Fine-mapping complex traits

in heterogeneous stock rats

Amelie Baud

Wadham College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2013

Abstract

The fundamental theme my thesis explores is the relationship between genetic variation and phenotypic variation. It addresses three main questions. What is the genetic architecture of traits in the HS? How can sequence information help identifying the sequence variants and genes responsible for phenotypic variation? Are the genetic factors contributing to phenotypic variation in the rat homologous to those contributing to variation in the same phenotype in the mouse?

To address these questions, I analysed data collected by the EURATRANS consortium on 1,407 Heterogeneous Stock (HS) rats descended from eight inbred strains through sixty generations of outbreeding. The HS rats were genotyped at 803,485 SNPs and 160 measures relevant to a number of models of disease (e.g. anxiety, type 2 diabetes, multiple sclerosis) were collected. The eight founders of the Stock were genotyped and sequenced. I identified loci in the genome that contribute to phenotypic variation (Quantitative Trait Loci, QTLs), and integrated sequence information with the mapping results to identify the genetic variants underlying the QTLs.

I made some important observations about the nature of genetic architecture in rats, and how this compares to mice and humans. I also showed how sequence information can be used to improve mapping resolution, and in some cases to identify causal variants. However, I report an unexpected observation: at the majority of QTLs, the genetic effect cannot be accounted for by a single variant. This finding suggests that genetic variation cannot be reduced to sequence variation. This complexity will need to be taken into account by studies that aim at unravelling the genetic basis of complex traits.

Acknowledgements

I would like to thank my supervisors, Professors Jonathan Flint and Richard Mott, for their exceptional guidance and support throughout my DPhil. I was very fortunate to have not only one, nor two, but a team of two brilliant scientists and caring supervisors to get me started in research. You are an inspiration. Thank you for everything you've done for me.

I would also like to thank present and past members of the Flint and Mott groups for creating such a friendly and stimulating environment at work and outside work. I had the best of times with Binnaz and Helen in Ambert. Polinka, Carme, Regina, Adam, and Martina, together with other friends in the EuRATools consortium, turned a not so nice job into a good experience (harvesting) and lovely time (dinners) in Barcelona. Amarjit, I really enjoyed our sweaty lunch breaks. Jerome, your support during my social adventure is much appreciated. Leo, ★ for knowing so many interesting things and telling me about them, and being so kind, and the chocolate. Jessie, I really enjoy our nights out and chats, let's keep doing it, and it's good to know we'll submit together tomorrow :) Na, thank you for your friendship, for showing me the good restaurants in Oxford, and providing me with an alternative to sleeping in the lab these last few days. Martha, it was good talking science and life and life in science with you over pizza. Caroline, see you soon my friend! GJ, I wish you were flying your helicopter in Jonathan's office right now, I wouldn't be alone in the lab. Rosie, thank you for your help many times. Mona, Tota, Jon, Martin, Xiangchao, thanks for good moments together.

Thank you to my fellow GMS students, Evan Harrell, Witold Czyz, Jared O'Connell, and Iain Mathieson, for their advice and help when I needed it, and a lot of good time spent together.

I would like to thank John Broxholme, Tim Bardsley, Mark Gibbons, Ruth Porter, and Robert Esnouf for their help and encouragement with computers, clusters, software, and the like, and teaching me some.

I am very grateful to the members of the Rat Genome Sequence and Mapping Consortium and the EURATRANS consortium for giving me the opportunity to work on a great project for my DPhil. I was thrilled to be part of the very dynamic and friendly rat community thanks to this collaboration. Samreen, thank you for your work on the Chr. 5 QTL - it gave a sense to my analysis - and for your friendship. Jonatan, working with you to try to make something out of the Chr. 9 QTL was extremely motivating, and it was great to have electronic company at night. Pernilla, thank you for your help on many things with the paper and your support while in Oxford - it helped me power through a not so comfortable time. Victor, it was good working closely with you on the paper, thank you.

I am very grateful to the Wellcome Trust for funding my DPhil. and giving me the opportunity to carry out exciting research and go to many conferences.

I thank my family for their love and support at all times, which give me so much strength.

Last but not least, thank you Loulou for supporting my decision to go to Oxford four years ago, and for making me so happy always, together and apart. Glad it's going to be together from now on :)

List of abbreviations

APT	affymetrix power tools
BAC	bacterial artificial chromosome
BMI	body mass index
BN	brown norway
CF	cystic fibrosis
CNG	centre national de genotypage
DGRP	drosophila genetic reference panel
DHS	deoxyribonuclease I hypersensitive
EAE	experimental autoimmune encephalomyelitis
EMSA	electrophoretic mobility shift assay
ENCODE	encyclopedia of DNA elements
ER	estrogen receptor
EVD	extreme value distribution
FDR	false discovery rate
GWA	genome-wide association
GWAS	genome-wide association study
HDL	high-density lipoprotein
HS	heterogeneous stock
IBD	identical by descent
IBS	identity by state
IQR	interquartile range
K	thousand
kb	kilobase
LD	linkage disequilibrium
LDL	low-density lipoprotein

logP	negative logarithm 10 of the p value
MAF	minor allele frequency
Mb	megabase
MCV	mean corpuscular volume
MDC	max delbruck center
MHC	major histocompatibility complex
NHGRI	national human genome research institute
NIH	national institute of health
pctB	proportion of B cells in WBC
QTL	quantitative trait locus
RATDIV	rat diversity
RFLP	restriction fragment length polymorphism
RGD	rat genome database
s.d.	standard deviation
SDP	strain distribution pattern
SNP	single nucleotide polymorphism
SSLP	simple sequence length polymorphism
TF	transcription factor
TSS	transcription start site
UCSC	university of california santa cruz

Contents

Table of figures

Table of tables

Chapter 1

Introduction

This study was carried out as part of the EURATRANS consortium[1]. Its overall aim was to identify the sequence variants and genes responsible for variation in many complex traits of biomedical relevance in the laboratory rat. The EURATRANS consortium collected data on 1,407 Heterogeneous Stock (HS) rats descended from eight inbred strains through sixty generations of outbreeding. The HS rats were genotyped at 803,485 SNPs and 160 measures relevant to a number of models of disease (e.g. anxiety, type 2 diabetes, multiple sclerosis) were collected. The eight founders of the stock were genotyped and sequenced. This thesis is concerned with the analysis of these data.

