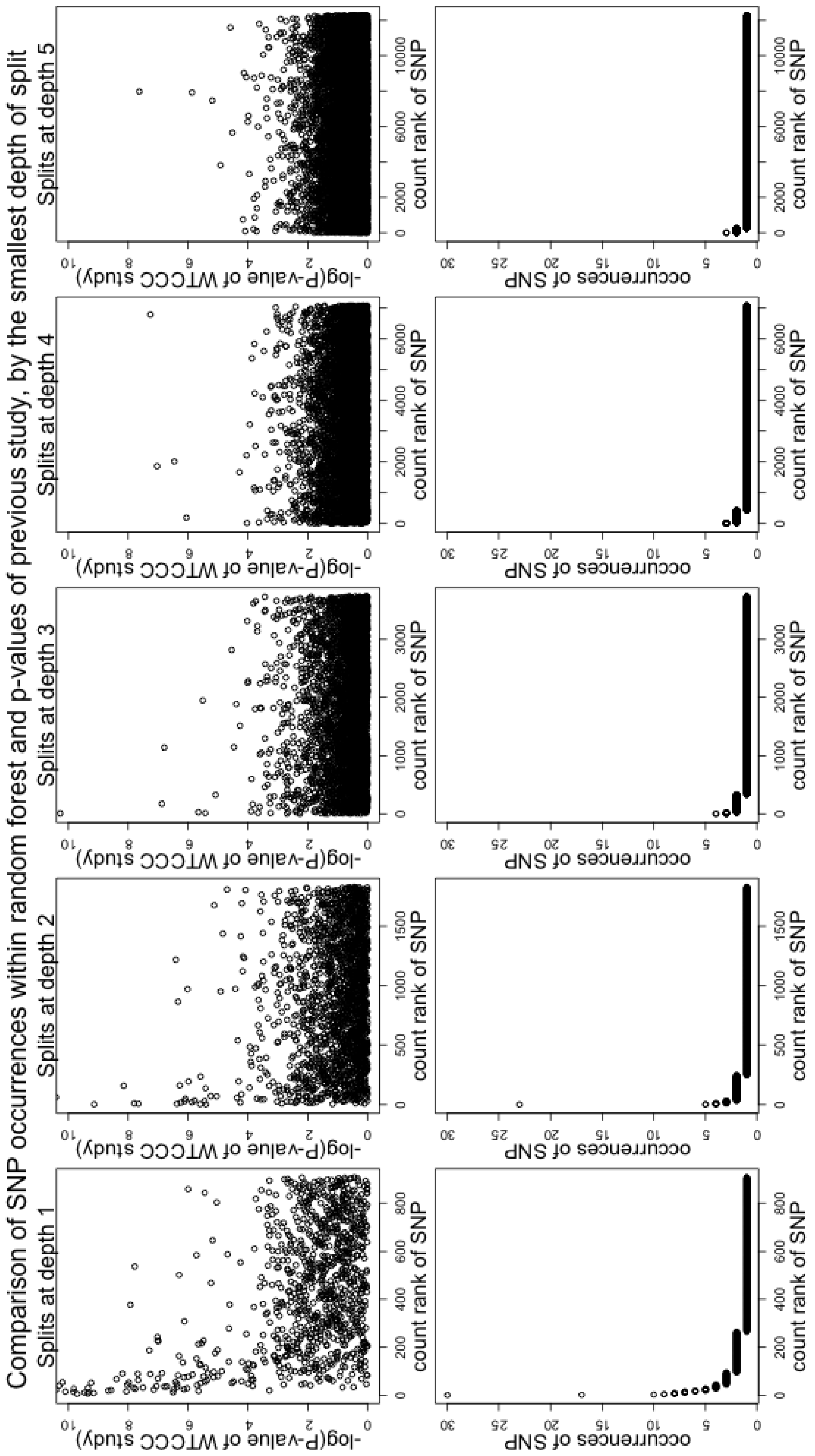


Figure 6.11: SNP importance as measured by the number of occurrences of a SNP within random forest and comparison with p-values of additive model from WTCCC study. SNPs are ranked by the number of occurrences.

building process could not be discovered with simple additive model previously. Nevertheless, the most important SNPs are still rather consistent among both studies and appear quite early in the tree building process.

Another, less elaborate and simpler, way of assessing the importance of a SNP within a random forest can be based on counting the occurrences of each SNP within ensemble of trees. If a SNP is included in the ensemble twice, it means that it has been the best predictor among nearly 600 SNPs two times ($[\sqrt{p}]$ used in the tree building algorithm is nearly 600) and also the best predictor for two different subsamples of individuals. It is very unlikely that a SNP gets included in the ensemble several times due to chance connected with individual subsampling and using only a subset of SNPs for splitting a leaf. Therefore such a repeated occurrence of the same SNP within ensemble can be strongly suggestive for an association. The comparison of p-values of WTCCC study and SNP occurrences within random forest in Fig.6.11 shows that larger number of occurrences is strongly correlated with high p-values of additive model. The correlation observed is more notable than the linear relationship between rank of scaled Gini impurity and p-values yielded by the same model. However, the correlation with p-values becomes very weak when the number of occurrences is low (2 or 1), as can be seen from the plots of deeper splits (from 3 onwards). Gini impurity measure seems to be more sensitive and precise, because even the tails of low rank SNPs show some correlation with p-values. The plot of SNP occurrences in Fig.6.11 is divided by the minimum depth of all splits induced by a particular SNP. It seems that the depth of a split does not affect the relationship between the number of SNP's occurrences and p-values of additive model.

Examination of p-value distributions for each number of SNP occurrences in Fig.6.12 confirms the strong relationship between both SNP importance metrics. The average p-value for a particular number of SNP occurrences monotonically increases together with increasing number of occurrences (the only exceptions are small jumps downwards from 7 to 8 and 9 to 10). However, more striking feature is the low variance of p-values and the lack of outliers for any particular occurrence number apart from the lowest ones. It means that SNP occurrences within random forest are a consistent predictor of an association strength. The most interesting are SNPs with 4 occurrences, because their p-values are a bit too

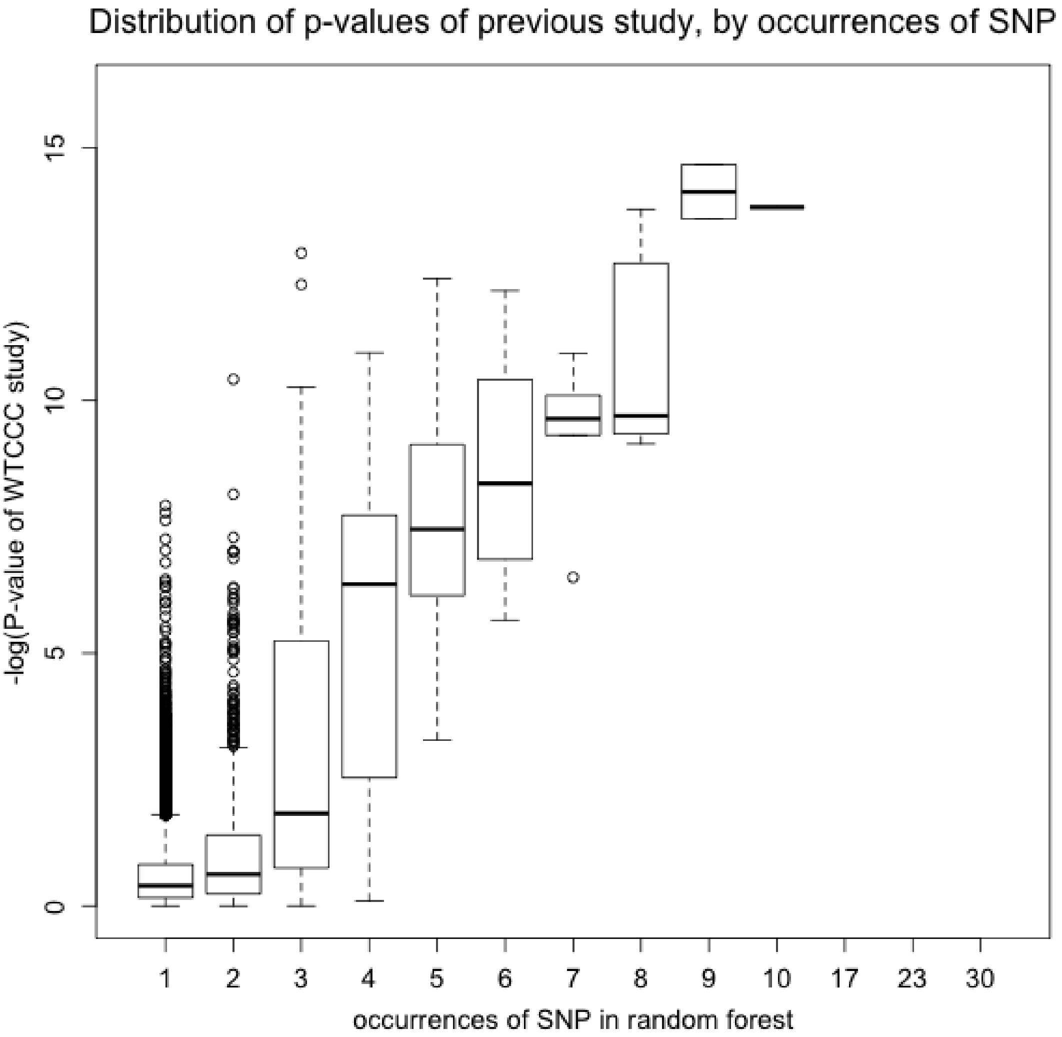


Figure 6.12: Boxplots of p-values of additive model from WTCCC study by the number of occurrences of a SNP within random forest.

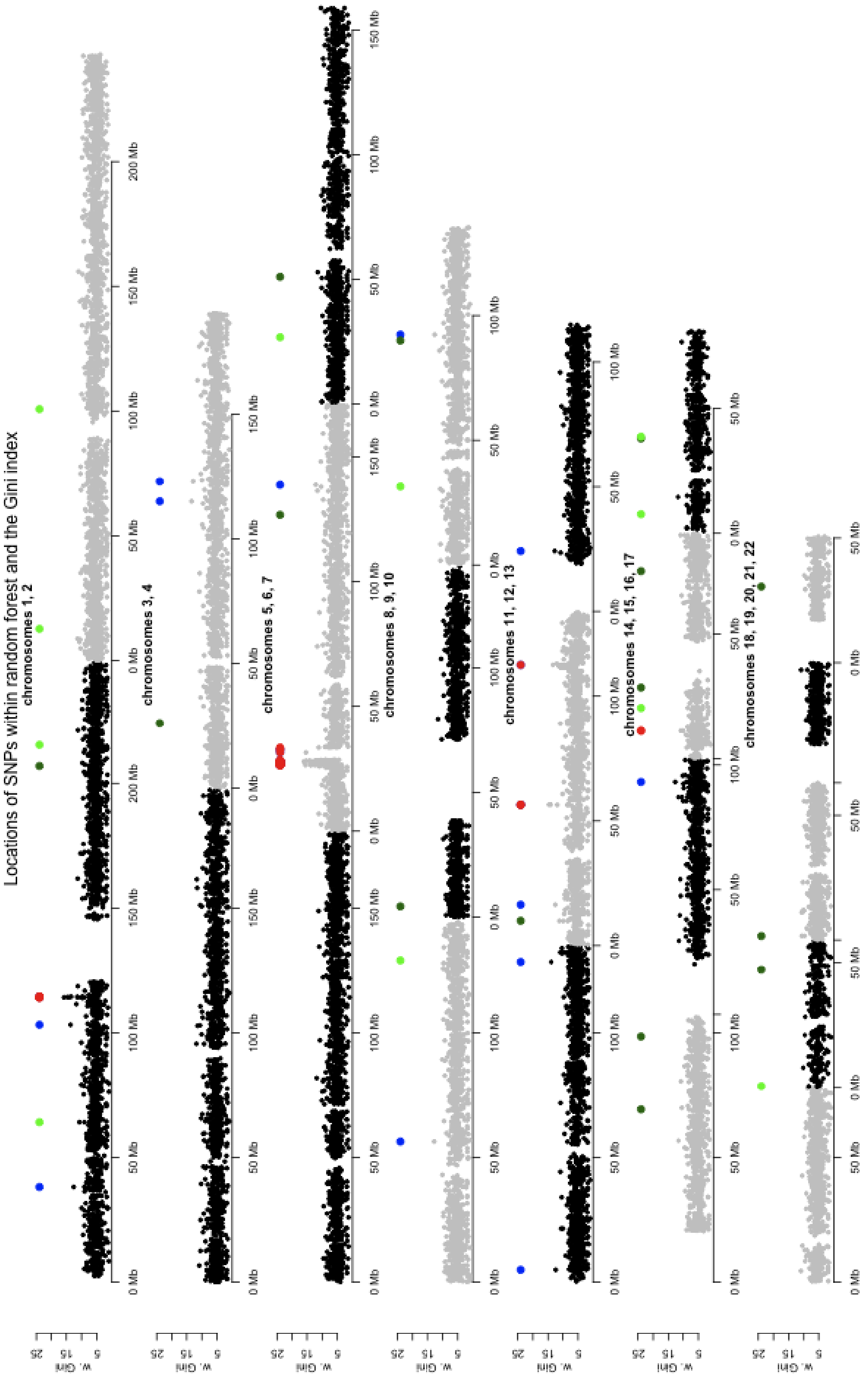


Figure 6.13: Locations of SNPs in random forest and the largest weighted improvement in Gini impurity induced by a SNP. Red dots indicate SNPs with p-values smaller than 10^{-8} in WTCCC study, blue dots correspond to the Gini index peaks, while green dots refer to the meta-analysis associations (dark green - replicated, light green - non-replicated).

high to be proclaimed significant after Bonferroni correction, nevertheless their consistent emergence in random forest is suggestive for a moderate association. These SNPs are candidates for association with T1D that have to be explored further.

Now that the SNP importance within random forest is assessed it is interesting to look at the genomic locations of the most significant SNPs. In Fig.6.13 a map of human genome is shown together with the improvement in Gini impurity yielded by each SNP of random forest. All genome regions having SNPs passing the quality filters are represented in the random forest. Homogeneous coverage of genome indicates that random forest algorithm is producing sufficient amount of variation within and among trees, and that there is no systematic bias generated by the method with respect to the genomic location. For most part of the genome Gini impurities do not exhibit any peaks, nevertheless there are a few notable clusters of SNPs associated with large improvements in Gini index. The densest cluster is located on chromosome 6, at approximately 30Mb. This cluster corresponds to the MHC region. Another important peak is located on chromosome 1, at 114Mb, which is the region of PTPN22 gene. Chromosome 12 has two regions of highly important SNPs: one close to 56Mb and another close to 112Mb (see Fig.6.13). Both regions are well known from previous association studies, including that of the first WTCCC data analysis. Chromosome 16 has one cluster of significant SNPs as well. It is located around 11Mb and has been reported in the results of previous WTCCC study. Overall the results of previous WTCCC study have been replicated successfully.

Apart from highly significant SNPs mentioned above there are also several clusters comprised of SNPs of moderate significance. There are two such regions on chromosome 1, around 35Mb and around 105Mb. Chromosome 4 has some evidence of association at a broader region spanning from 120Mb to 130Mb (Fig.6.13). Chromosome 6 has some outlying SNPs at 140Mb, however, the corresponding peak is not very high. There is a single SNP giving large improvement in Gini impurity on chromosome 8, at 55Mb. Both chromosome 10 and 11 have distinct association regions. They are situated at 90Mb (chromosome 10) and at 4Mb and 132Mb (chromosome 11) correspondingly. Chromosome 12 has an additional region of significant SNPs at 15Mb. The SNPs are not clustered as

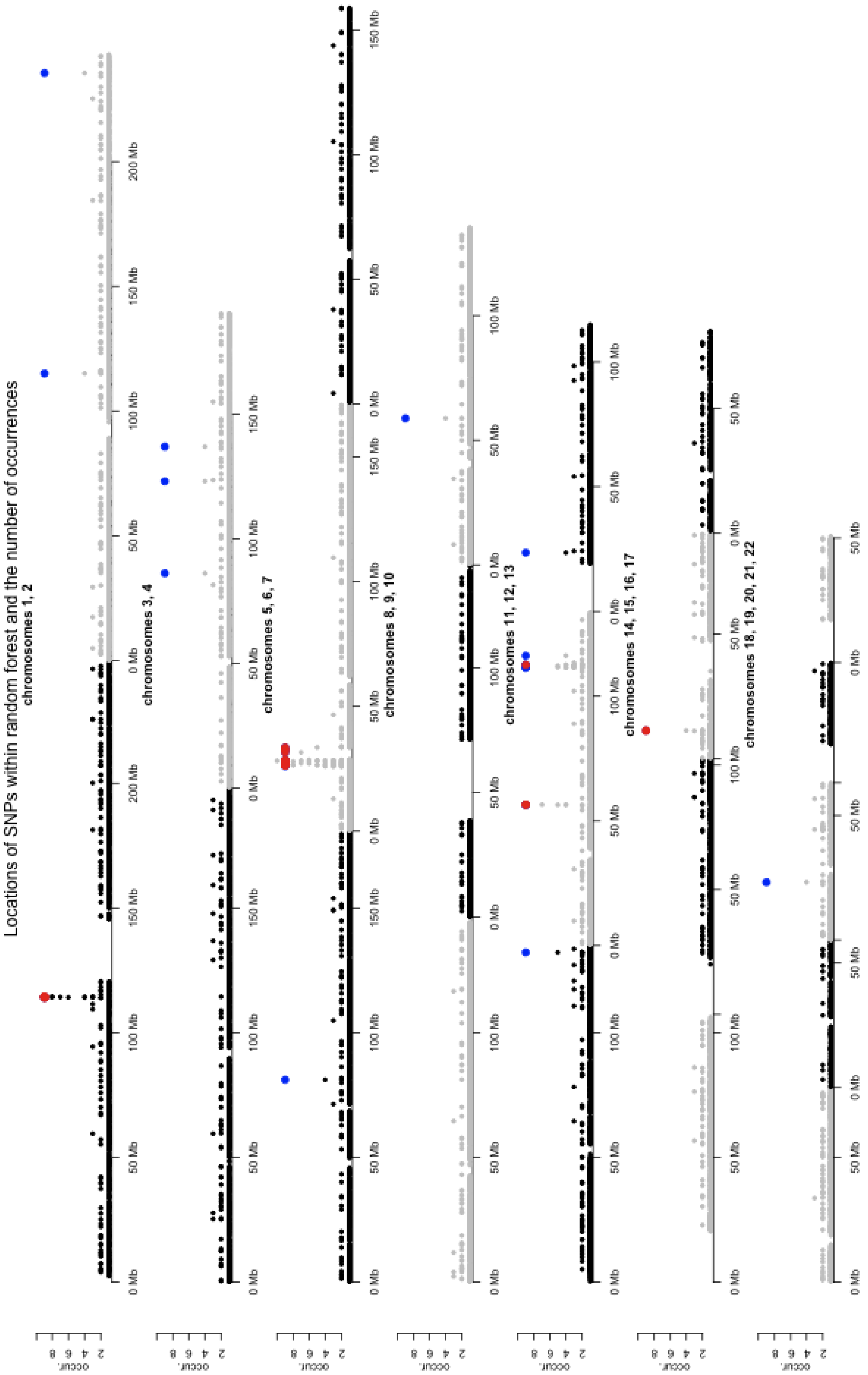


Figure 6.14: Locations of SNPs in random forest and the number of occurrences of a SNP within forest. Red dots indicate SNPs with p-values smaller than 10^{-8} in WTCCC study, blue dots correspond to the loci with high number of occurrences within random forest.

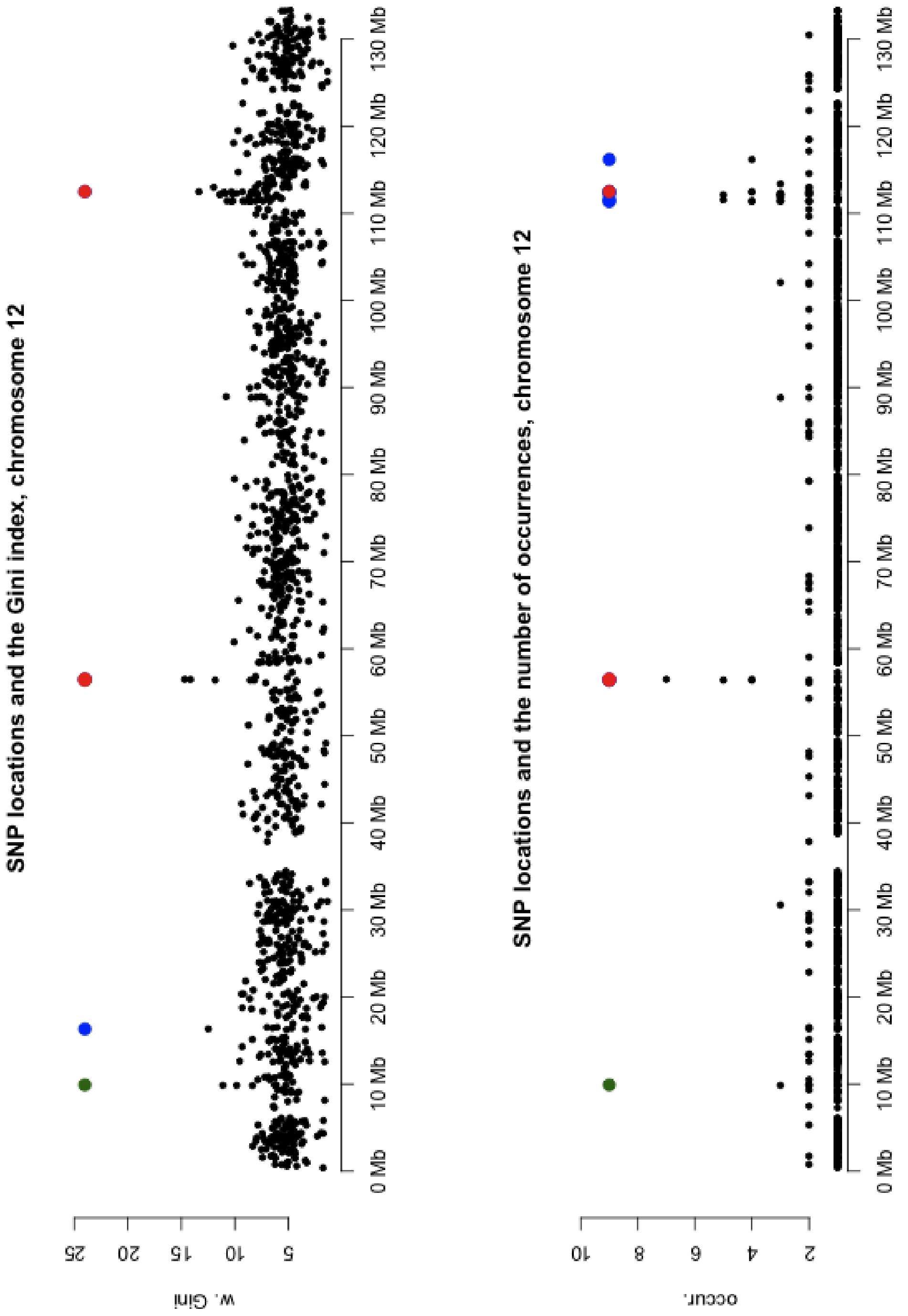


Figure 6.15: Associations on chromosome 12: comparison of different approaches towards evaluating the significance of a locus. The colours of the dots are as in previous plots.

densely here as they are within other association regions on chromosome 12, and the region is rather broad, starting with a small peak around 10Mb. There is a notable area of important SNPs at 23Mb on chromosome 13. As for the remaining chromosomes, there is a moderate evidence of association on chromosome 15 at 90Mb.

It is interesting to compare the set of associations found by random forest algorithm not only with the initial analysis of WTCCC data set, but also with subsequent meta-analyses that incorporate the case/control data from WTCCC. The meta-analysis done by (Barrett et al., 2009) was chosen for the comparison due to the large amount of novel loci yielded by this analysis. In total 24 novel loci were proposed and 18 of them were validated by a replication study (Barrett et al., 2009).

There are a few intriguing overlaps between peaks yielded by random forest method and the meta-analysis. The tightest overlap can be observed on chromosome 10 at 90Mb. Interestingly enough, the p-value obtained from the analysis of the WTCCC data set alone is as large as 1.4×10^{-3}. Even association p-values yielded by separate analysis of other data sets participating in the meta-analysis are smaller. Another case of closely located peaks can be seen on chromosome 12. The region from 10Mb to 15Mb encompasses both an association proposed by meta-analysis and also an association yielded by random forest algorithm. Slightly more distant are peaks on the chromosome 6 that do not belong to the MHC region: the meta-analysis has highlighted the region around 127Mb, while the current random forest analysis has yielded a cluster of SNPs close to 140Mb. Despite the fact that there are no other coincidences with the results of meta-analysis, the set of overlapping peaks suggests that on some occasions random forest approach can be equally powerful to the meta-analyses, even in cases when the p-value from the standard additive model is rather insignificant.

The analysis of locations of SNPs with large number of occurrences in random forest confirms most of the association regions already seen in the Fig.6.13 showing the improvement in Gini impurity. The most important SNPs are even more distinct from the background in Fig.6.14, among them MHC region, region on chromosome 1 around 114Mb and another region on chromosome 12 around 112Mb. Previously known association region on chromosome 16 at 11Mb appears to be less important by the number of occurrences than by the Gini index, though it is still noticeable in Fig.6.14.

Not all moderate associations according to the improvement in Gini impurity reemerge in Fig.6.14 according to the number of SNP occurrences. For example, moderate associations on chromosomes 1, 6, 8, 12 and 15 are lost in Fig.6.14. Nevertheless, associations on chromosomes 4, 11 and 13 clearly hold according to the number of occurrences, with an exception of the association around 4Mb on chromosome 11 that was confirmed by Gini index only. In addition, there are several novel association regions on chromosome 2, around 115Mb and 235Mb, on chromosome 4 at 86Mb and on chromosome 5 (81Mb). In comparison with Fig.6.13 the peak on chromosome 10 is shifted from 90Mb to 60Mb (most probably these are two different regions). Yet another novel association can be observed on chromosome 20 at 23Mb.

The example of chromosome 12 in Fig.6.15 illustrates how both approaches towards ascertaining the SNP significance relate to each other. There is no difference between both criteria in replicating the strongest associations. However, there can be subtle discrepancies between the sensitivity of these methods towards detecting more moderate associations. Although there is a peak at the locus identified by meta-analysis in both plots within Fig.6.15, the association seems to be weaker according to the number of occurrences.

Table 6.3: SNPs with at least 4 occurrences in random forest: location and type of a SNP

SNP name	Location	Gini	Occ.	SNP type	closest gene
rs17013326	1, 114089316	9.72	4	intronic	MAGI3
rs1113523	1, 114129474	11.18	6	intronic	MAGI3
rs1230658	1, 114218451	13.94	8	intronic	MAGI3
rs1230649	1, 114244177	14.69	7	intronic	PHTF1
rs6679677	1, 114303808	27.41	17	upstream	RP11-426L16.1
rs1217396	1, 114338236	11.42	4	intronic	RSBN1
rs2488457	1, 114415368	13.45	8	within non coding gene	AL137856.3
rs3132057	2, 115023640	8.72	4	upstream	AC016683.1

continued on the next page. . .

Table 6.3 – continued

SNP name	Location	Gini	Occ.	SNP type	closest gene
rs10167927	2, 235550321	7.80	4	intergenic	UGT1A8
rs425196	4, 86255297	10.71	4	intergenic	NKX6-1
rs10015924	4, 123019620	12.86	4	intergenic	TNIP3
rs17048453	4, 137005752	7.55	4	intergenic	U6
rs1566630	5, 81207413	9.55	4	intergenic	RASGRF2
rs523383	6, 25869848	8.26	4	intronic	SLC17A3
rs9379851	6, 26354780	12.44	4	upstream	LRRC16A
rs6933583	6, 26355283	6.14	4	upstream	LRRC16A
rs9393708	6, 26362643	15.48	5	upstream	LRRC16A
rs9358932	6, 26362705	8.81	4	upstream	LRRC16A
rs10456045	6, 26404958	13.81	8	intronic	BTN3A1
rs742090	6, 26415637	11.48	8	downstream	LRRC16A
rs7763910	6, 26472655	10.66	5	intronic	BTN2A1
rs4358615	6, 26998567	11.17	6	downstream	HIST1H1PS2
rs12190473	6, 27024687	13.96	7	downstream	HIST1H4B
rs7451149	6, 27057518	12.36	4	intergenic	HFE
rs200481	6, 27773832	16.12	8	upstream	AL591044.3
rs201002	6, 27808192	13.19	9	upstream	AL133255.3
rs200991	6, 27815494	13.84	6	intergenic	AL133255.3
rs149946	6, 27970031	11.55	5	intergenic	AL590062.3
rs149970	6, 27980220	10.50	4	upstream	AL590062.3
rs9393881	6, 28023751	14.30	7	upstream	AL590062.4
rs203884	6, 28077374	15.33	10	intergenic	U2
rs17711344	6, 28077602	9.95	5	intergenic	U2
rs1225709	6, 28103648	8.02	4	intronic	AL358933.2
rs1233704	6, 28166923	12.28	7	intergenic	AL121934.1
rs1233699	6, 28169158	7.27	4	intergenic	AL121934.1
rs17720293	6, 28214698	10.95	5	intronic	ZKSCAN4
rs2859365	6, 28391465	10.37	6	intergenic	ZNF184
rs2523691	6, 31420687	15.33	7	within non coding gene	HCP5
rs9272723	6, 32609427	109.62	23	intronic	HLA-DQA1
rs10807124	6, 33404064	11.14	5	intronic	SYNGAP1
rs7922854	10, 58750903	7.13	4	intergenic	ZWINT
rs1065	11, 132273416	9.63	5	intergenic	OPCML
rs1873914	12, 56379427	8.55	4	intronic	RAB5B

continued on the next page. . .

Table 6.3 – continued

SNP name	Location	Gini	Occ.	SNP type	closest gene
rs705702	12, 56390636	11.83	5	upstream	NEUROD4
rs11171739	12, 56470625	14.13	4	upstream	OR9K1P
rs2292239	12, 56482180	14.69	7	intronic	ERBB3
rs886125	12, 111365324	10.32	4	intergenic	GIT2
rs4766442	12, 111401693	8.57	4	intergenic	GIT2
rs10774613	12, 111546165	7.70	5	intronic	CUX2
rs659964	12, 112130199	11.40	5	intronic	ACAD10
rs7114	12, 112460749	9.28	4	3 prime UTR	ERP29
rs4767293	12, 112463296	10.33	4	downstream	CUX2
rs17696736	12, 112486818	13.34	9	intronic	C12orf30
rs1965297	12, 116173802	5.09	4	intergenic	AC026765.1
rs9634385	13, 23638279	10.73	4	intergenic	AL136962.2
rs12924729	16, 11187783	10.79	4	intronic	CLEC16A
rs844888	20, 23106737	8.65	4	within non coding gene	RP4-737E23.1

The coincidence of association regions according to the number of SNP occurrences with those yielded by the meta-analysis by (Barrett et al., 2009) is less pronounced than the coincidence with Gini impurity peaks. Therefore the results from the meta-analysis are not highlighted in Fig.6.14. The loss of correlation could be explained by the relationship between WTCCC p-values and the number of occurrences within random forest (see Fig.6.12). As the meta-analysis associations that were discovered by Gini impurity have insignificant p-values in the initial analysis, it is of no surprise that these locations were not significant due to the number of occurrences.

It is interesting to look at the candidate genes located near associated SNPs and their possible functional roles. In Table 6.3 the gene that is closest to the associated locus is depicted in a separate column. Significant SNPs on chromosome 1 are very close to the association region known before and encompassing PTPN22 gene. There are two distinct regions on chromosome 2 included in the Table 6.3. The first of these regions hits the locus of non-coding RNA. The second region is slightly more interesting, because it is located nearby gene coding UDP-glucuronosyltransferase 1-8 precursor. CTLA4 gene is not very close to any of the SNPs on chromosome 2 emerging from the current study. Also chromosome 4 has three association regions that are not too far from each other. The middle region practically replicates association with genes IL2-IL21 and has gene coding TNFAIP3-interacting protein 3 nearby. Close to the first region on chromosome 4 there is a gene coding NKX6-1 homeobox protein Nkx-6.1. Chromosome 5 yields a single candidate gene coding Ras-specific guanine nucleotide-releasing factor 2. This region is clearly distinct from chromosome 5p13 yielding associations both with T1D and Graves' disease. A multitude of associations on chromosome 6 can be divided into two broad blocks, the first being related to the well known HLA-DQA1 gene and another spanning over the histone cluster next to the MHC region. The associated SNP on chromosome 10 is located close to a gene coding ZW10-interacting protein 1 and far from previously reported moderate/suggestive association on 10p15 with both T1D and RA. Quite unexpectedly, association on chromosome 11 hits gene coding OPCML neurotrimin precursor. At the same time insulin locus INS on chromosome 11 is not included into association list. Previous WTCCC study was not able to replicate association with INS locus as well due to poor coverage of SNP genotyping. SNPs on chromosome 12 can be divided into two distinct regions, where the most interesting gene within the first region is that which is coding neurogenic differentiation factor 4 (the associated SNP is located upstream from this gene). This region replicates known association on chromosome 12q13. The second cluster of SNPs on chromosome 12 is situated next to another association discovered previously, namely that on chromosome 12q24. The association on chromosome 13 is novel though not very interesting, because it hits intergenic region without any genes nearby. The region on chromosome 16 is a replication of an association within chromosome 16p13 discovered in WTCCC study. At last a novel region on chromosome 20 is located within a non coding gene.

By taking the genes within 20KB of SNPs with at least 4 occurrences in random forest, I then applied two different functional classification systems; the DAVID system (DAVID) and the g:profiler system (g:Profiler; Reimand et al., 2007).

The most significant functional cluster consisted of genes from the MHC region, which have previously been reported to be associated with T1D risk (Cooper et al., 2008; Todd et al., 2007; Wellcome Trust Case Control Consortium, 2007). The p-values for the genes inside this cluster reached value as high as 4.6×10^{-8} and the enrichment score was 4.95. MHC molecules have a significant role in specific immunity and in certain types of autoimmunity, because they present antigens to the T cells. The presentation is done via T cell receptors. After successful T cell receptor cross-linking and antigen presentation T cell activation can take place, where activation means proliferation and production of certain cytokines (IL-2 and others).

The second strongest functional association (highest p-value 4.02×10^{-5}, enrichment score 2.46) was to chromatin related function, but this is due almost entirely to a histone cluster sitting close alongside the MHC region in chromosome 6. Nucleosome assembly, chromatin assembly, DNA packaging, protein-DNA complex assembly, methylation, DNA-binding and some other similar functions are included in this enrichment set. The histone cluster mentioned above seems to be involved in gene regulation via histone modifications. The amount of MHC molecules available for capturing and presenting antigens can be regulated thus. The regulatory aspect could explain the emergence of this function in the context of the genetic associations with T1D. Interestingly enough, the gene with the smallest p-value within this cluster is associated with systemic lupus erythematosus (SLE) risk. SLE is characterized by the immune damage mediated by type III hypersensitivity. This type of hypersensitivity is caused by autoantibodies produced by B2 cells. B2 cells work together with T cells and incorporate high affinity IgG antibodies, which are narrowly specialized. Thus, despite different aetiology of T1D and SLE, some of the disease causing mechanisms (specialized antibodies and involvement of T cells) seem to be related and therefore might yield similar functional enrichments.

A number of other associations, such as a structural propensity for immunoglobulin domains are also driven by this large association on the boundary of the MHC. Immunoglobulins IgG, IgM, IgA, IgD and IgE are part of B lymphocytes and determine the specificity of binding for B cell to antigen. This is directly related to the structural propensity for immunoglobulin domains, a function for which the cluster is enriched. B cells typically present antigens to CD4+ T cells via MHC. CD4+ T cells in turn are involved into the autoimmune destruction mechanism characteristic for T1D (Gorodezky et al., 2006). This seems to be the link between this functional enrichment cluster and associations with T1D.

An intriguing additional association was to neuronal/glial development (4 genes in total, spread over different chromosomes). The enrichment score for this cluster was 1.52 and the largest p-value for a gene was 0.0043 making it the third most significant functional category. Genes within this cluster are related to glial cell differentiation, gliogenesis, neuron differentiation and cell surface receptor linked signal transduction. Defects in gliogenic pathways are characteristic for multiple sclerosis (MS). MS is typically defined by the depletion of myelin that is necessary for axon insulation and defects in myelin sheaths around axons. The depletion is caused by T cells attacking myelin cells according to the autoimmune pathway. It can be hypothesized that this functional cluster encompasses genes and autoimmune pathways that are shared by T1D and MS. Genetic factors in common for MS and T1D are well known. For example, MHC region encompasses variants increasing risk of both T1D and MS (Alcina et al., 2009). This is not surprising considering the role of T cells in the aetiology of both diseases. Also IL2RA gene has distinct causative variants for T1D and MS (Alcina et al., 2009; Lowe et al., 2007; Qu et al., 2007b; Vella et al., 2005). A shared allele within CD226 gene is related to both diseases as well (Hafler et al., 2009).

Comparison of results obtained by random forest algorithm and stepwise model selection shows that both methods are good at replicating known associations. Of course, this statement is true if the specific data set permits to discover them by sufficient genotyping density (see discussion about INS locus). Random forest method produces more variation within discovered associations, however it lacks the precision of stepwise model selection. It can be seen both from the replicated associations and the fact that novel associations hit intergenic regions and non-coding genes. As the random forest algorithm uses individual subsampling and smaller subsets of SNPs, hitting non-causative variants within linkage disequilibrium by chance can be expected. Thus, it can be suggested as an efficient and fast method to discover novel candidate loci, however, more detailed follow up studies are necessary or, alternatively, other statistical methods should be used to postprocess the output of random forest algorithm.

Chapter 7

Conclusions

The thesis is devoted to the understanding of the genetic components underlying various phenotypes. First, traits characterizing early development of fruit fly are studied, with a particular focus upon oogenesis and nurse cell development throughout oogenesis. Second, the disease aetiology of type 1 diabetes in humans is elucidated using various model selection approaches and case/control data from WTCCC.

The quantitative characterization of fruit fly development requires deciding upon the method of the phenotype collection and then upon the approach towards phenotype measurement. Trait quantification is a particularly subtle task, because the nearly isogenic lines studied here are all wild type and therefore do not exhibit severe mutations or easily distinguishable phenotypes. A novel approach towards defining developmental traits is proposed, which is based on the optical microscopy techniques and a pipeline of image analysis algorithms. A set of suitable metrics for the quantitative characterization of a phenotype is suggested, most of them based on rotationally invariant image moments, for instance Zernike moments, and others describing geometric features and texture.

The association study for the oogenesis phenotypes is among the first ones for *Drosophila melanogaster* that is making use of nearly 6 million SNPs now available from (Drosophila population genomics project, 2010). Two types of models are used for linking the genotype and phenotype information, namely the non-parametric Mann-Whitney test and the mixed model nested Anova test that takes into account the hierarchical structure of trait measurements (levels of a single line, image and nurse cell). The results yielded by the non-parametric Mann-Whitney test imply that the number of fly lines is insufficient for a test that ignores several levels of nested trait measurements, because there is strong evidence for low statistical power. The results obtained by the parametric Anova models show that this approach has much more statistical power than the Mann-Whitney test. However, some false positives were noticed among the highest association peaks, which turned out to be a consequence of the violations of the model assumptions. To avoid claiming false associations due to the incorrect asymptotic behaviour of the model the most promising candidate loci were subjected to the permutation tests shuffling the line labels. Interestingly enough, traits with at least one association confirmed by the permutation test also had small p-values by the non-parametric Kruskal-Wallis test. In addition, traits with false positive associations due to the violated model assumptions had large Kruskal-Wallis test p-values. It means that the Kruskal-Wallis test can serve as a robust and reliable predictor of the phenotype's potential in the association study.

The association study for early developmental phenotypes in *Drosophila melanogaster* has yielded 5 candidate loci for 5 different traits. Most of the associated traits characterize the distribution of DNA within the cell nucleus, apart from the cell nuclei compactness. The majority of loci discovered either directly hit or are nearby important developmental genes. This finding together with results of permutation tests confirms the significance and validity of the associations discovered. However, the study is still somehow lacking in statistical power to serve as a final proof of the genotype/phenotype relationship. Therefore an extra round of experiments to augment the set of phenotype measurements is suggested and will be carried out within the near future. The aim of the additional work is to analyse 40 new *Drosophila melanogaster* strains coming from the same isogenic population. The analysis itself can be improved by using the newly resequenced *Drosophila melanogaster* genomes, which can provide more accurate SNP information and thus add both to statistical power in association discovery and also to the precision of fine mapping of loci.

The full potential of the phenotypes collected has not yet been exploited. The current study aims at in depth analysis of traits relevant to nurse cells. However, more information can be extracted and additional traits can be defined from *oskar* localization patterns and characteristics of oocyte. As the image analysis pipeline is fully established, adding several new traits is not such a complicated task and will be done in the near future as well.

Type 1 diabetes was selected as the focus of the reanalysis of the WTCCC case/control data set due to the large amount of genetic factors affecting this disease and its complicated aetiology. The factors influencing T1D risk, various model selection techniques and their potential in improving the discovery of the associations of moderate strength were studied.

The association study methods considered in this work exploit the idea of conditional modeling of the effects of a locus upon a disease, where conditioning on other loci already known to affect the disease is used. Composite models incorporating several loci are important when a few distinct variants have to be selected within a linkage disequilibrium region, because conditional genome scans avoid allele patterns similar to those already present in the model. Also the power of association discovery can be increased thus, because moderate associations are enhanced after taking into account the effect of the strongest associations. These properties of conditional methods become even more favourable as the SNP maps become denser.

Two distinct modeling approaches were applied to the WTCCC case/control data, namely stepwise regression and the random forest algorithm. The first of these methods is a deterministic model selection algorithm, while the second method relies on a probabilistic model selection mechanism based on the subsampling of individuals from case and control subsets. Deterministic model selection techniques are known from association studies on model organisms exploiting microsatellite markers and sparse marker maps. However, these methods are traditionally avoided in large scale association studies, mostly due to the computational challenges they present. Probabilistic model selection is explored only within a few association studies.

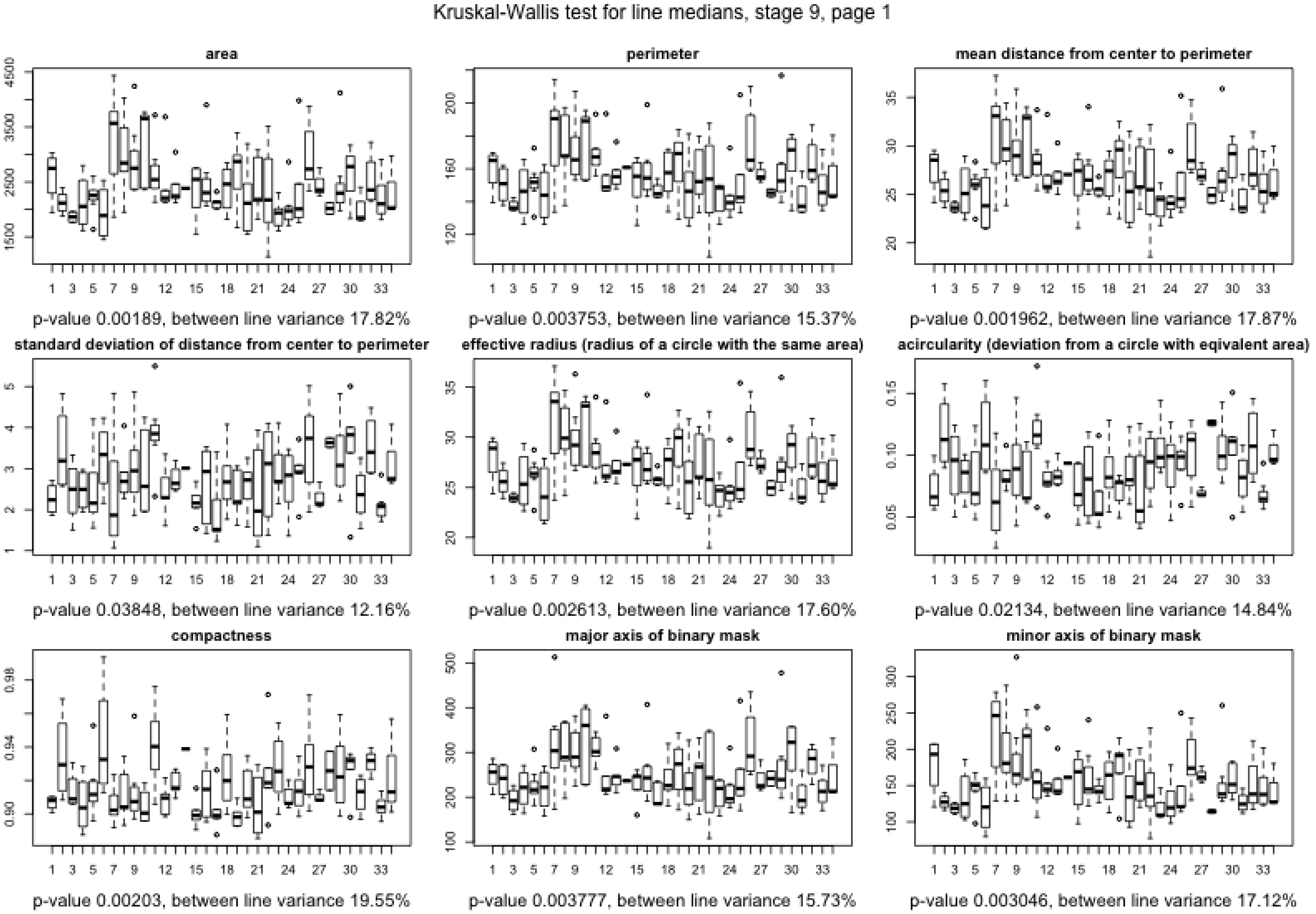
The application of these model selection methods to the analysis of case/control data for type 1 diabetes has led to the conclusion that various methodologies of model building do not influence the success in replication of previous associations considerably, because all models replicated previous associations well. It means that the main factors affecting the results of replication studies are population stratification and specific characteristics of the population studied instead of particular approach taken for statistical analysis. It is worth noting that the random forest algorithm successfully replicated moderate associations as well as the five strongest associations. Apparently, probabilistic model selection is more sensitive towards weaker association signals due to the repeated sampling. However, stepwise model selection exhibits more precision in replicating.

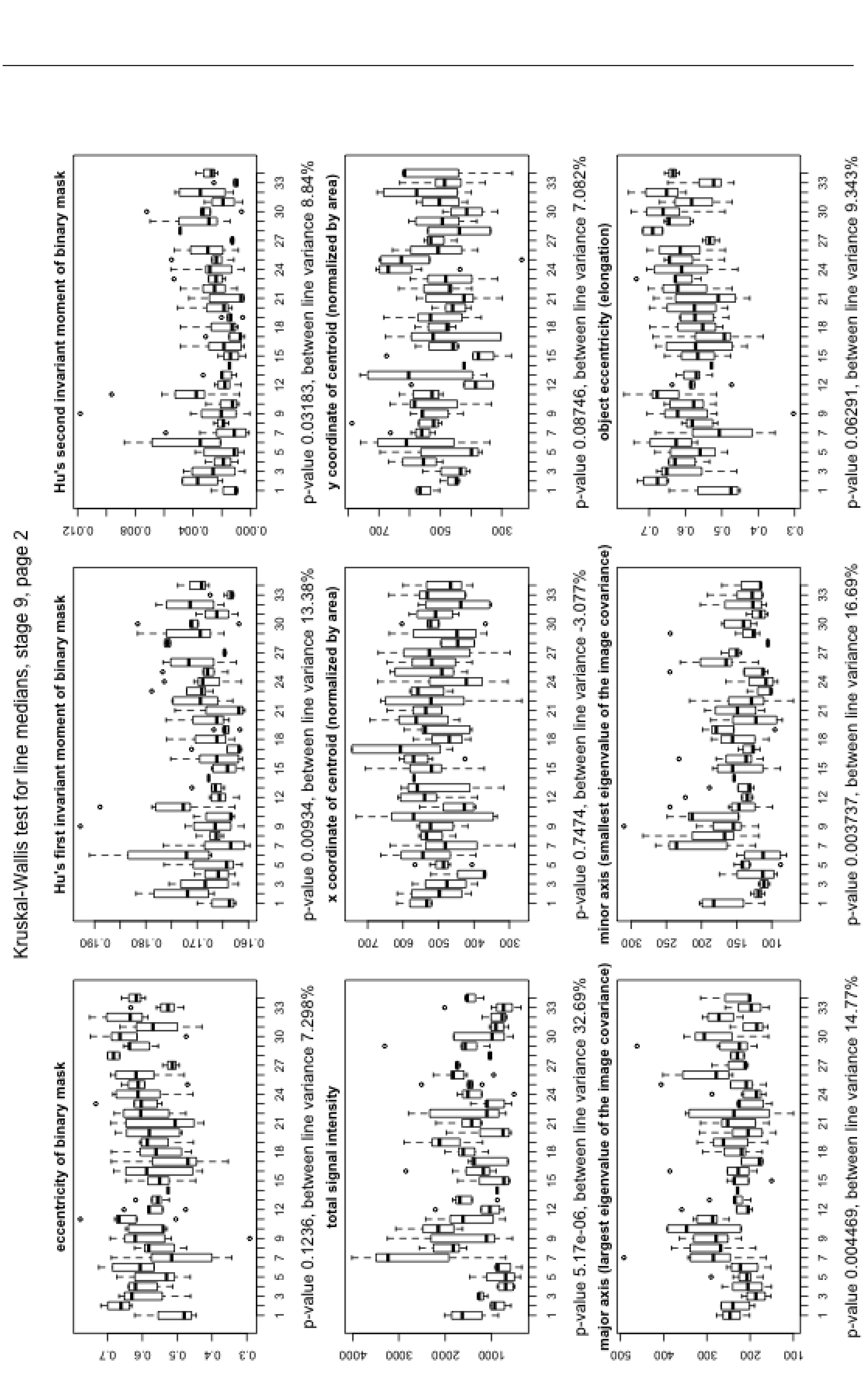
How many novel associations can be added via using particular model selection methods? The stepwise regression added at least 5 novel loci to the replicated associations. All of the new associations exhibit relatively small p-values in the previous analysis of WTCCC data set, which nevertheless are not good enough for these loci to be discovered by a simple additive model. Thus the novel loci have emerged due to increased statistical power after conditioning upon loci already known. The significance of SNPs within the model yielded by the random forest algorithm is rather consistent with the p-values of the previous analysis. However, there are a few novelties, for instance the SNPs on chromosomes 5, 11, 13 and 20. Three association regions that were discovered by the random forest method coincide with the replicated results of the meta-analysis incorporating WTCCC data. It is a good indicator of the sensitivity of the proposed method. Rather surprisingly some of the discovered associations are functionally related to neuronal and glial development. Further explorations of the location of these SNPs and corresponding candidate genes is necessary.

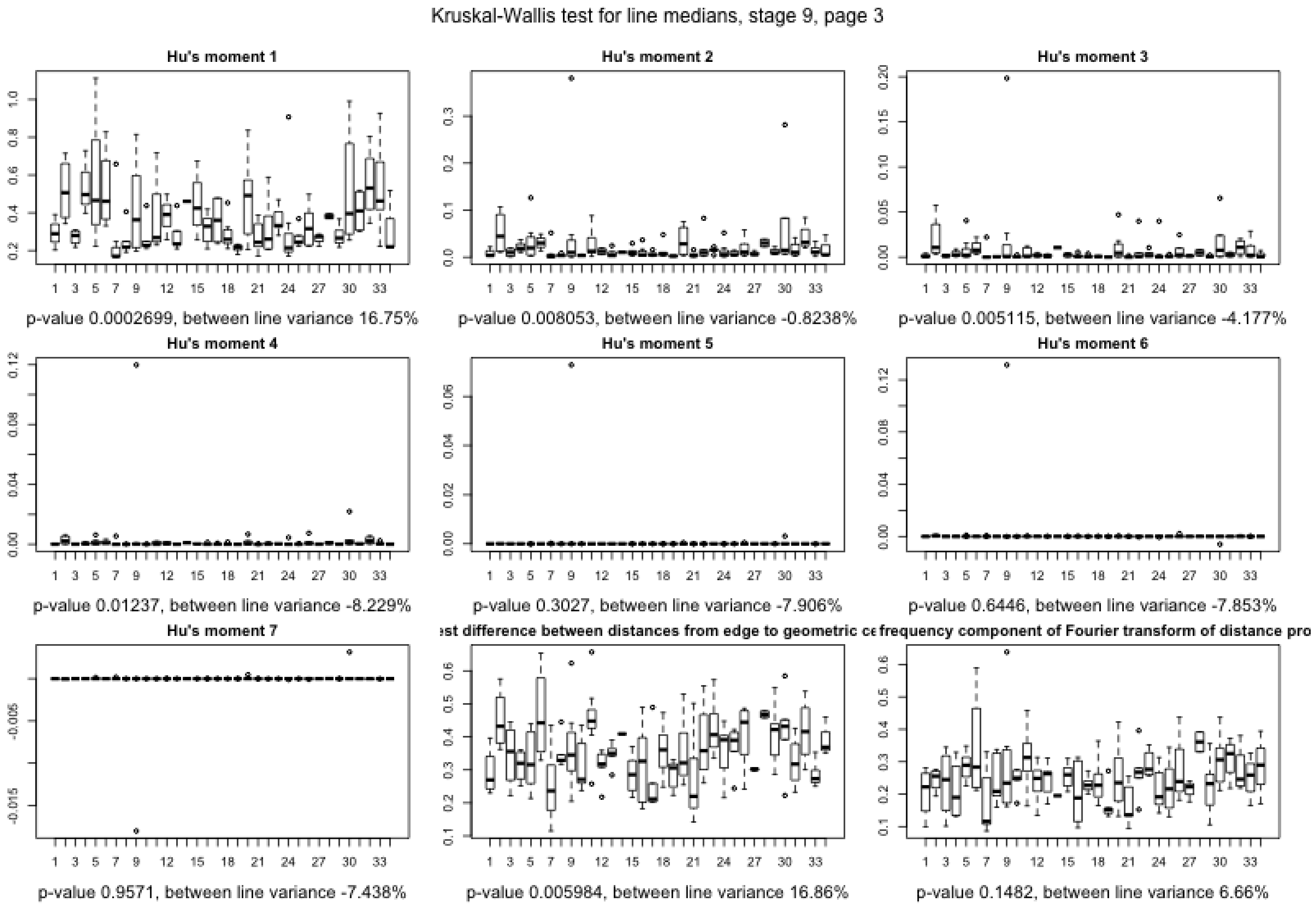
Overall the model selection methods applied to the study of T1D show enough potential in association discovery to be applied to other common diseases within the WTCCC data set. It means that using better models can be helpful in elucidation of more meaning from the same data.

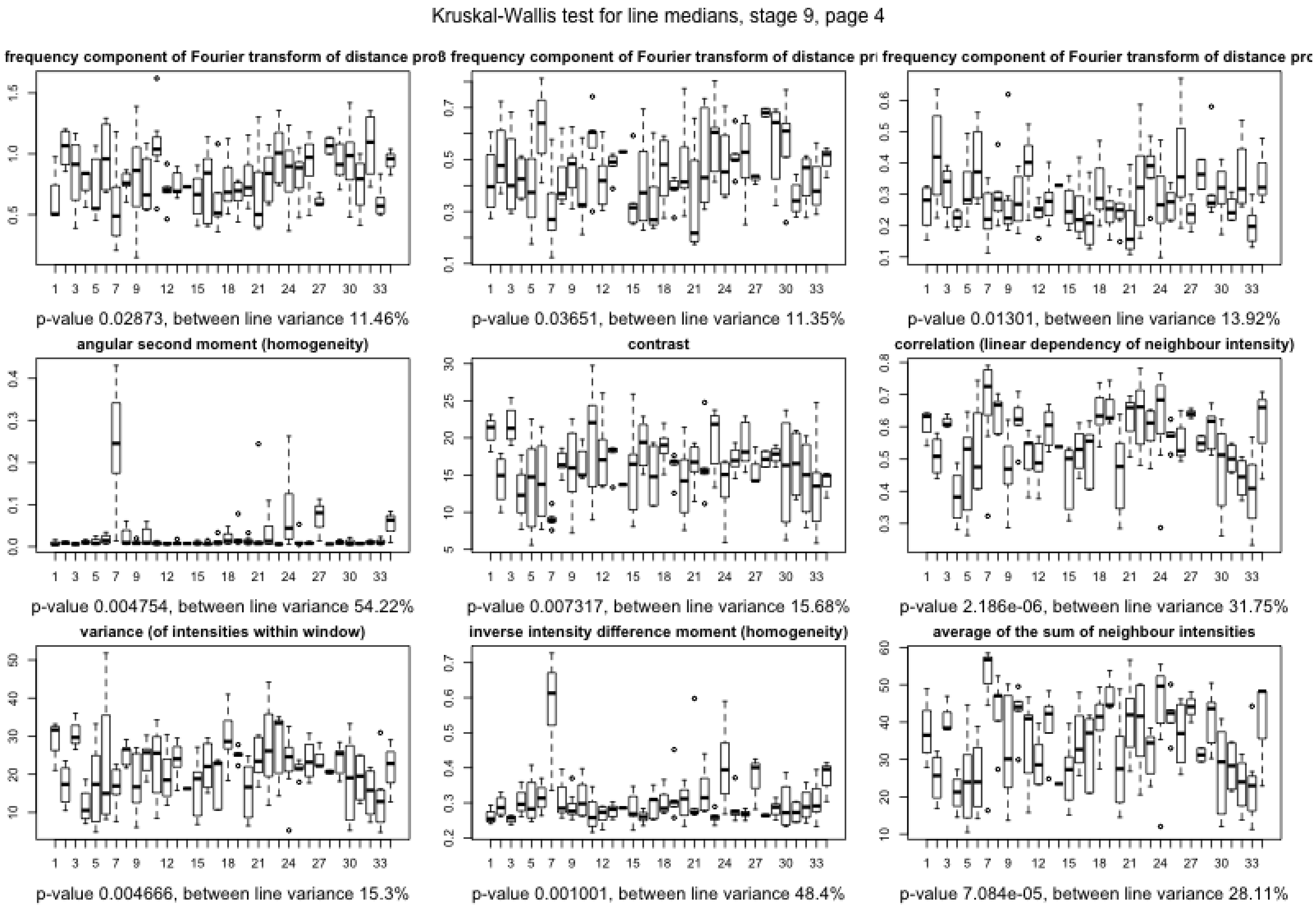
Appendix A

Trait Variability within and between *Drosophila melanogaster* Lines, Split by Developmental Stage

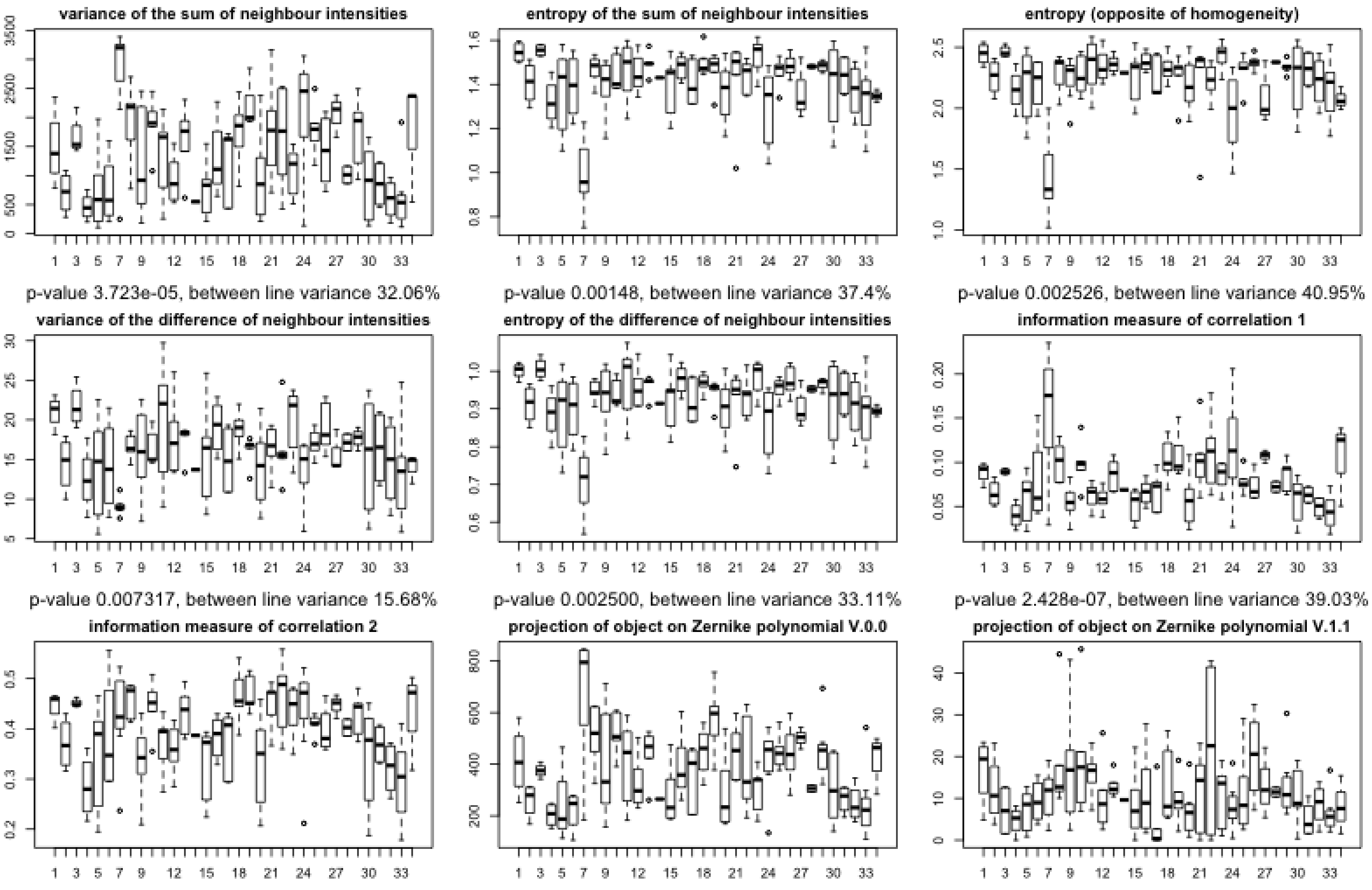

Kruskal-Wallis test for line medians, stage 9, page 1
area
1500
2500
3500
4500
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.00189, between line variance 17.82%
perimeter
120
160
200
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.003753, between line variance 15.37%
mean distance from center to perimeter
20
25
30
35
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.001962, between line variance 17.87%
standard deviation of distance from center to perimeter
1
2
3
4
5
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.03848, between line variance 12.16%
effective radius (radius of a circle with the same area)
20
25
30
35
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.002613, between line variance 17.60%
acircularity (deviation from a circle with eqivalent area)
0.05
0.10
0.15
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.02134, between line variance 14.84%
compactness
0.90
0.94
0.98
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.00203, between line variance 19.55%
major axis of binary mask
100
200
300
400
500
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.003777, between line variance 15.73%
minor axis of binary mask
100
150
200
250
300
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.003046, between line variance 17.12%


Kruskal-Wallis test for line medians, stage 9, page 2
eccentricity of binary mask
0.3 0.4 0.5 0.6 0.7
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.1236, between line variance 7.298%
Hu's first invariant moment of binary mask
0.160 0.170 0.180 0.190
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.00934, between line variance 13.38%
Hu's second invariant moment of binary mask
0.000 0.004 0.008 0.012
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.03183, between line variance 8.84%
total signal intensity
1000 2000 3000 4000
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 5.17e-06, between line variance 32.69%
x coordinate of centroid (normalized by area)
300 400 500 600 700
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.7474, between line variance -3.077%
y coordinate of centroid (normalized by area)
300 500 700
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.08746, between line variance 7.082%
major axis (largest eigenvalue of the image covariance)
100 200 300 400 500
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.004469, between line variance 14.77%
minor axis (smallest eigenvalue of the image covariance)
100 150 200 250 300
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.003737, between line variance 16.69%
object eccentricity (elongation)
0.3 0.4 0.5 0.6 0.7
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.06291, between line variance 9.343%


Kruskal-Wallis test for line medians, stage 9, page 3
Hu's moment 1
p-value 0.0002699, between line variance 16.75%
Hu's moment 2
p-value 0.008053, between line variance -0.8238%
Hu's moment 3
p-value 0.005115, between line variance -4.177%
Hu's moment 4
p-value 0.01237, between line variance -8.229%
Hu's moment 5
p-value 0.3027, between line variance -7.906%
Hu's moment 6
p-value 0.6446, between line variance -7.853%
Hu's moment 7
p-value 0.9571, between line variance -7.438%
est difference between distances from edge to geometric ce
p-value 0.005984, between line variance 16.86%
frequency component of Fourier transform of distance pro
p-value 0.1482, between line variance 6.66%
1 3 5 7 9 12 15 18 21 24 27 30 33


Kruskal-Wallis test for line medians, stage 9, page 4
frequency component of Fourier transform of distance proß
frequency component of Fourier transform of distance pr
frequency component of Fourier transform of distance prc
p-value 0.02873, between line variance 11.46%
p-value 0.03651, between line variance 11.35%
p-value 0.01301, between line variance 13.92%
angular second moment (homogeneity)
contrast
correlation (linear dependency of neighbour intensity)
p-value 0.004754, between line variance 54.22%
p-value 0.007317, between line variance 15.68%
p-value 2.186e-06, between line variance 31.75%
variance (of intensities within window)
inverse intensity difference moment (homogeneity)
average of the sum of neighbour intensities
p-value 0.004666, between line variance 15.3%
p-value 0.001001, between line variance 48.4%
p-value 7.084e-05, between line variance 28.11%
1 3 5 7 9 12 15 18 21 24 27 30 33

Kruskal-Wallis test for line medians, stage 9, page 5



variance of the sum of neighbour intensities

p-value 3.723e-05, between line variance 32.06%

entropy of the sum of neighbour intensities

p-value 0.00148, between line variance 37.4%

entropy (opposite of homogeneity)

p-value 0.002526, between line variance 40.95%

variance of the difference of neighbour intensities

p-value 0.007317, between line variance 15.68%

entropy of the difference of neighbour intensities

p-value 0.002500, between line variance 33.11%

information measure of correlation 1

p-value 2.428e-07, between line variance 39.03%

information measure of correlation 2

p-value 4.156e-06, between line variance 29.43%

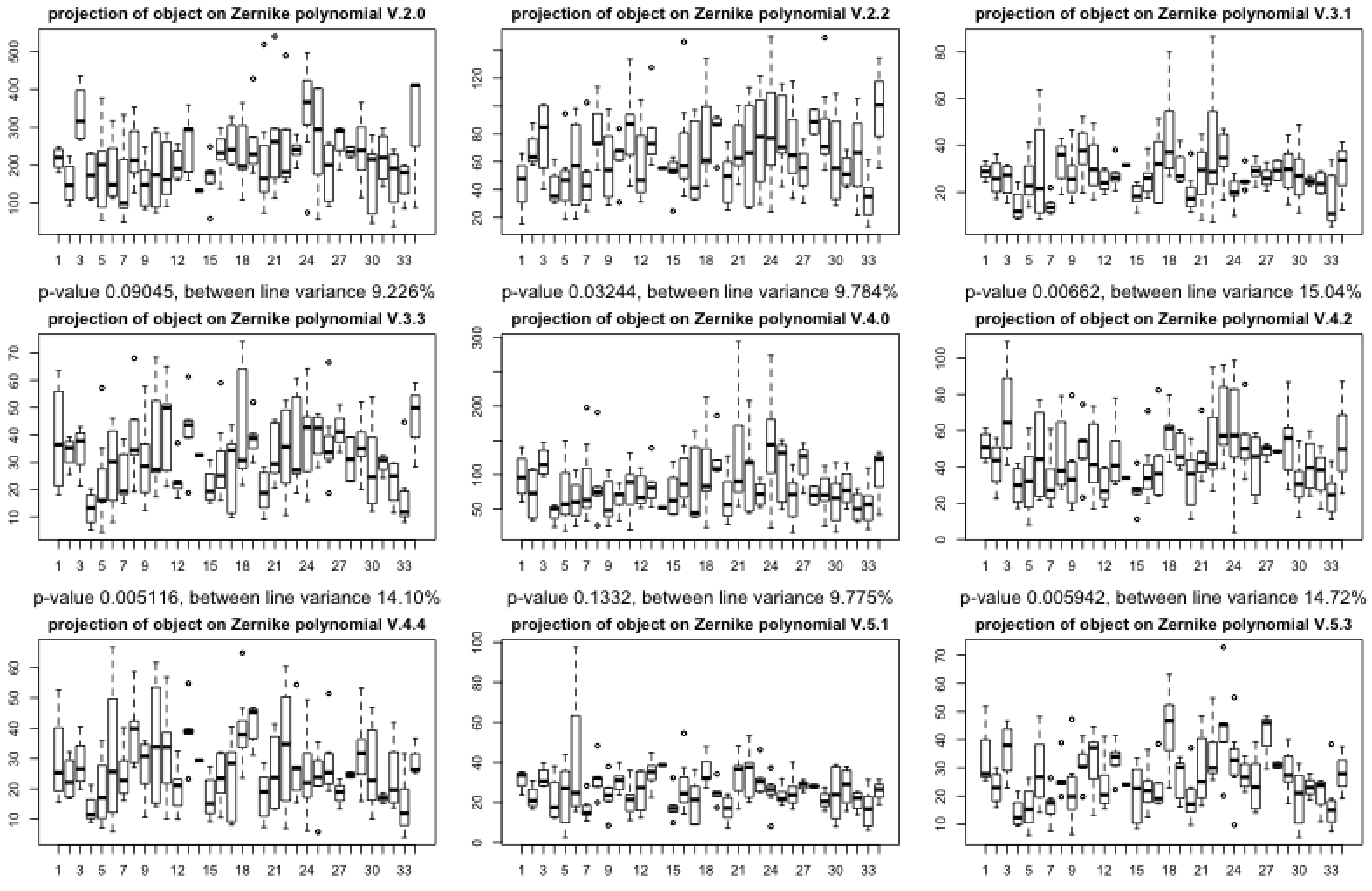
projection of object on Zernike polynomial V.0.0

p-value 1.612e-05, between line variance 33.04%

projection of object on Zernike polynomial V.1.1

p-value 0.1205, between line variance 8.67%

Kruskal-Wallis test for line medians, stage 9, page 6



projection of object on Zernike polynomial V.2.0

p-value 0.09045, between line variance 9.226%

projection of object on Zernike polynomial V.2.2

p-value 0.03244, between line variance 9.784%

projection of object on Zernike polynomial V.3.1

p-value 0.00662, between line variance 15.04%

projection of object on Zernike polynomial V.3.3

p-value 0.005116, between line variance 14.10%

projection of object on Zernike polynomial V.4.0

p-value 0.1332, between line variance 9.775%

projection of object on Zernike polynomial V.4.2

p-value 0.005942, between line variance 14.72%

projection of object on Zernike polynomial V.4.4

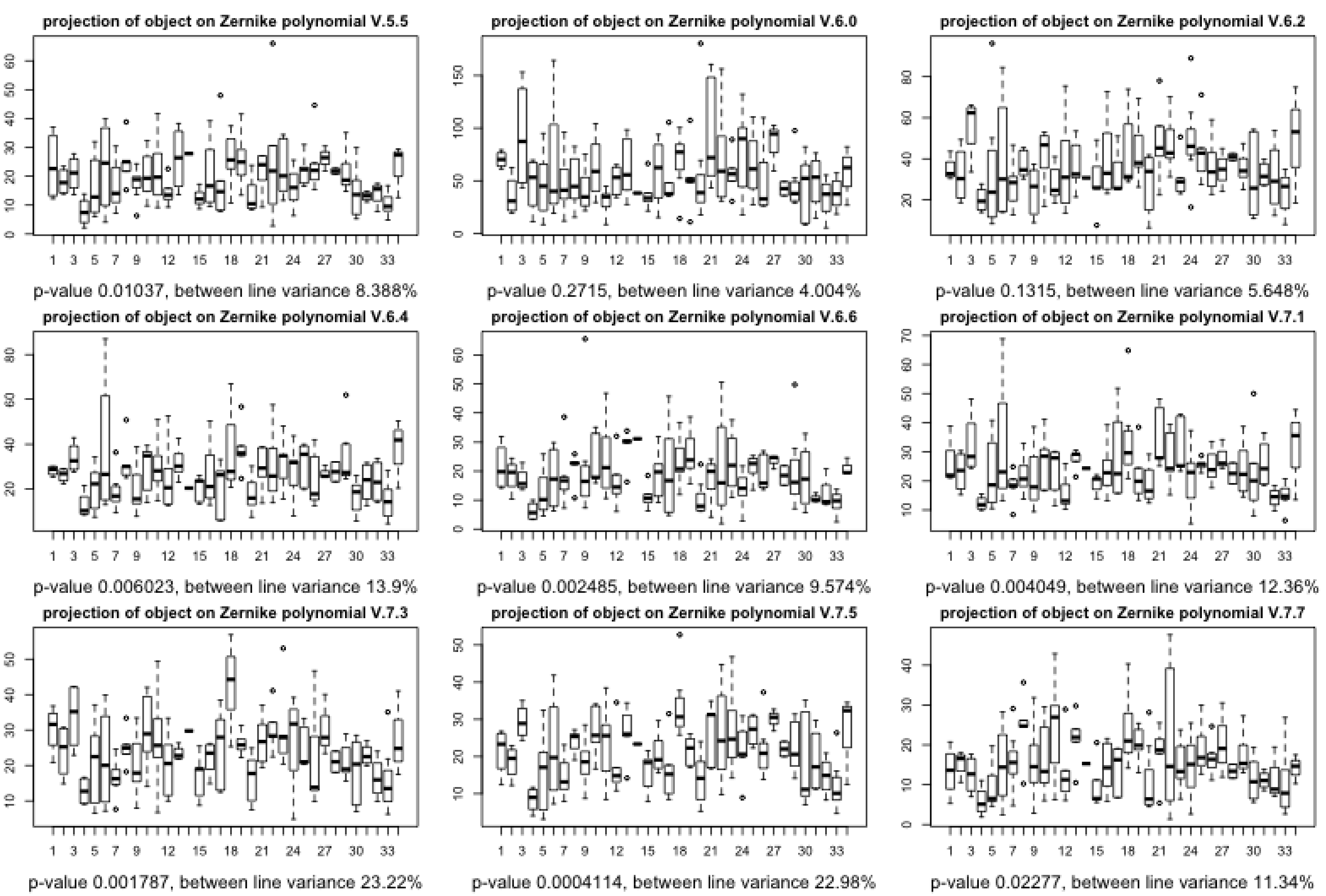
p-value 0.01028, between line variance 14.28%

projection of object on Zernike polynomial V.5.1

p-value 0.002489, between line variance 13.39%

projection of object on Zernike polynomial V.5.3

p-value 5.988e-05, between line variance 29.5%



Kruskal-Wallis test for line medians, stage 9, page 7

projection of object on Zernike polynomial V.5.5

p-value 0.01037, between line variance 8.388%

projection of object on Zernike polynomial V.6.0

p-value 0.2715, between line variance 4.004%

projection of object on Zernike polynomial V.6.2

p-value 0.1315, between line variance 5.648%

projection of object on Zernike polynomial V.6.4

p-value 0.006023, between line variance 13.9%

projection of object on Zernike polynomial V.6.6

p-value 0.002485, between line variance 9.574%

projection of object on Zernike polynomial V.7.1

p-value 0.004049, between line variance 12.36%

projection of object on Zernike polynomial V.7.3

p-value 0.001787, between line variance 23.22%

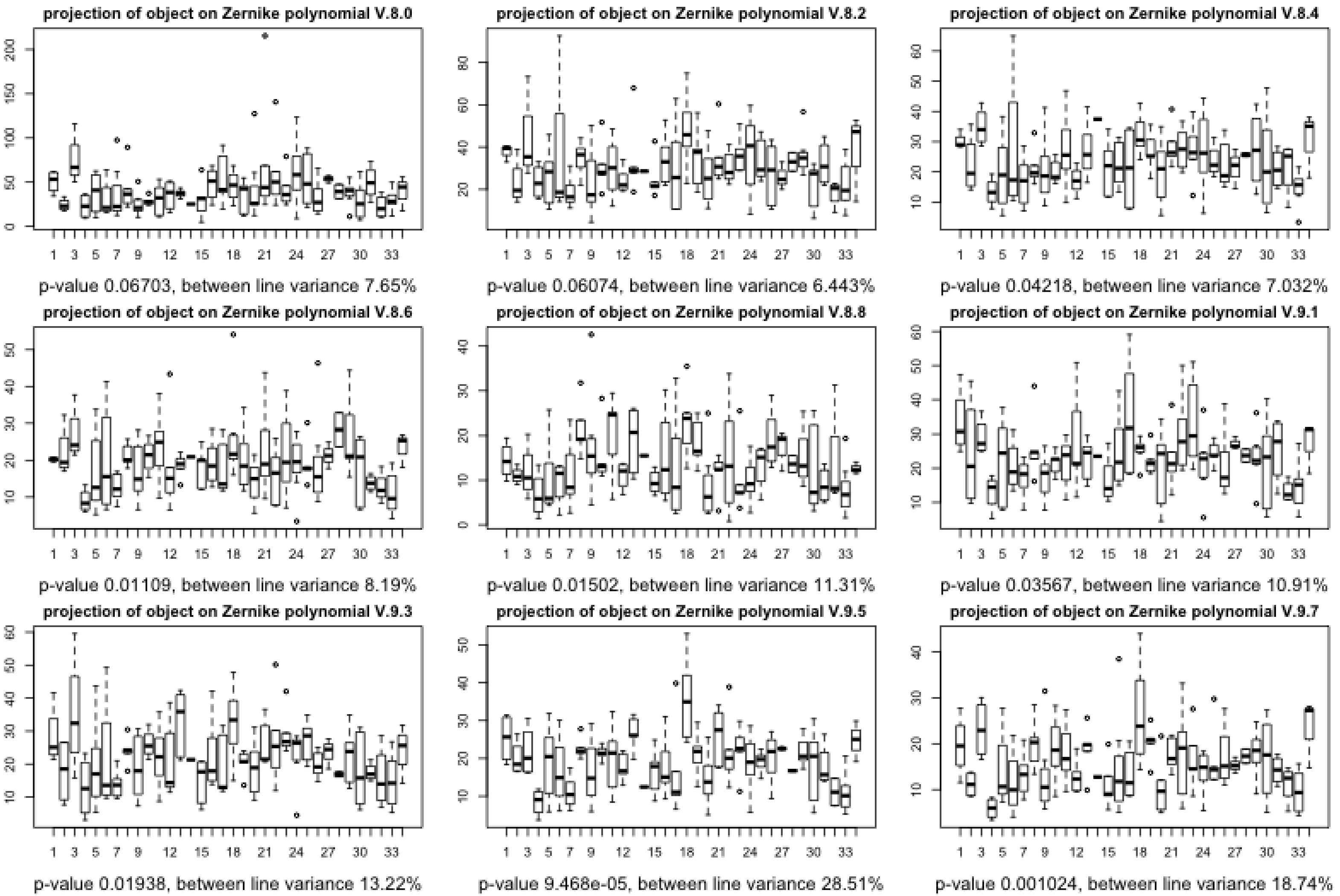
projection of object on Zernike polynomial V.7.5

p-value 0.0004114, between line variance 22.98%

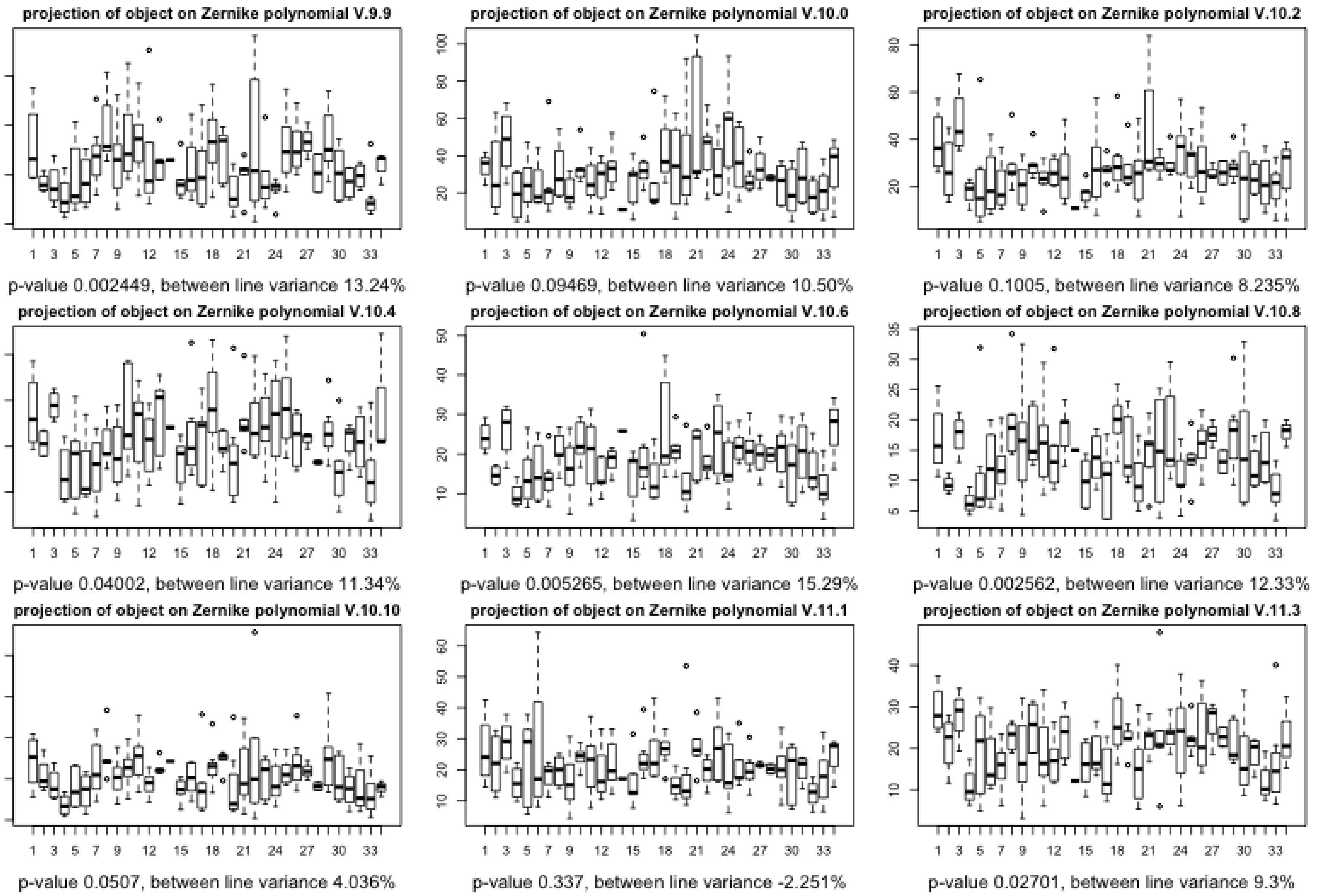
projection of object on Zernike polynomial V.7.7

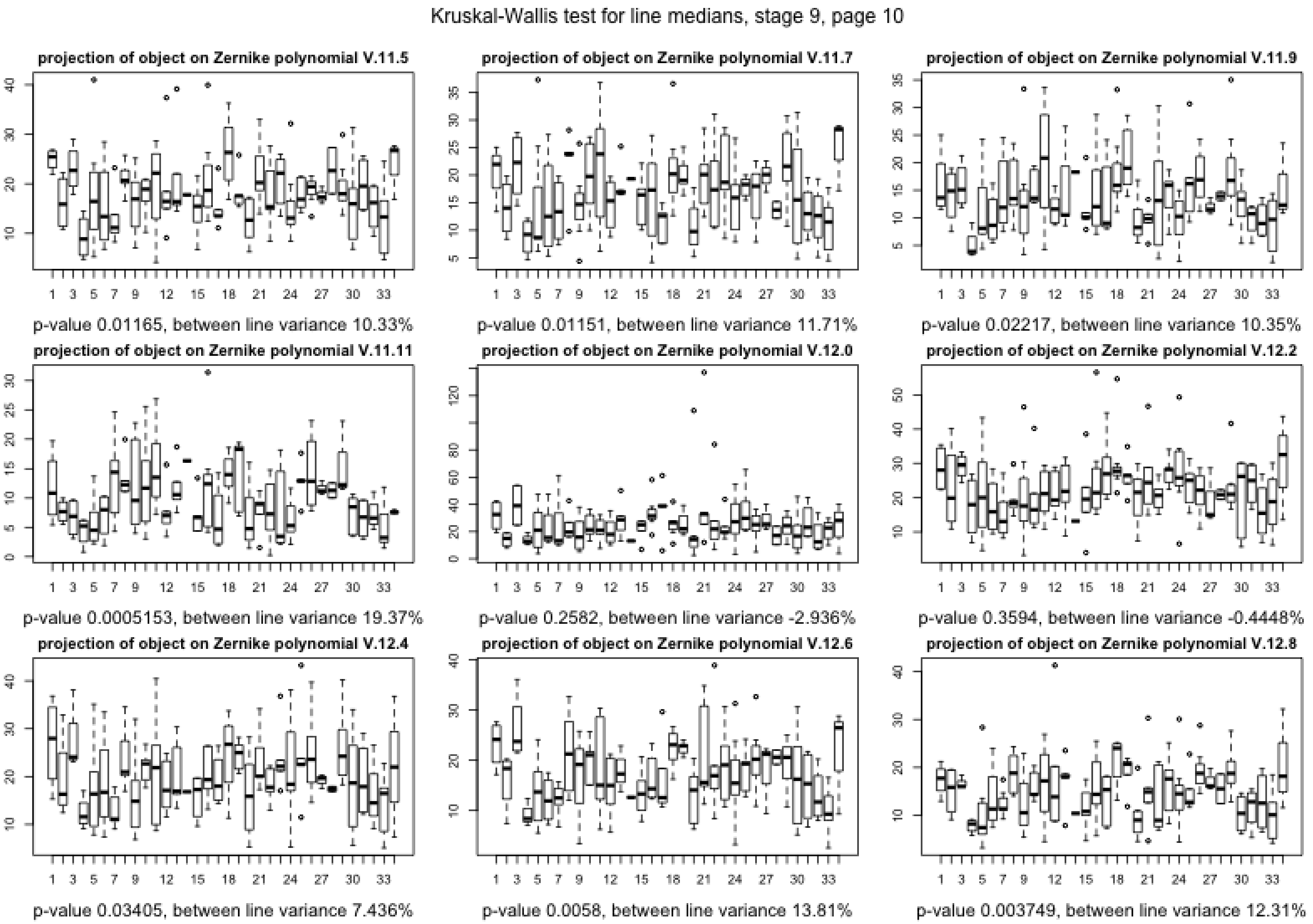
p-value 0.02277, between line variance 11.34%

Kruskal-Wallis test for line medians, stage 9, page 8



Kruskal-Wallis test for line medians, stage 9, page 9





Kruskal-Wallis test for line medians, stage 9, page 10

projection of object on Zernike polynomial V.11.5

p-value 0.01165, between line variance 10.33%

projection of object on Zernike polynomial V.11.7

p-value 0.01151, between line variance 11.71%

projection of object on Zernike polynomial V.11.9

p-value 0.02217, between line variance 10.35%

projection of object on Zernike polynomial V.11.11

p-value 0.0005153, between line variance 19.37%

projection of object on Zernike polynomial V.12.0

p-value 0.2582, between line variance -2.936%

projection of object on Zernike polynomial V.12.2

p-value 0.3594, between line variance -0.4448%

projection of object on Zernike polynomial V.12.4

p-value 0.03405, between line variance 7.436%

projection of object on Zernike polynomial V.12.6

p-value 0.0058, between line variance 13.81%

projection of object on Zernike polynomial V.12.8

p-value 0.003749, between line variance 12.31%

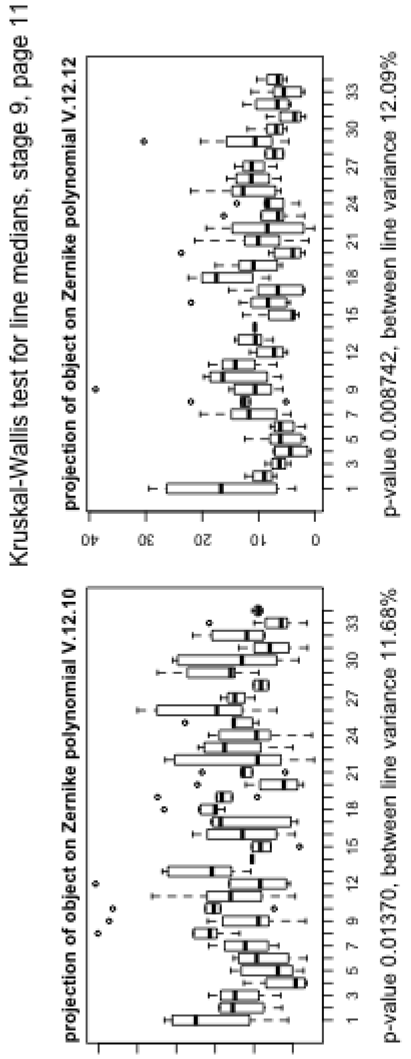
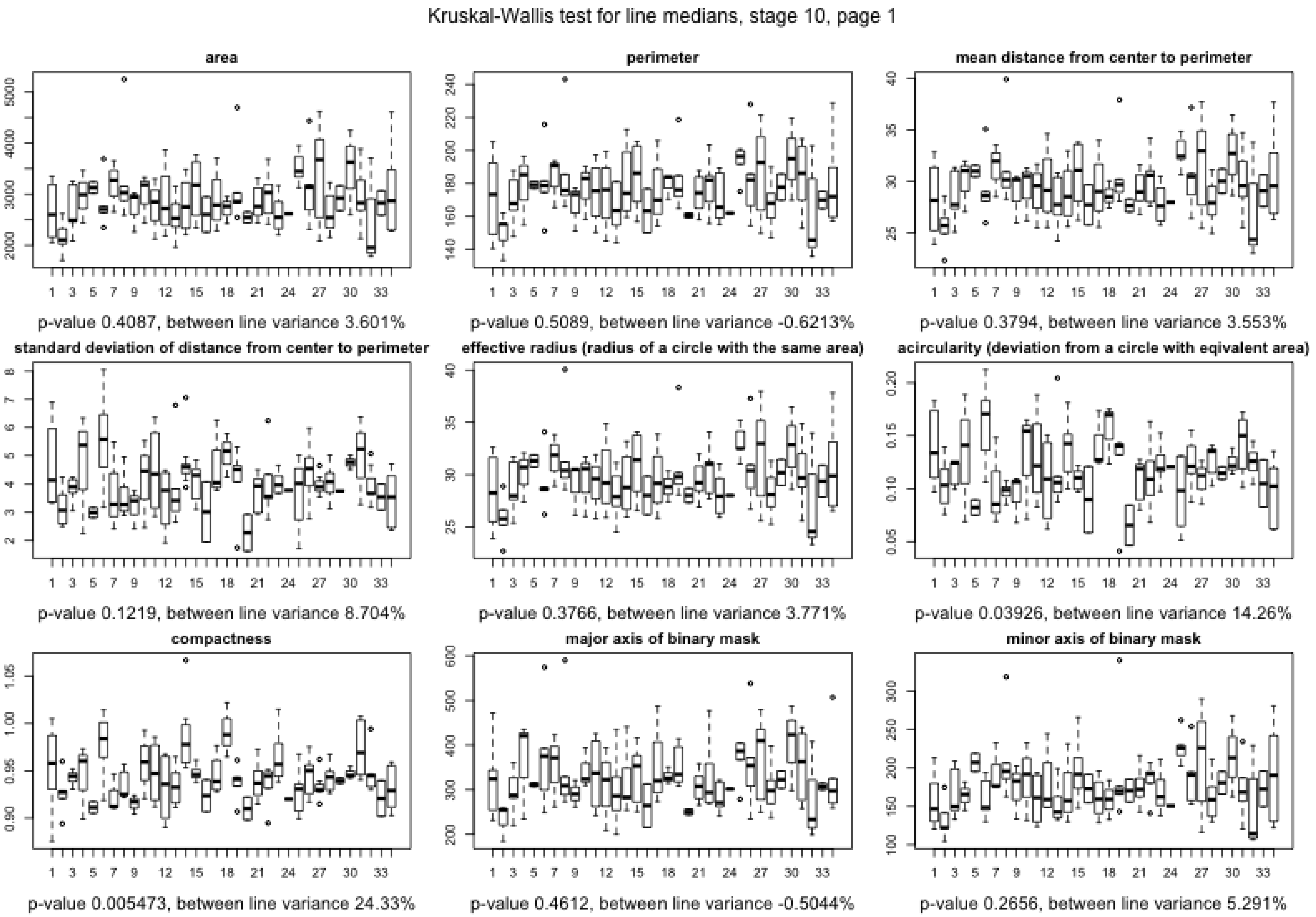


Figure A.1: Traits characterizing nurse cells, stage 9. Numbers on x axis correspond to *Drosophila melanogaster* lines, while y axis denotes trait value. Kruskal-Wallis test p-value and between line variance is indicated below each trait panel.



Kruskal-Wallis test for line medians, stage 10, page 1

area

p-value 0.4087, between line variance 3.601%

perimeter

p-value 0.5089, between line variance -0.6213%

mean distance from center to perimeter

p-value 0.3794, between line variance 3.553%

standard deviation of distance from center to perimeter

p-value 0.1219, between line variance 8.704%

effective radius (radius of a circle with the same area)

p-value 0.3766, between line variance 3.771%

acircularity (deviation from a circle with eqivalent area)

p-value 0.03926, between line variance 14.26%

compactness

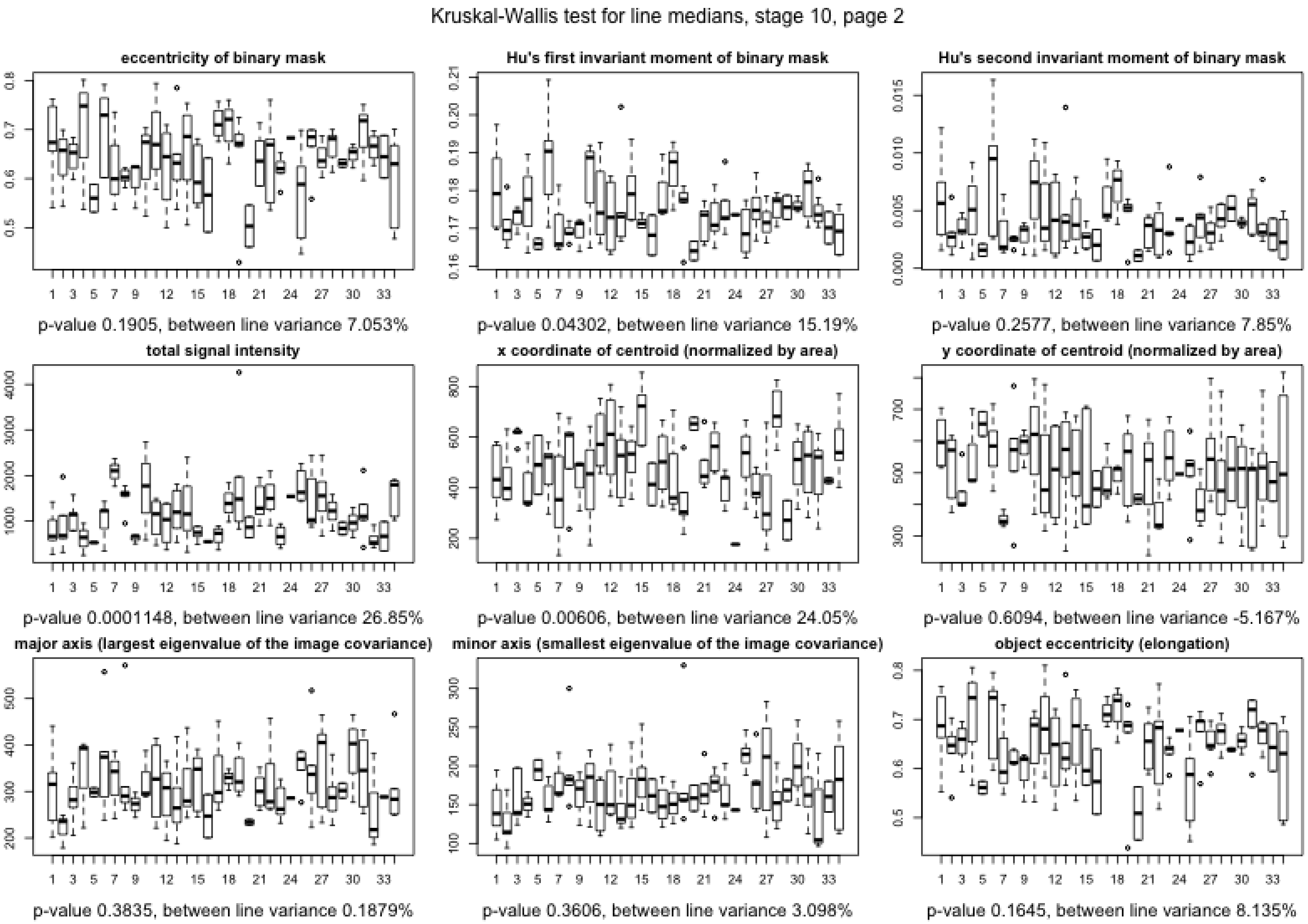
p-value 0.005473, between line variance 24.33%

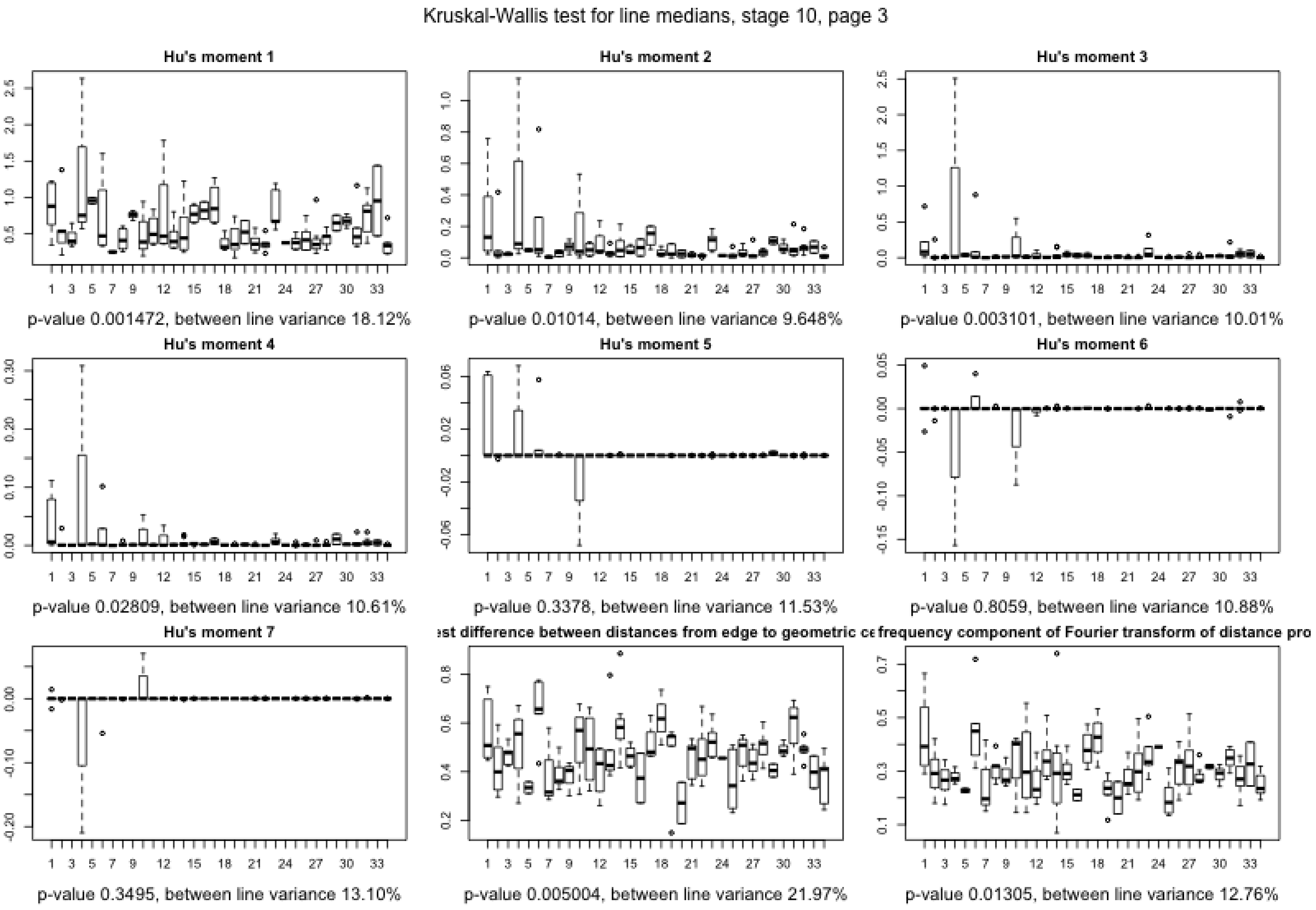
major axis of binary mask

p-value 0.4612, between line variance -0.5044%

minor axis of binary mask

p-value 0.2656, between line variance 5.291%


Kruskal-Wallis test for line medians, stage 10, page 2
eccentricity of binary mask
p-value 0.1905, between line variance 7.053%
Hu's first invariant moment of binary mask
p-value 0.04302, between line variance 15.19%
Hu's second invariant moment of binary mask
p-value 0.2577, between line variance 7.85%
total signal intensity
p-value 0.0001148, between line variance 26.85%
x coordinate of centroid (normalized by area)
p-value 0.00606, between line variance 24.05%
y coordinate of centroid (normalized by area)
p-value 0.6094, between line variance -5.167%
major axis (largest eigenvalue of the image covariance)
p-value 0.3835, between line variance 0.1879%
minor axis (smallest eigenvalue of the image covariance)
p-value 0.3606, between line variance 3.098%
object eccentricity (elongation)
p-value 0.1645, between line variance 8.135%
1 3 5 7 9 12 15 18 21 24 27 30 33



Kruskal-Wallis test for line medians, stage 10, page 3

Hu's moment 1

p-value 0.001472, between line variance 18.12%

Hu's moment 2

p-value 0.01014, between line variance 9.648%

Hu's moment 3

p-value 0.003101, between line variance 10.01%

Hu's moment 4

p-value 0.02809, between line variance 10.61%

Hu's moment 5

p-value 0.3378, between line variance 11.53%

Hu's moment 6

p-value 0.8059, between line variance 10.88%

Hu's moment 7

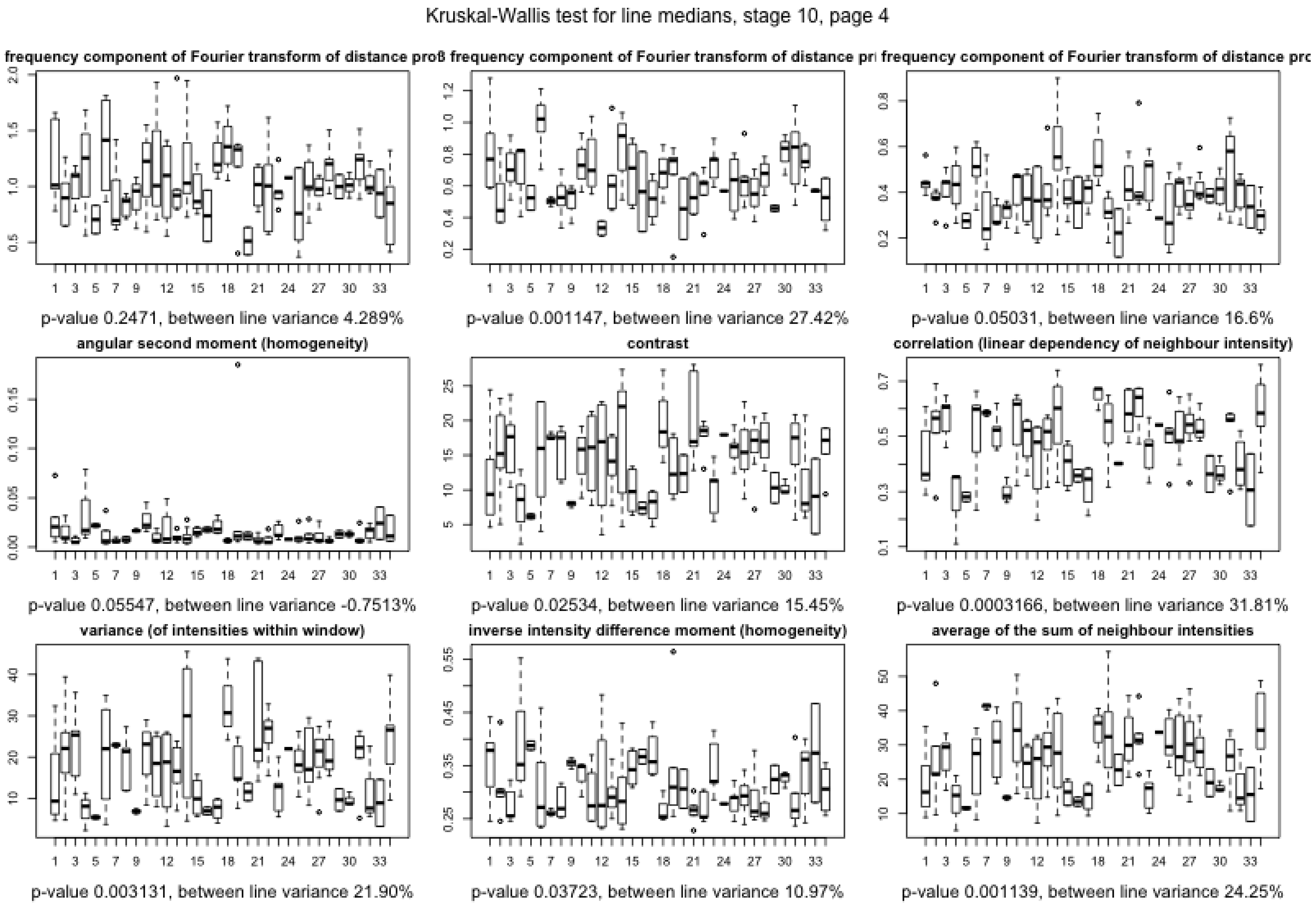
p-value 0.3495, between line variance 13.10%

est difference between distances from edge to geometric ce

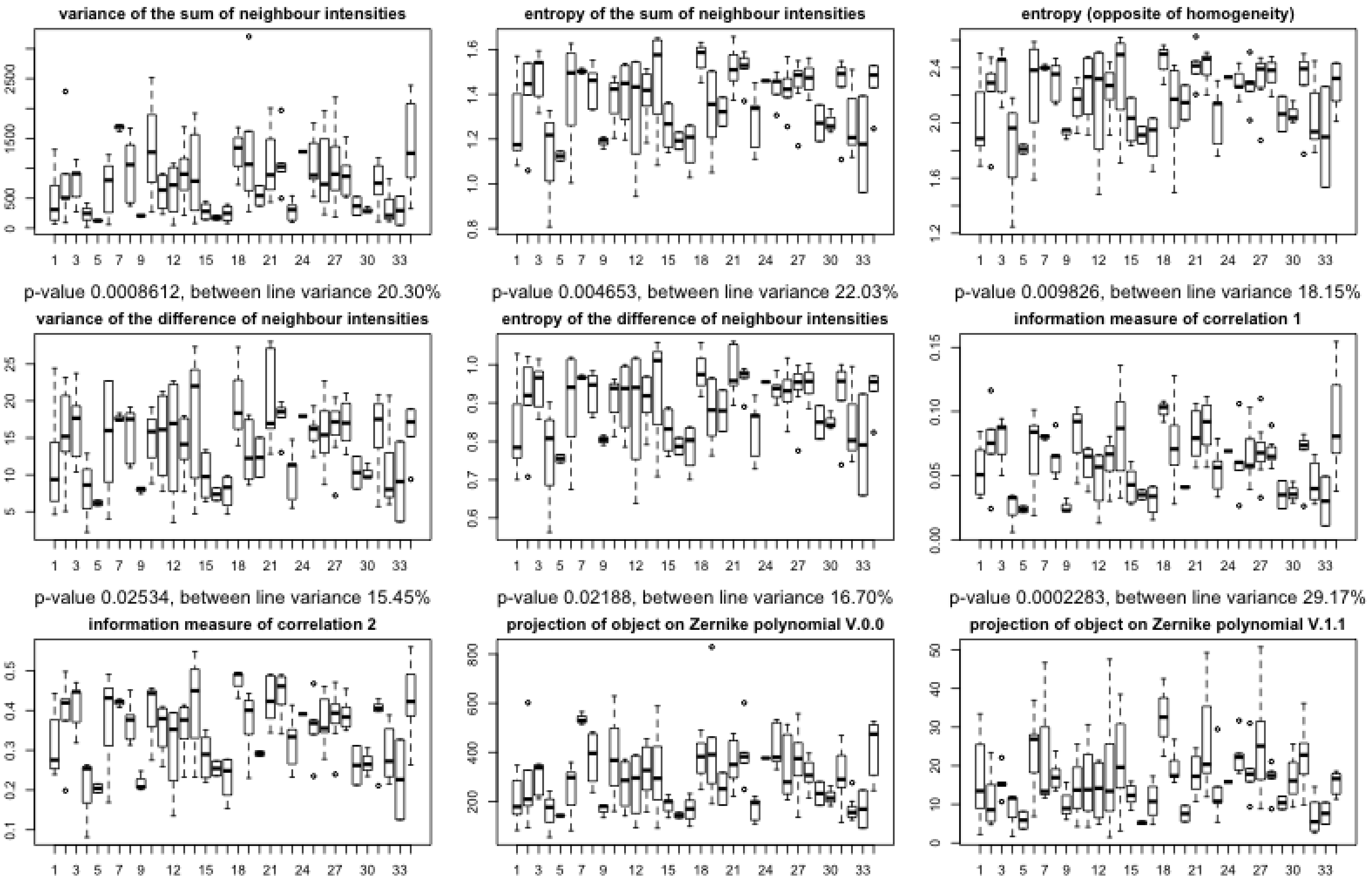
p-value 0.005004, between line variance 21.97%

frequency component of Fourier transform of distance pro

p-value 0.01305, between line variance 12.76%


Kruskal-Wallis test for line medians, stage 10, page 4
frequency component of Fourier transform of distance proß
frequency component of Fourier transform of distance pri
frequency component of Fourier transform of distance prc
2.0
1.5
1.0
0.5
1.2
1.0
0.8
0.6
0.4
0.2
0.8
0.6
0.4
0.2
1 3 5 7 9 12 15 18 21 24 27 30 33
p-value 0.2471, between line variance 4.289%
p-value 0.001147, between line variance 27.42%
p-value 0.05031, between line variance 16.6%
angular second moment (homogeneity)
contrast
correlation (linear dependency of neighbour intensity)
0.15
0.10
0.05
0.00
25
20
15
10
5
0.7
0.5
0.3
0.1
p-value 0.05547, between line variance -0.7513%
p-value 0.02534, between line variance 15.45%
p-value 0.0003166, between line variance 31.81%
variance (of intensities within window)
inverse intensity difference moment (homogeneity)
average of the sum of neighbour intensities
40
30
20
10
0.55
0.45
0.35
0.25
50
40
30
20
10
p-value 0.003131, between line variance 21.90%
p-value 0.03723, between line variance 10.97%
p-value 0.001139, between line variance 24.25%

Kruskal-Wallis test for line medians, stage 10, page 5



variance of the sum of neighbour intensities

p-value 0.0008612, between line variance 20.30%

entropy of the sum of neighbour intensities

p-value 0.004653, between line variance 22.03%

entropy (opposite of homogeneity)

p-value 0.009826, between line variance 18.15%

variance of the difference of neighbour intensities

p-value 0.02534, between line variance 15.45%

entropy of the difference of neighbour intensities

p-value 0.02188, between line variance 16.70%

information measure of correlation 1

p-value 0.0002283, between line variance 29.17%

information measure of correlation 2

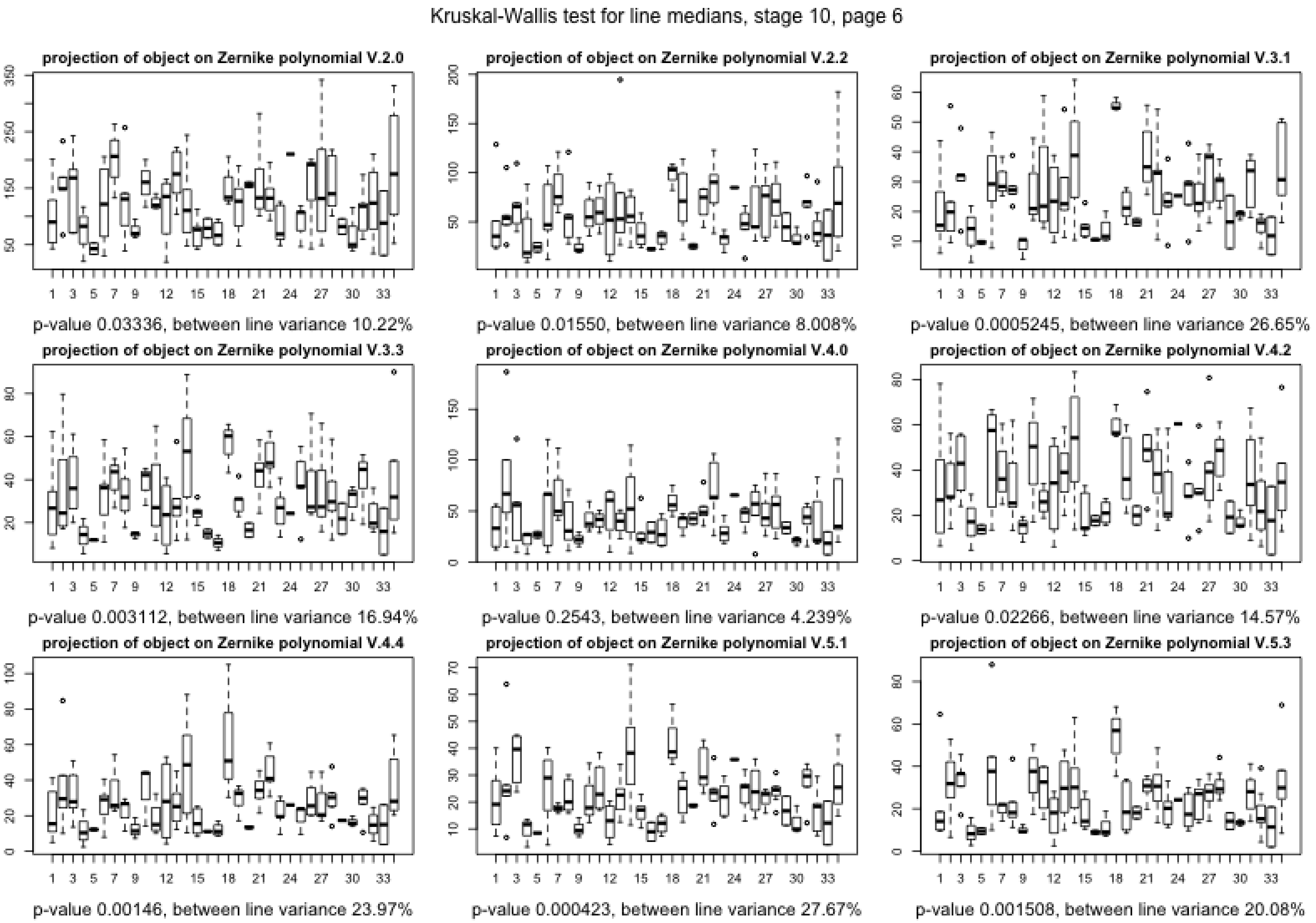
p-value 0.0003961, between line variance 30.9%

projection of object on Zernike polynomial V.0.0

p-value 0.000138, between line variance 27.94%

projection of object on Zernike polynomial V.1.1

p-value 0.004505, between line variance 16.06%



Kruskal-Wallis test for line medians, stage 10, page 6

projection of object on Zernike polynomial V.2.0

p-value 0.03336, between line variance 10.22%

projection of object on Zernike polynomial V.2.2

p-value 0.01550, between line variance 8.008%

projection of object on Zernike polynomial V.3.1

p-value 0.0005245, between line variance 26.65%

projection of object on Zernike polynomial V.3.3

p-value 0.003112, between line variance 16.94%

projection of object on Zernike polynomial V.4.0

p-value 0.2543, between line variance 4.239%

projection of object on Zernike polynomial V.4.2

p-value 0.02266, between line variance 14.57%

projection of object on Zernike polynomial V.4.4

p-value 0.00146, between line variance 23.97%

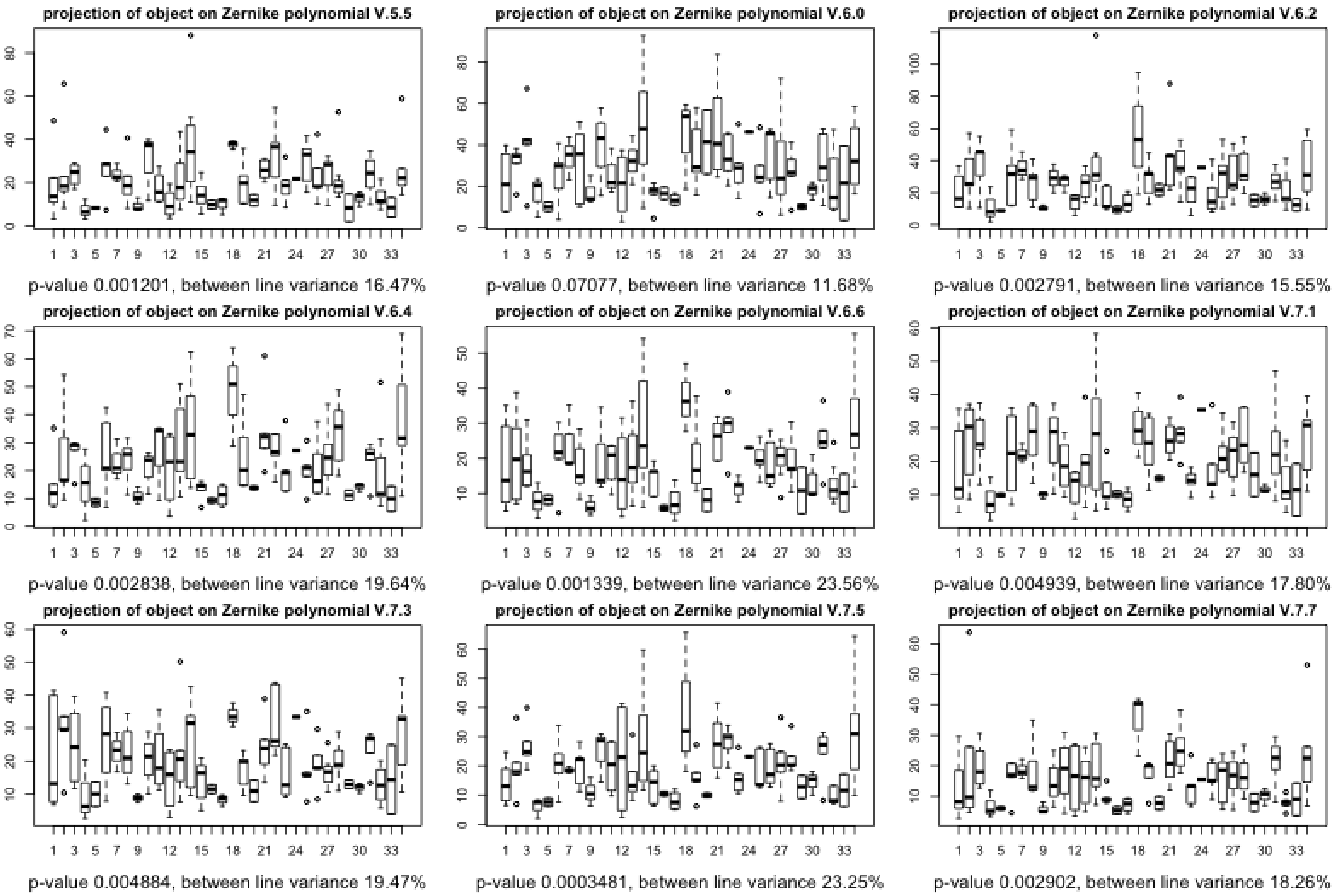
projection of object on Zernike polynomial V.5.1

p-value 0.000423, between line variance 27.67%

projection of object on Zernike polynomial V.5.3

p-value 0.001508, between line variance 20.08%

Kruskal-Wallis test for line medians, stage 10, page 7



projection of object on Zernike polynomial V.5.5

p-value 0.001201, between line variance 16.47%

projection of object on Zernike polynomial V.6.0

p-value 0.07077, between line variance 11.68%

projection of object on Zernike polynomial V.6.2

p-value 0.002791, between line variance 15.55%

projection of object on Zernike polynomial V.6.4

p-value 0.002838, between line variance 19.64%

projection of object on Zernike polynomial V.6.6

p-value 0.001339, between line variance 23.56%

projection of object on Zernike polynomial V.7.1

p-value 0.004939, between line variance 17.80%

projection of object on Zernike polynomial V.7.3

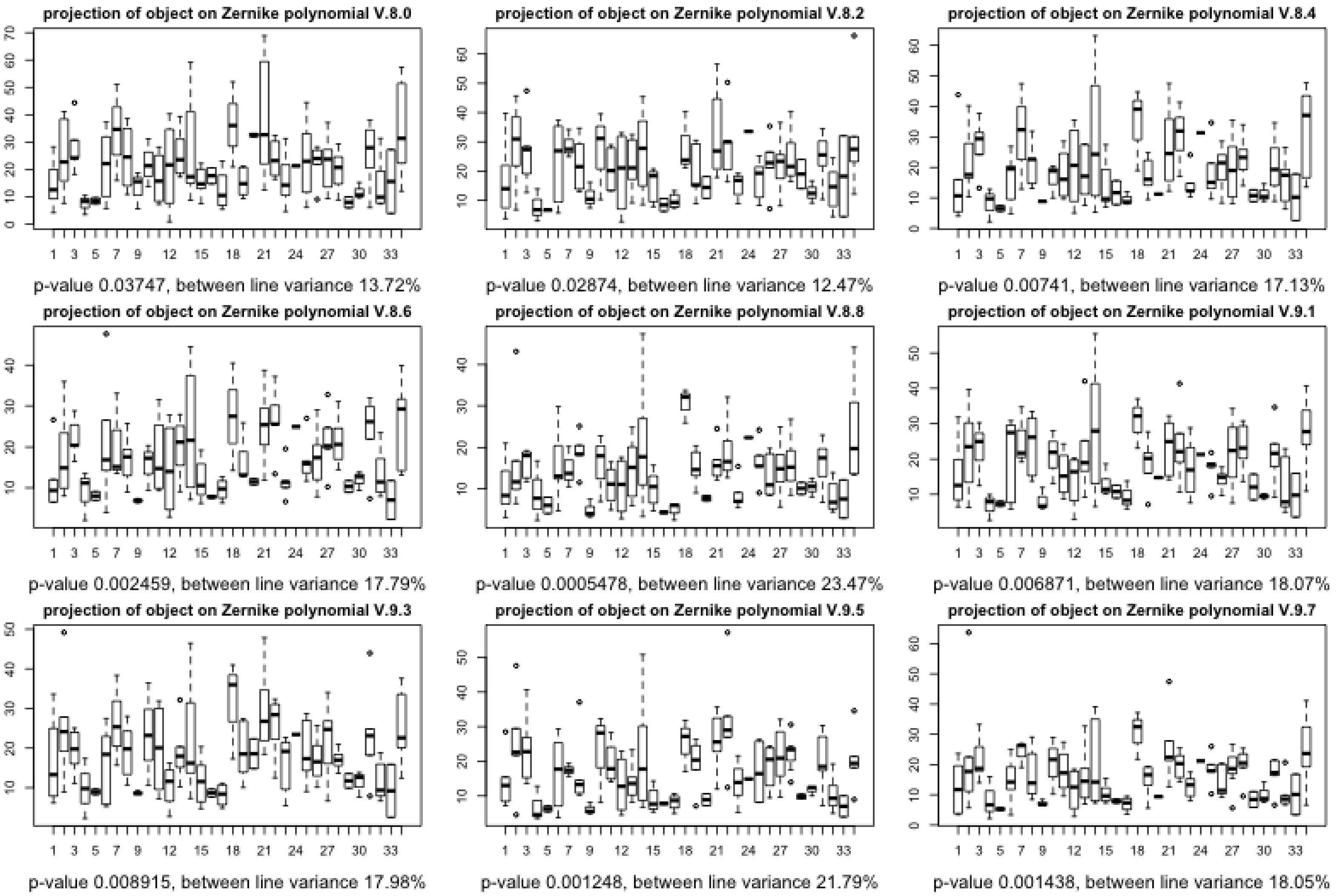
p-value 0.004884, between line variance 19.47%

projection of object on Zernike polynomial V.7.5

p-value 0.0003481, between line variance 23.25%

projection of object on Zernike polynomial V.7.7

p-value 0.002902, between line variance 18.26%



Kruskal-Wallis test for line medians, stage 10, page 8

projection of object on Zernike polynomial V.8.0

p-value 0.03747, between line variance 13.72%

projection of object on Zernike polynomial V.8.2

p-value 0.02874, between line variance 12.47%

projection of object on Zernike polynomial V.8.4

p-value 0.00741, between line variance 17.13%

projection of object on Zernike polynomial V.8.6

p-value 0.002459, between line variance 17.79%

projection of object on Zernike polynomial V.8.8

p-value 0.0005478, between line variance 23.47%

projection of object on Zernike polynomial V.9.1

p-value 0.006871, between line variance 18.07%

projection of object on Zernike polynomial V.9.3

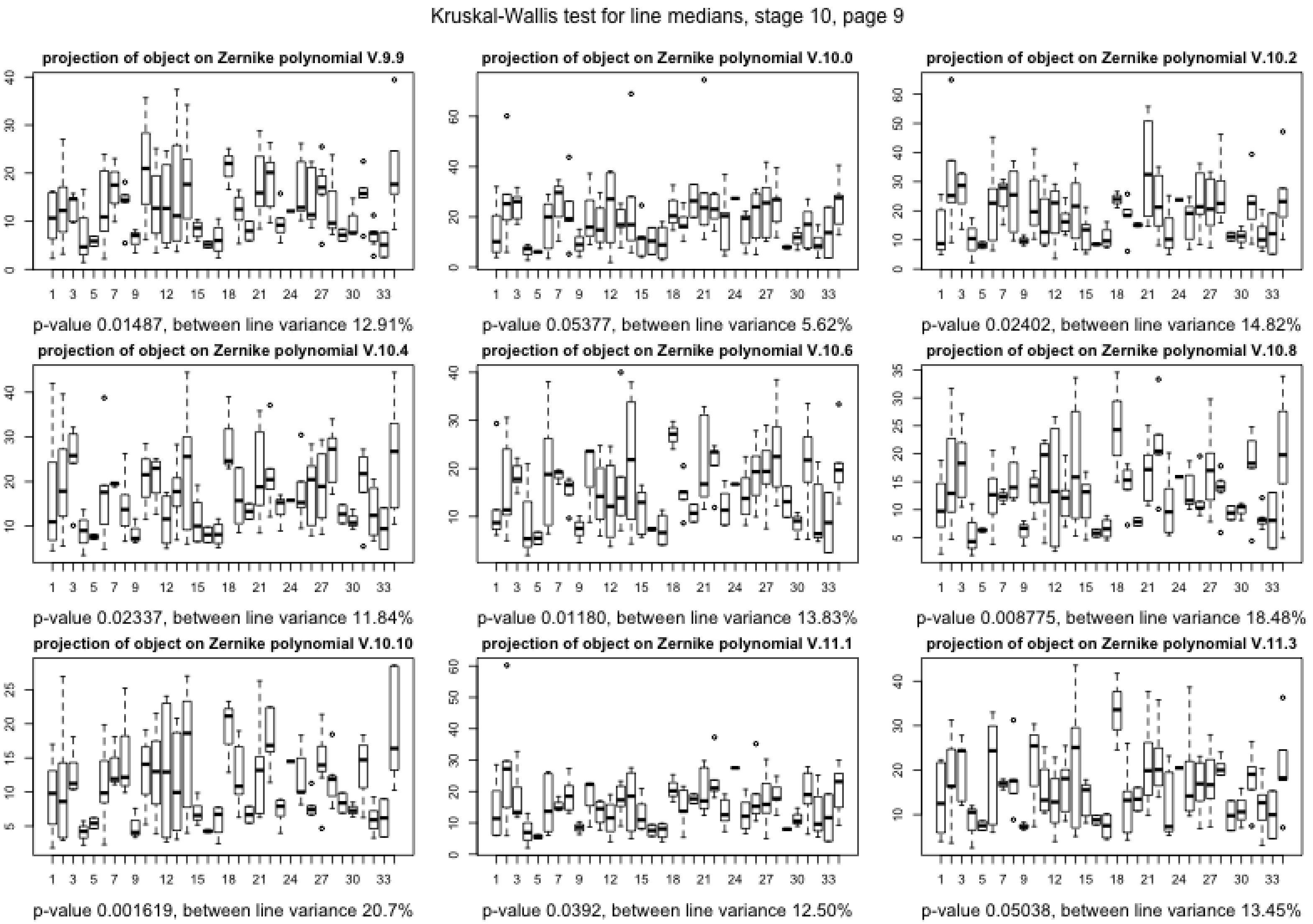
p-value 0.008915, between line variance 17.98%

projection of object on Zernike polynomial V.9.5

p-value 0.001248, between line variance 21.79%

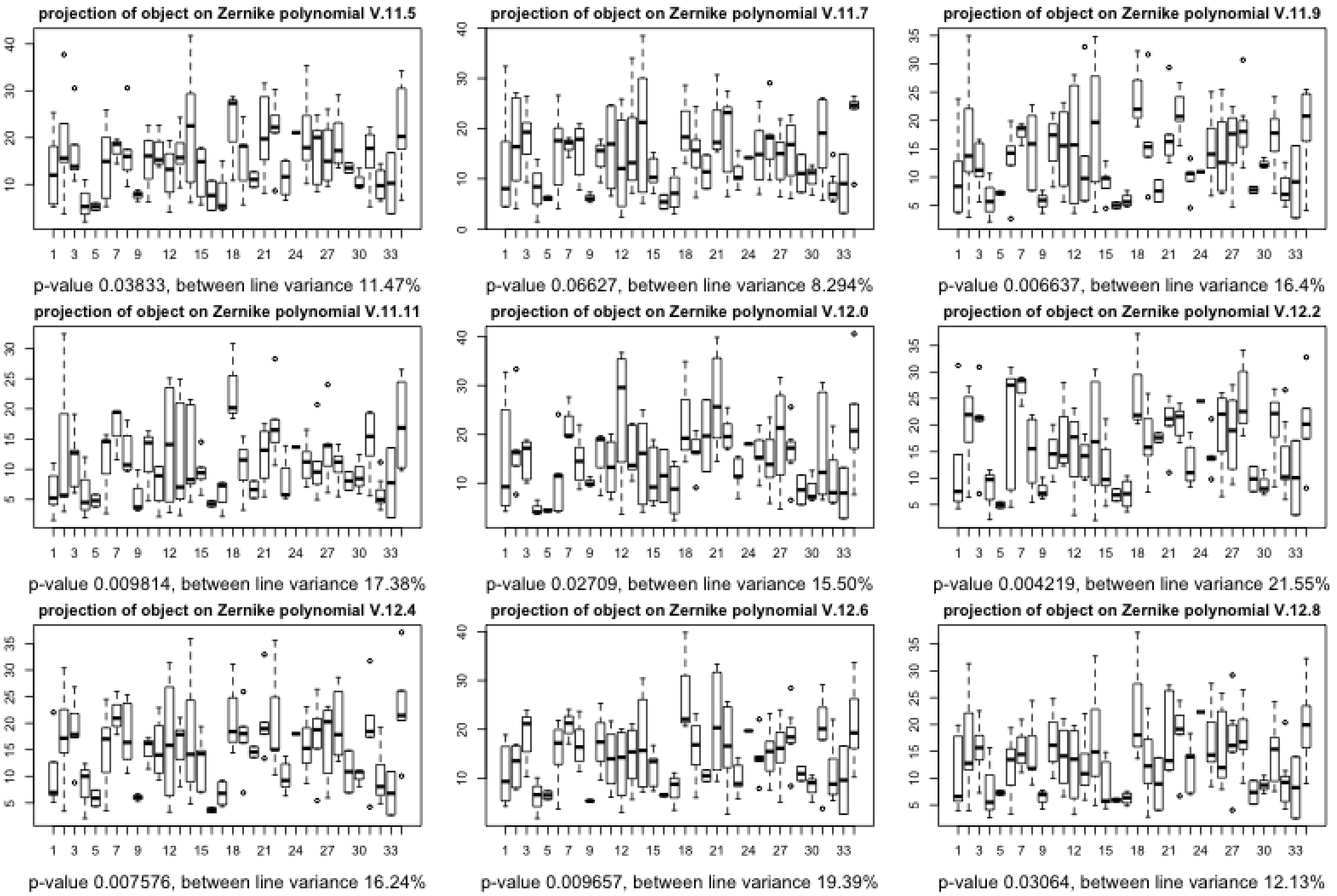
projection of object on Zernike polynomial V.9.7

p-value 0.001438, between line variance 18.05%



Kruskal-Wallis test for line medians, stage 10, page 9

Kruskal-Wallis test for line medians, stage 10, page 10



projection of object on Zernike polynomial V.11.5

p-value 0.03833, between line variance 11.47%

projection of object on Zernike polynomial V.11.7

p-value 0.06627, between line variance 8.294%

projection of object on Zernike polynomial V.11.9

p-value 0.006637, between line variance 16.4%

projection of object on Zernike polynomial V.11.11

p-value 0.009814, between line variance 17.38%

projection of object on Zernike polynomial V.12.0

p-value 0.02709, between line variance 15.50%

projection of object on Zernike polynomial V.12.2

p-value 0.004219, between line variance 21.55%

projection of object on Zernike polynomial V.12.4

p-value 0.007576, between line variance 16.24%

projection of object on Zernike polynomial V.12.6

p-value 0.009657, between line variance 19.39%

projection of object on Zernike polynomial V.12.8

p-value 0.03064, between line variance 12.13%

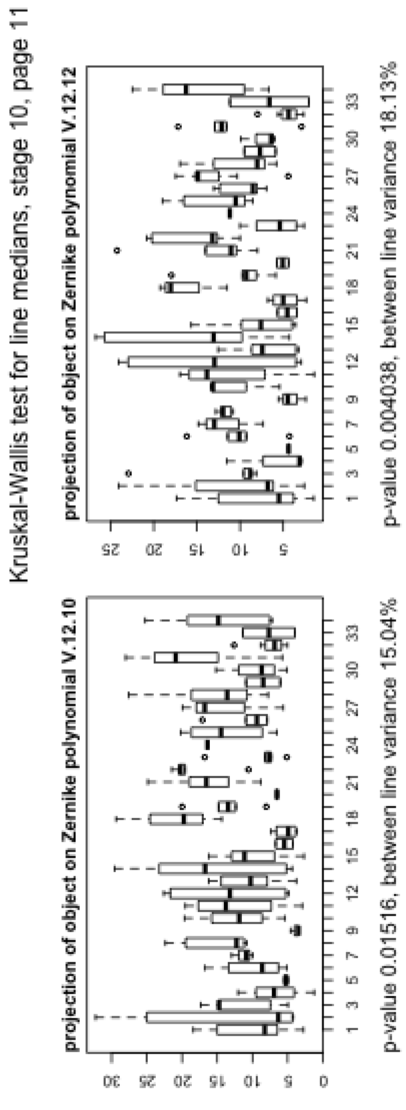


Figure A.2: Traits characterizing nurse cells, stage 10. Numbers on x axis correspond to *Drosophila melanogaster* lines, while y axis denotes trait value. Kruskal-Wallis test p-value and between line variance is indicated below each trait panel.

References

T. Aigaki, K.-H. Seong, and T. Matsuo. Longevity determination genes in *Drosophila melanogaster*. *Mech. Ageing Dev.*, 123:1531–1541, 2002. 28

A. Alcina, M. Fedetz, D. Ndagire, O. Fernandez, L. Leyva, et al. IL2RA/CD25 gene polymorphisms: uneven association with multiple sclerosis (MS) and type 1 diabetes (T1D). *PLoS ONE*, 4, 2009. 37, 156

S.M. Anjos, M.C. Tessier, and C. Polychronakos. Association of the cytotoxic T lymphocyte-associated antigen 4 gene with type 1 diabetes: evidence for independent effects of two polymorphisms on the same haplotype block. *J. Clin. Endocrinol. Metab.*, 89:6257–6265, 2004. 34

M. Ashburner, K. Golic, and S. Hawley. *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory press, Cold Spring Harbor, NY, second edition, 2005. 29, 30, 41

O.T. Avery. A further study on the biologic classification of pneumococci. *JEM*, 22:804–819, 1915. 3

J.C. Barrett, D.G. Clayton, P. Concannon, B. Akolkar, J.D. Cooper, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, 41:703–707, 2009. 34, 35, 36, 150, 153

A. Bashirullah, R. Cooperstock, and H. Lipshitz. RNA localization in development. *Annu. Rev. Biochem.*, 67:335–394, 1998. 30

A.E. Baum, N. Akula, M. Cabanero, I. Cardona, W. Corona, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry*, 13:197–207, 2008. 35

A.B. Begovich, V.E. Carlton, L.A. Honigberg, S.J. Schrodi, A.P. Chokkalingam, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.*, 75:330–337, 2004. 37

J.K. Belknap. Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. *Behavior Genetics*, 28:29–38, 1998. 8

J.K. Belknap, S.R. Mitchell, L.A. O'Toole, M.L. Helms, and J.C. Crabbe. Type I and Type II error rates for quantitative trait loci (QTL) mapping studies using recombinant inbred mouse strains. *Behavior Genetics*, 26:149–160, 1996. 16

J.K. Belknap, S.P. Richards, L.A. O'Toole, M.L. Helms, and T.J. Phillips. Shoort-term selective breeding as a tool for QTL mapping: ethanol preference drinking in mice. *Behavior Genetics*, 27:55–66, 1997. 6

G.I. Bell, S. Horita, and J.H. Karam. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes*, 33:176–183, 1984. 34

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Society B.*, 57:289–300, 1995. 19

S.T. Bennett, A.M. Lucassen, S.C.L. Gough, E.E. Powell, D.E. Undlien, et al. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat. Genet.*, 9:284–292, 1995. 34

W. Bodmer and C. Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genet.*, 40:695–701, 2008. 11

M. Boehnke and C.D. Langefeld. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.*, 62:950–961, 1998. 16

N. Bottini, L. Musumeci, A. Alonso, S. Rahmouni, K. Nika, et al. A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nat. Genet.*, 36:337–338, 2004. 34, 37

N. Bottini, T. Vang, F. Cucca, and T. Mustelin. Role of PTPN22 in type 1 diabetes and other autoimmune diseases. *Seminars in Immunology*, 18:207–213, 2006. 37, 38

A.L. Boulesteix. Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal*, 48:451–462, 2006a. 118

A.L. Boulesteix. Maximally selected chi-square statistics and binary splits of nominal variables. *Biometrical Journal*, 48:838–848, 2006b. 118

O.J. Brand, C.E. Lowe, J.M. Heward, J.A. Franklyn, J.D. Cooper, et al. Association of the interleukin-2 receptor alpha (IL-2Ralpha)/CD25 gene region with Graves' disease using a multilocus test and tag SNPs. *Clin. Endocrinol. (Oxf.)*, 66:508–512, 2007. 37

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 24, 121

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. New York: Chapman and Hall, 1984. 121

R. Brendza, L. Serbus, J. Duffy, and W. Saxton. A function for kinesin I in the posterior transport of *oskar* mRNA and Staufen protein. *Science*, 289: 2120–2122, 2000. 31, 32

T.L. Bugawan, W. Klitz, M. Alejandrino, J. Ching, A. Panelo, et al. The association of specific HLA class I and II alleles with type 1 diabetes among Filipinos. *Tissue Antigens*, 59:452–469, 2002. 34

S. Bullock and D. Ish-Horowicz. Conserved signals and machinery for RNA transport in *Drosophila* oogenesis and embryogenesis. *Nature*, 414:611–616, 2001. 31

A. Bureau, J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, et al. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28: 171–182, 2005. 24

I. Canton, S. Akhtar, N.G. Gavalas, D.J. Gawkrodger, A. Blomhoff, et al. A single-nucleotide polymorphism in the gene encoding lymphoid protein tyrosine phosphatase (PTPN22) confers susceptibility of generalised vitiligo. *Genes Immun.*, 6:584–587, 2005. 37

D.G. Chatziplis and C.S. Haley. Selective genotyping for QTL detection using sib pair analysis in outbred populations with hierarchical structures. *Genet. Sel. Evol.*, 32:547–560, 2000. 7

C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *Data Mining and Knowledge Discovery*, 3:197–217, 1999. 121

G.A. Churchill and R.W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994. 19

A. Clark, C. Meignin, and I. Davis. A Dynein-dependent shortcut rapidly delivers axis determination transcripts into the *Drosophila* oocyte. *Development*, 134: 1955–1965, 2007. 31

I. Clark, E. Giniger, H. Ruohola-Baker, L.Y. Jan, and Y.N. Jan. Transient posterior localization of a kinesin fusion protein reflects anteroposterior polarity of the *Drosophila* oocyte. *Curr. Biol.*, 4:289–300, 1994. 31

I. Clark, L.Y. Jan, and Y.N. Jan. Reciprocal localization of Nod and kinesin fusion proteins indicates microtubule polarity in the *Drosophila* oocyte, epithelium, neuron and muscle. *Development*, 124:461–470, 1997. 31

D. Clayton, J. Chapman, and J. Cooper. Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology*, 27:415–428, 2004. 17

D.G. Clayton, N.M. Walker, D.J. Smyth, R. Pask, J.D. Cooper, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genet.*, 37:695–701, 2005. 11

M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48:253–285, 2002. 23

P. Concannon, S. Onengut-Gumuscu, J.A. Todd, F. Smyth, D.J. Pociot, et al. A human type 1 diabetes susceptibility locus maps to chromosome 21q11.3. *Diabetes*, 57:2858–2861, 2008. 35

J.D. Cooper, D.J. Smyth, A.M. Smiles, V. Plagnol, N.M. Walker, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.*, 40:1399–1401, 2008. 35, 36, 154

H.J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11:2463–2468, 2002. 4

H.J. Cordell and D.G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.*, 70:124–141, 2002. 13

H.J. Cordell, J.A. Todd, N.J. Hill, C.J. Lord, P.A. Lyons, et al. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Genetics*, 158:357–367, 2001. 13

N.M. Cowen. Multiple linear regression analysis of RFLP data sets used in mapping QTLs. In *Development and application of molecular markers to problems in plant genetics*, pages 113–116. Cold Spring Harbor Laboratory press, Cold Spring Harbor, NY, 1989. 14

F. Cucca, F. Dudbridge, M. Loddo, A.P. Mulargia, R. Lampis, et al. The HLA-DPB1-associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1. *Diabetes*, 50:1200–1205, 2001a. 34

F. Cucca, R. Lampis, M. Congia, E. Angius, S. Nutland, et al. A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.*, 10:2025–2037, 2001b. 34

D. Curtis. Use of siblings as controls in case-control studies. *Annals of Human Genetics*, 61:319–333, 1997. 16

A. Darvasi and M. Soller. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.*, 85: 353–359, 1992. 16

A. Darvasi and M. Soller. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141:1999–1207, 1995. 16

A. Darvasi and M. Soller. A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics*, 27:359–371, 1997. 15

A. Darvasi, A. Weinreb, V. Minke, J.I. Weller, and M. Soller. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*, 134:943–951, 1993. 13

DAVID. David: gene functional classification. URL `http://david.abcc.ncifcrf.gov/`. 154

A.C. Davison. *Statistical models*. Cambridge University Press, 2003. 13, 20, 118

A.G. Demaine, M.L. Hibberd, D. Mangles, and B.A. Millward. A new marker in the HLA class I region is associated with the age at onset of IDDM. *Diabetologia*, 38:623–628, 1995. 34

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999. 11, 17

R. Diaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:5–32, 2006. 24

J. Dorman and C.H. Bunker. HLA-DQ locus of the human leukocyte antigen complex and type 1 diabetes mellitus: a HuGHE review. *Epidemiol. Rev.*, 22: 218–227, 2000. 33

Drosophila population genomics project. Drosophila population genomics project, 2010. URL http://www.dpgp.org/. 41, 42, 83, 157

A. Ephrussi and R. Lehmann. Induction of germ cell formation by *oskar*. *Nature*, 358:387–392, 1992. 31, 32

A. Ephrussi, L.K. Dickinson, and R. Lehmann. *oskar* organizes the germ plasm and directs localization of the posterior determinant *nanos*. *Cell*, 66:37–50, 1991. 31, 32

M.P. Epstein and G.A. Satten. Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet.*, 73:1316–1329, 2003. 17

M. Erdélyi, A. Michon, A. Guichet, J. Bogucka Glotzer, and A. Ephrussi. A requirement for *Drosophila* cytoplasmic tropomyosin in *oskar* mRNA localization. *Nature*, 377:524–527, 1995. 33

H.A. Erlich, J.I. Rotter, J.D. Chang, S.D. Shaw, L.J. Raffel, et al. Association of HLA-DPB1*0301 with insulin dependent diabetes mellitus in Mexican-Americans. *Diabetes*, 45:610–614, 1996. 34

G.J. Evans, E. Giuffra, A. Sanchez, S. Kerje, G. Davalos, et al. Identification of quantitative trait loci for production traits in commercial pig populations. *Genetics*, 164:621–627, 2003. 7

D.S. Falconer. *Introduction to quantitative genetics*. Longman, New York, 1989. 15

D.S. Falconer and T.F.C. Mackay. *Introduction to quantitative genetics*. Addison Wesley Longman, Harlow, Essex, UK, fourth edition, 1996. 4

C.T. Falk and P. Rubinstein. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, 51:227–233, 1987. 16

J.J. Fanara, K.O. Robinson, S.M. Rollmann, R.R.H. Anholt, and T.F.C. Mackay. *Vanaso* is a candidate quantitative trait gene for *Drosophila* olfactory behaviour. *Genetics*, 162:1321–1328, 2002. 8, 26, 28

K.A. Fitzpatrick, S.M. Gorski, Z. Ursuliak, and J.V. Price. Expression of protein tyrosine phosphatase genes during oogenesis in *Drosophila melanogaster*. *Mech. Dev.*, 53:171–183, 1995. 107

FlyBase. FlyBase: a database of *Drosophila* genes and genomes. URL `http://flybase.org/`. 104, 109

J. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001. 117

J. Friedman and P. Hall. On bagging and nonlinear estimation, 1999. URL `http://www-stat.stanford.edu/~jhf/`. 121

J.H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2002. 25

J.H. Friedman and B.E. Popescu. Importance sampled learning ensembles, 2003. 120

J.D. Fry, S.V. Nuzhdin, E.G. Pasyukova, and T.F.C. Mackay. QTL mapping of genotype-environment interaction for fitness in *Drosophila melanogaster*. *Genet. Res.*, 71:133–141, 1998. 25

T. Fujisawa, H. Ikegami, Y. Kawaguchi, E. Yamato, K. Takekawa, et al. Class I HLA is associated with age-at-onset of IDDM, while class II HLA confers susceptibility to IDDM. *Diabetologia*, 38:1493–1495, 1995. 34

E. Y. M. G. Fung, D.J. Smyth, J.M.M. Howson, J.D. Cooper, N.M. Walker, et al. Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immun.*, 10:188–191, 2009. 35

D.J. Futuyma. *Evolutionary biology*. Sinauer Associates, Sunderland, Massachussetts, third edition, 1997. 10

G.L. Geiger-Thornsberry and T.F.C. Mackay. Quantitative trait loci affecting natural variation in *Drosophila* longevity. *Mechanisms of ageing and development*, 125:179–189, 2004. 25, 28, 29

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. CRC Press, second edition, 2004. 18

A. Gonzalez-Reyes, H. Elliott, and D. St Johnston. Polarization of both major body axes in *Drosophila* by *gurken-torpedo* signaling. *Nature*, 375:654–658, 1995. 31

C. Gorodezky, C. Alaez, A. Murguia, A. Rodriguez, S. Balladares, et al. HLA and autoimmune diseases: type 1 diabetes (T1D) as an example. *Autoimmunity Reviews*, 5:187–194, 2006. 33, 34, 156

g:Profiler. g:Profiler. URL `http://biit.cs.ut.ee/gprofiler/`. 154

R.R. Graham, S.V. Kozyrev, E.C. Baechler, M.V.P.L. Reddy, R.M. Plenge, et al. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat. Genet.*, 38:550–555, 2006. 37

M.C. Gurganus, J.D. Fry, S.V. Nuzhdin, E.G. Pasyukova, R.F. Lyman, et al. Genotype-environment interaction at quantitative trait loci affecting sensory bristle number in *Drosophila melanogaster*. *Genetics*, 149:1883–1898, 1998. 25, 26

M.C. Gurganus, S.V. Nuzhdin, J.W. Leips, and T.F.C. Mackay. High-resolution mapping of quantitative trait loci for sternopleural bristle number in *Drosophila melanogaster*. *Genetics*, 152:1585–1604, 1999. 28

J.F. Gusella, N.S. Wexler, P.M. Conneally, S.L. Naylor, M.A. Anderson, et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306:234–238, 1983. 9

O. Hachet and A. Ephrussi. *Drosophila* Y14 shuttles to the posterior of the oocyte and is required for *oskar* mRNA transport. *Curr. Biol.*, 11:1666–1674, 2001. 33

O. Hachet and A. Ephrussi. Splicing of *oskar* RNA in the nucleus is coupled to its cytoplasmic localization. *Nature*, 428:959–963, 2004. 33

J.P. Hafler, L.M. Maier, J.D. Cooper, V. Plagnol, A. Hinks, et al. CD226 *Gly307Ser* association with multiple autoimmune diseases. *Genes Immun.*, 10:5–10, 2009. 38, 156

H. Hakonarson, S.F.A. Grant, J.P. Bradfield, L. Marchand, C.E. Kim, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature*, 448:591–594, 2007. 35

C.S. Haley and S.A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324, 1992. 13, 14

C.S. Haley, S.A. Knott, and J.-M. Elsen. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136:1999–1207, 1994. 15

B.A. Hamilton, A. Ho, and K. Zinn. Targeted mutagenesis and genetic analysis of a *Drosophila* receptor-linked protein tyrosine phosphatase gene. *Roux's Arch. Dev. Biol.*, 204:187–192, 1995. 107

R.M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979. 56

R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE*, 3:610–621, 1973. 56

S.T. Harbison, S. Chang, K.P. Kamdar, and T.F.C. Mackay. Quantitative genomics of starvation stress resistance in *Drosophila*. *Genome Biology*, 6, 2005. 25

D.L. Hartl and A.G. Clark. *Principles of population genetics*. Sinauer Associates, Sunderland, Massachussetts, fourth edition, 2007. 58

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag, second edition, 2008. 19, 21, 23, 122, 138

S.C. Heath. Markov Chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.*, 61:748–760, 1997. 25

J.M. Heward, A. Allahabadia, M. Armitage, A. Hattersley, P.M. Dodson, et al. The development of Graves' disease and the CTLA-4 gene on chromosome 2q33. *J. Clin. Endocrinol. Metab.*, 84:2398–2401, 1999. 38

X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, et al. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics*, 6:547–557, 2005. 24

L. Jacobsson, H.-B. Park, P. Wahlberg, M. Fredriksson, R. Perez-Enciso, et al. Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. *Genet. Res., Camb.*, 86: 115–125, 2005. 6

R.C. Jansen. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor. Appl. Genet.*, 85:252–260, 1992. 14

R.C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135: 205–211, 1993. 14, 113

R.C. Jansen and P. Stam. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136:1447–1455, 1994. 14

C.-H. Kao and Z.-B. Zeng. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics*, 53:359–371, 1997. 14

C.-H. Kao, Z.-B. Zeng, and R.D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152:1203–1216, 1999. 14

M.J. Kearsey and H.S. Pooni. *The genetical analysis of quantitative traits*. Chapman and Hall, London, 1996. 8

A. Khotanzad and Y.H. Hong. Invariant image recognition by Zernike moments. *Pattern Analysis and Machine Intelligence*, 12:489–497, 1990. 55, 56

W.-Y. Kim and Y.-S. Kim. A region-based shape descriptor using Zernike moments. *Signal Processing: Image Communication*, 16:95–102, 2000. 55

J. Kim-Ha, J.L. Smith, and P.M. Macdonald. *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell*, 66:23–25, 1991. 31, 32, 40

R.C. King. *Ovarian development in* Drosophila melanogaster. New York: Academic Press, 1970. 29

S. Kirkpatrick. Optimization by simulated annealing: quantitative studies. *Journal of Statistical Physics*, 34:975–986, 1984. 25

S.J. Knapp. Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theor. Appl. Genet.*, 79:583–592, 1991. 14

S.A. Knott, L. Marklund, C.S. Haley, K. Andersson, W. Davies, et al. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics*, 149:1069–1080, 1998. 7

I. Kononenko. On biases in estimating multi-valued attributes. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1034–1040. Montreal, Canada, 1995. 118

O.P. Kristiansen, Z.M. Larsen, and F. Pociot. CTLA-4 in autoimmune diseases - a general susceptibility gene to autoimmunity? *Genes Immun.*, 53:170–184, 2000. 34

L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*, 22:139–144, 1999. 10

C. Kyogoku, C.D. Langefeld, W.A. Ortmann, A. Lee, S. Selby, et al. Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am. J. Hum. Genet.*, 75:504–507, 2004. 37

M.B. Ladner, N. Bottini, A.M. Valdes, and J.A. Noble. Association of the single nucleotide polymorphism C1858T of the PTPN22 gene with type 1 diabetes. *Hum. Immunol.*, 66:60–64, 2005. 37

E. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, 11:165–177, 1995. 18

E.S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199, 1989. 13, 15

G. Landini. Auto local threshold, 2010. URL `http://pacific.mpi-cbg.de/wiki/index.php/Auto_Local_Threshold`. 48, 51

C. Lange and N.M. Laird. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genetic Epidemiology*, 23:165–180, 2002. 16

P. Lasko. RNA sorting in *Drosophila* oocytes and embryos. *FASEB J.*, 13:421–433, 1999. 31

A. Liaw and M. Wiener. Classification and regression by randomForest. *Machine Learning*, 45:5–32, 2002. 121

D.Y. Lin and D. Zeng. Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the Americal Statistical Association*, 101:179–190, 2006. 17

J. Liu, J.M. Mercer, L.F. Stam, G.C. Gibson, Z.-B. Zeng, et al. Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritana*. *Genetics*, 102/103:199–215, 1998. 26

K.E. Lohmueller, C.L. Pearce, M. Pike, E.S. Lander, and J.N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.*, 33:466–477, 2003. 12

A.D. Long and C.H. Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.*, 9:720–731, 1999. 10

A.D. Long, S.L. Mullaney, T.F.C. Mackay, and C.H. Langley. Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*. *Genetics*, 14:1497–1510, 1996. 28

A.D. Long, R.F. Lyman, C.H. Langley, and T.F.C. Mackay. Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics*, 149:999–1017, 1998. 25

A.D. Long, R.F. Lyman, A.H. Morgan, C.H. Langley, and T.F.C. Mackay. 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*, 154:1255–1269, 2000. 25

C.E. Lowe, J.D. Cooper, T. Brusko, N.M. Walker, D.J. Smyth, et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.*, 39:1074–1082, 2007. 34, 36, 37, 156

K.L. Lunetta, L.B. Hayward, J. Segal, and P.V. Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5:171–182, 2004. 24, 116

T.F.C. Mackay and R.F. Lyman. Polygenic mutation in *Drosophila melanogaster*: genotype × environment interaction for spontaneous mutations affecting bristle number. *Genetica*, 94:9734–9739, 1998. 26

M.P. Marron, A. Zeidler, L.J. Raffel, S.E. Eckenrode, J.J. Yang, et al. Genetic and physical mapping of a type 1 diabetes susceptibility gene (IDDM12) to a 100-kb phagemid artificial chromosome clone containing D2S72-CTLA4-D2S105 on chromosome 2q33. *Diabetes*, 49:492–499, 2000. 38

O. Martinez and R.N. Curnow. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.*, 85:480–488, 1992. 14

A. McConkie-Rosell, Y.-T. Chen, D. Harris, M.C. Speer, M.A. Pericak-Vance, et al. Mild cystic fibrosis linked to chromosome 7q22 markers with an uncommon haplotype. *Ann. Intern. Med.*, 111:797–801, 1989. 9

J.G. Mezey, D. Houle, and S.V. Nuzhdin. Naturally segregating quantitative trait loci affecting wing shape of *Drosophila melanogaster*. *Genetics*, 169:2101–2113, 2005. 25, 26, 27

D.R. Micklem, R. Dasgupta, H. Elliott, F. Gergely, C. Davidson, et al. The *mago nashi* gene is required for the polarization of the oocyte and the formation of perpendicular axes in *Drosophila*. *Curr. Biol.*, 7:468–478, 1997. 32

D.R. Micklem, J. Adams, S. Grunert, and D. St Johnston. Distinct roles of two conserved Staufen domains in *oskar* mRNA localization and translation. *EMBO J.*, 19:1366–1377, 2000. 32

modENCODE. Model organism encyclopedia of DNA elements (modENCODE). URL `http://www.modencode.org/`. 104, 107, 110

S.E. Mohr, S.T. Dillon, and R.E. Boswell. The RNA-binding protein Tsunagi interacts with *mago nashi* to establish polarity and localize *oskar* mRNA during *Drosophila* oogenesis. *Genes Dev.*, 15:2886–2899, 2001. 32, 33

P. Moncada, C.P. Martinez, J. Borrero, M. Chatel, Jr. Gauch, H., et al. Quantitative trait loci for yield and yield components in an *Oryza sativa* × *Oryza rufipogon* bc_2f_2 population evaluated in an upland environment. *Theor. Appl. Genet.*, 102:41–52, 2001. 5

R.F. Murphy, M.V. Boland, and M. Velliste. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:251–259, 2000. 57, 110

P.A. Newmark, S.A. Mohr, L. Gong, and R.E. Boswell. *mago nashi* mediates the posterior follicle cell-to-oocyte signal to organize axis formation in *Drosophila*. *Development*, 124:3197–3207, 1997. 32

K. Nicodemus and Y.Y. Shugart. Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case-control studies. In *Abstract volume of the Sixteenth Annual Meeting of the International Genetic Epidemiology Society*, page 611. North Yorkshire, UK, 2007. 118

L. Nistico, R. Buzzetti, L.E. Pritchard, B. van der Auwera, C. Giovannini, et al. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.*, 5:1075–1080, 1996. 34

J.A. Noble, A.M. Valdes, M. Cook, W. Klitz, G. Thomson, et al. The role of HLA class II genes in insulin-dependent diabetes mellitus: mlecular analysis of 180 Caucasian, multiplex families. *Am. J. Hum. Genet.*, 59:1134–1148, 1996. 34

J.A. Noble, A.M. Valdes, G. Thomson, and H.A. Erlich. The HLA class II locus DPB1 can influence susceptibility to type 1 diabetes. *Diabetes*, 49:121–125, 2000. 34

K.K. Norga, M.C. Gurganus, C.L. Dilda, A. Yamamoto, R.F. Lyman, et al. Quantitative analysis of bristle number in *Drosophila* mutants identifies genes involved in neural development. *Current Biology*, 13:1388–1397, 2003. 28

S.V. Nuzhdin, E.G. Pasyukova, C.L. Dilda, Z.-B. Zeng, and T.F.C. Mackay. Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, 94:9734–9739, 1997. 25, 26

S.V. Nuzhdin, L.G. Harshman, M. Zhou, and K. Harmon. Genome-enabled hitch-hiking mapping identifies QTLs for stress resistance in natural *Drosophila*. *Heredity*, 99:313–321, 2007. 25, 27

L. O'Gorman, M.J. Sammon, and M. Seul. *Practical algorithms for image analysis.* Cambridge University Press, second edition, 2008. 46, 54, 55

C. Ovilo, M. Perez-Enciso, C. Barragan, A. Clop, C. Rodriguez, et al. A QTL for intramuscular fat and backfat thickness is located on porcine chromosome 6. *Mammalian Genome*, 11:344–346, 2000. 7

I. Palacios, D. Gatfield, D. St Johnston, and E. Izaurralde. An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. *Nature*, 427:753–757, 2004. 33

A. Pallson and G. Gibson. Association between nucleotide variation in *Egfr* and wing shape in *Drosophila melanogaster*. *Genetics*, 167:1187–1198, 2004. 25, 26, 27

G. Pau, F. Fuchs, O. Sklyar, M. Boutros, and W. Huber. EBImage - an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, 26:979–981, 2010. 53

M. Perez-Enciso, A. Clop, J.L. Noguera, C. Ovilo, A. Coll, et al. A QTL on pig chromosome 4 affects fatty acid metabolism: evidence from an Iberian by Landrace intercross. *J. Anim. Sci.*, 78:2525–2531, 2000. 7

N.J. Pokrywka and E.C. Stephenson. Microtubules are a general component of mRNA localization systems in *Drosophila* oocytes. *Dev. Biol.*, 167:363–370, 1995. 31

D.N. Politis, J.P. Romano, and M. Wolf. *Subsampling.* New York: Springer, 1999. 121

Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63:490–500, 2006. 24

H.-Q. Qu and C. Polychronakos. The TCF7L2 locus and type 1 diabetes. *BMC Medical Genetics*, 8:51–55, 2007. 36

H.-Q. Qu, L. Marchand, R. Grabs, and C. Polychronakos. The IRF5 polymorphism in type 1 diabetes. *J. Med. Genet.*, 44:670–672, 2007a. 37

H.-Q. Qu, A. Montpetit, B. Ge, T.J. Hudson, and C. Polychronakos. Toward further mapping of the association between the IL2RA locus and type 1 diabetes. *Diabetes*, 56:1174–1176, 2007b. 36, 37, 156

H.-Q. Qu, J.P. Bradfield, A. Belisle, S.F.A. Grant, H. Hakonarson, et al. The type 1 diabetes association of the IL2RA locus. *Genes and Immunity*, 10:42–48, 2009. 37

D. Rabinowitz and N. Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, 50:211–223, 2000. 16

J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:Profiler - a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, 35:193–200, 2007. 154

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 255:1516–1517, 1996. 10

C. Robin, R.F. Lyman, A.D. Long, C.H. Langley, and T.F.C. Mackay. *hairy*: a quantitative trait locus for *Drosophila* sensory bristle number. *Genetics*, 162: 155–164, 2002. 25

B.O. Roep. The role of T-cells in the pathogenesis of type 1 diabetes: from cause to cure. *Diabetologia*, 46:305–321, 2003. 33

G.A. Rohrer and J.W. Keele. Identification of quantitative trait loci affecting carcass composition in swine. I. Fat deposition traits. *J. Anim. Sci.*, 76:2247–2254, 1998a. 7

G.A. Rohrer and J.W. Keele. Identification of quantitative trait loci affecting carcass composition in swine. I. Muscling and wholesale product yield traits. *J. Anim. Sci.*, 76:2255–2262, 1998b. 7

S. Roth, F.S. Neuman-Silberberg, G. Barcelo, and T. Schüpbach. *cornichon* and the EGF receptor signaling process are necessary for both anterior-posterior and dorsal-ventral pattern formation in *Drosophila*. *Cell*, 81:967–978, 1995. 31

J.M. Satapogan, B.S. Yandell, M.A. Newton, and T.C. Osborn. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816, 1996. 14

G.A. Satten, W.D. Flanders, and Q. Yang. Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.*, 68:466–477, 2001. 11

Y. Shih. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18:547–557, 2005. 24

S. Sigurdsson, G. Nordmark, H.H.H. Göring, K. Lindroos, A.-C. Wiman, et al. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am. J. Hum. Genet.*, 76: 528–537, 2005. 37

M.J. Sillanpaa and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148:1373–1388, 1998. 14

D. Smyth, J.D. Cooper, J.E. Collins, J.M. Heward, J.A. Franklyn, et al. Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes*, 53:3020–3023, 2004. 34, 37

D.J. Smyth, J.D. Cooper, R. Bailey, S. Field, O. Burren, et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.*, 38:617–619, 2006. 35

R.R. Sokal and F.J. Rohlf. *Biometry: the principles and practice of statistics in biological research*. W. H, Freeman and Co., New York, third edition, 1995. 78, 80

M. Soller and J.S. Beckmann. Marker-based mapping of quantitative trait loci using replicated progenies. *Theor. Appl. Genet.*, 80:205–208, 1990. 8

R.S. Spielman and W.J. Ewens. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.*, 59:983–989, 1996. 16

R.S. Spielman and W.J. Ewens. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.*, 62: 450–458, 1998. 16

R.S. Spielman, R. McGinnis, and W.J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, 52:506–516, 1993. 16

A.C. Spradling. Development genetics of oogenesis. In *The development of* Drosophila melanogaster, pages 1–70. Cold Spring Harbor Laboratory press, Cold Spring Harbor, NY, 1993. 29

D. St Johnston. Moving messages: the intracellular localization of mRNAs. *Nat. Rev. Mol. Cell Biol.*, 6:363–375, 2005. 30

D. St Johnston, W. Driever, T. Berleth, S. Richstein, and C. Nüsslein-Vorhard. Multiple steps in the localization of *bicoid* RNA to the anterior pole of the *Drosophila* oocyte. *Development (Suppl.)*, 107:13–19, 1989. 31

P. Stam. Some aspects of QTL analysis. In *Proceedings of the eighth meeting of the Eucarpia section biometrics in plant breeding*. BRNO, 1991. 14

M. Stephens and D.J. Balding. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genetics*, 10:681–690, 2009. 19

D.O. Stram, C.L. Pearce, P. Bretsky, M. Freedman, J.N. Hirschhorn, et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity*, 55: 179–190, 2003. 17

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:1189–1232, 2007. 118, 121

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:838–848, 2008. 118

A.H. Sturtevant, C.B. Bridges, and T.H. Morgan. The spatial relations of genes. *Proc. Natl. Acad. Sci. USA*, 5:168–173, 1919. 3

S.D. Tanksley and J.C. Nelson. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor. Appl. Genet.*, 92:191–203, 1996. 5, 6

S.D. Tanksley, S. Grandillo, T.M. Fulton, D. Zamir, Y. Eshed, et al. Advanced backcross QTL analysis in a cross between elite processing line of tomato and its wild relative *L. pimpinellifolium*. *Theor. Appl. Genet.*, 92:213–224, 1996. 5

The modENCODE Consortium, S. Roy, J. Ernst, P.V. Kharchenko, P. Kheradpour, et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330:1787–1797, 2010. 104, 107, 110

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996. 22

J.A. Todd, N.M. Walker, J.D. Cooper, D.J. Smyth, K. Downes, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.*, 39:857–864, 2007. 35, 155

J. Tower. Transgenic methods for increasing *Drosophila* lifespan. *Mech. Ageing Dev.*, 118:1–14, 2000. 28

D.A. Tregouet, S. Barbaux, S. Escolano, N. Tahri, J.L. Golmard, et al. Specific haplotypes of the P-selectin gene are associated with myocardial infarction. *Hum. Mol. Genet.*, 11:2015–2023, 2002. 18

D.A. Tregouet, S. Escolano, L. Tiret, A. Mallet, and J.L. Golmard. A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Ann. Hum. Genet.*, 68:165–177, 2004. 18

M. Tsui, L.C. Buchwald, D. Barker, J.C. Braman, R. Knowlton, et al. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, 230:1054–1057, 1985. 9

S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. McQuilton, et al. FlyBase: (e)nhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Research*, 37:555–559, 2008. 104, 109

H. Ueda, J.M.M. Howson, L. Esposito, J. Heward, G. Chamberlain, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, 423:506–511, 2003. 34, 38

P. Vafiadis, S.T. Bennett, J.A. Todd, J. Nadeau, R. Grabs, et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nat. Genet.*, 15:289–292, 1997. 34

W. Valdar, L.C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, 38:879–887, 2006. 24

W. Valdar, C.C. Holmes, R. Mott, and J. Flint. Mapping in structured populations by resample model averaging. *Genetics*, 182:1263–1277, 2009. 24, 117

A.M. Valdes, G. Thomson, H.A. Erlich, and J.A. Noble. Association between type 1 diabetes age of onset and HLA among sibling pairs. *Diabetes*, 48:1658, 1999. 34

A.M. Valdes, H.A. Erlich, and J.A. Noble. Human leukocyte antigen class I B and C loci contribute to type 1 diabetes (T1D) susceptibility and age at T1D onset. *Human immunology*, 66:301–313, 2005. 34

A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998. 77, 78

F.J.M. van Eeden, I.M. Palacios, M. Petronczki, M.J.D. Weston, and D. St Johnston. Barentsz is essential for the posterior localization of *oskar* mRNA and colocalizes with it to posterior. *J. Cell. Biol.*, 154:511–524, 2001. 32, 41

M. van Oene, R.F. Wintle, X. Liu, M. Yazdanpanah, X. Gu, et al. Association of the lymphoid tyrosine phosphatase R620W variant with rheumatoid arthritis, but not Crohn's disease, in Canadian populations. *Arthritis Rheum.*, 52:1993–1998, 2005. 37

M.R. Velaga, V. Wilson, C.E. Jennings, C.J. Owen, S. Herington, et al. The codon 620 tryptophan allele of the lymphoid tyrosine phosphatase (LYP) gene is a major determinant of Graves' disease. *Clin. Endocrinol. Metab.*, 89:5862–5865, 2004. 37

A. Vella, J.D. Cooper, C.E. Lowe, N. Walker, S. Nutland, et al. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, 76:773–779, 2005. 36, 37, 156

C. Vieira, E.G. Pasyukova, Z.B. Zeng, J. Brant Hackett, R.F. Lyman, et al. Genotype-environment interaction for quantitative trait loci affecting life span in *Drosophila melanogaster*. *Genetics*, 154:213–227, 2000. 25, 26

P.M. Visscher, R. Thompson, and C.S. Haley. Confidence intervals in QTL mapping by bootstrapping. *Genetics*, 143:1013–1020, 1996. 15

D. Wang, G.L. Graef, A.M. Procopiuk, and B.W. Diers. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theor. Appl. Genet.*, 108:458–467, 2004. 5

J.D. Watson and F.H.C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:82–84, 1953. 3

K. Weber, R. Eisman, L. Morey, A. Patty, J. Sparks, et al. An analysis of polygenes affecting wing shape on chromosome 3 in *Drosophila melanogaster*. *Genetics*, 153:773–786, 1999. 26

K. Weber, R. Eisman, S. Higgins, L. Morey, A. Patty, et al. An analysis of polygenes affecting wing shape on chromosome 2 in *Drosophila melanogaster*. *Genetics*, 159:1045–1057, 2001. 26

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007. 11, 34, 35, 123, 124, 155

J.E. Wilhelm, J. Mansfield, N. Hom-Booher, S. Wang, C.W. Turck, et al. Isolation of a ribonucleoprotein complex involved in mRNA localization in *Drosophila* oocytes. *J. Cell Biol.*, 148:427–439, 2000. 30, 41

H. Wu, R.M. Cantor, D.S. Graham, C.M. Lingren, L. Farwell, et al. Association analysis of the R620W polymorphism of protein tyrosine phosphatase PTPN22 in systemic lupus erythematosus families: increased T allele frequency in systemic lupus erythematosus patients with autoimmune thyroid disease. *Arthritis Rheum.*, 52:2396–2402, 2005. 37

T. Xia, H. Zhu, S. Huazhong, P. Haigron, and L. Luo. Image description with generalized pseudo-zernike moments. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.*, 24:50–59, 2007. 56, 110

D.V. Zaykin, P.H. Westfall, S.S. Young, M.A. Karnoub, M.J. Wagner, et al. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, 53:79–91, 2002. 17, 18

Z.-B. Zeng. Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. *Proc. Natl. Acad. Sci. USA*, 90:10972–10976, 1993. 14

Z.-B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136:1457–1468, 1994. 14

L.P. Zhao, S.S. Li, and N. Khalid. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am. J. Hum. Genet.*, 72:1231–1250, 2003. 17

E. Zimmerman, A. Pallson, and G. Gibson. Quantitative trait loci affecting components of wing shape in *Drosophila melanogaster*. *Genetics*, 155:671–683, 2000. 26

V.L. Zimyanin, N. Lowe, and D. St Johnston. An *oskar*-dependent positive feedback loop maintains the polarity of the *Drosophila* oocyte. *Curr. Biol.*, 17: 353–359, 2007. 31

V.L. Zimyanin, K. Belaya, J. Pecreaux, M.J. Gilchrist, A. Clark, et al. In vivo imaging of *oskar* mRNA transport reveals the mechanism of posterior localization. *Cell*, 134:843–853, 2008. 31

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320, 2005. 22