1.1 Thesis aims

The theme of my thesis is the relationship between genetic and phenotypic variation. It addresses three main questions.

First, what is the genetic architecture of traits in the rat HS, and how does it compare to that in a mouse HS[2] and humans? Knowledge of the genetic architecture of traits is important in its own right but is also necessary to improve the design of experiments that seek to identify the genetic factors contributing to phenotypic variation.

Second, how can sequence information help identify the sequence variants and genes responsible for phenotypic variation? I will consider two components of sequence information. First, annotation, whereby function is attributed to DNA elements (e.g. gene[3], 5' untranslated region[3], etc.) or consequence is attributed to a genetic variant (e.g. missense single nucleotide polymorphisms (SNP)[4]). Second, sequence variation, whereby not only a subset of markers but all the variants that segregate in the mapping population have been genotyped. The known ancestry of the HS makes it possible to infer genotypes of all the variants that segregate in the HS. I will describe how sequence information helped identify causal genetic variants, and the limitations of this approach.

Third, are the causal genetic factors in the rat homologous to those contributing to variation in the same phenotype in the mouse? Asking this question is a step towards evaluating the assumption that underlies the use of model organisms in biomedical research, namely that the same genes or pathways contribute to variation in a phenotype across species. Thirty eight phenotypes were collected both in the rat HS and a mouse HS[2], using similar protocols. They offer a unique opportunity for a genome-wide test of the conservation of causal variation.

1.2 Genetic architecture

The genetic architecture of a trait refers to its overall heritability, the number of alleles contributing to phenotypic variation, their frequencies, effect sizes, and the way in which they combine together (additivity, dominance, gene-gene interactions, genotypic correlations, pleiotropy) and with the environment (gene-environment interactions, gene-environment correlations)[5].

In this section, I first review the genetic architecture in humans because this is the population in which the most information is available. I discuss the following features of genetic architecture: (i) genetic variation contributes to natural phenotypic variation, (ii) multiple genes are involved, (iii) they contribute to phenotypic variation in a non-additive way, (iv) individual genes impact multiple traits, (vi) genotypes and environment contribute to phenotypic variation through their main effects, correlations and interactions. Then I investigate how genetic architecture in humans compares with that in other species, in particular laboratory populations of model organisms. I show how the breeding of these populations shapes their genetic architecture. Finally, I outline the genetic architecture of complex traits in a mouse HS[2], as a basis for comparison with the rat HS.

Information about the genetic architecture of traits can be obtained without knowledge of the individual loci that contribute to phenotypic variation, by analysing families and large pedigrees.

1.2.1 Insights from quantitative genetics

The joint contribution of all the causal loci to a trait is summarized by its heritability. It can be estimated without knowing which loci contribute to phenotypic variation, based on correlations between phenotypic similarity and degree of genetic relatedness in pedigrees[6-8]. Narrow-sense heritability - the proportion of phenotypic variation attributable to additive genetic effects - is of primary interest to quantitative geneticists[8]. It enables predictions of the response to both artificial and natural selection, and is a key parameter in predicting the efficiency of gene-mapping studies[8]. Heritability in humans ranges from moderate (e.g. 37% for depression[9] and 38% for type 2 diabetes[10]) to complete (Mendelian traits).

Biomedical research focuses on disease, which is a binary phenotype, but variation in most other traits is quantitative, suggesting that multiple genes might be at stake[11]. The non-Mendelian segregation of common diseases in large pedigrees suggests they are controlled by multiple genes[12] in the same way as quantitative traits. 'Complex traits' - traits controlled by multiple genes as well as the environment - are the focus of the next paragraphs.

Broad-sense heritability includes non-additive genetic effects[8]. These arise from interactions between the alleles of a gene (dominance), from interactions between genomic loci (gene-gene interactions, or 'epistasis'), and from genotypic correlations between loci[6,8]. Non-additive genetic effects are important to predict phenotype from genotypes[13].

The contribution of dominance to variation in complex traits can be significant. It has been estimated using various designs, for example human twin studies. In the twin design, because only three variance components can be estimated from resemblance among monozygotic twins, resemblance among dizygotic twins, and total variance, the contribution of the environment shared by twins has to be ignored if one wants to estimate dominance in addition to additive genetic effects, and effects of the individual environment[14]. Not accounting for a source of variation may lead to biased estimates[8]. Therefore, dominance is better estimated in pedigrees. For example, from 25,000 twin pairs and 50,000 biological and adoptive relatives, Eaves and colleagues estimated that dominance contributed half of the 67% of variation in body mass index (BMI) attributable to genetic effects[15]. Some human populations (e.g. Hutterites) are particularly well suited to estimate dominance because every pair of individuals has a non-zero probability of sharing two alleles identical by descent as a result of inbreeding, and because individuals are related through multiple lines of descent[16]. In Hutterites, dominance accounts for 60%, 49%, 45%, and 31% of the variance in low-density lipoprotein (LDL), serotonin levels, systolic blood pressure, and fat free mass respectively. It was not significant for high-density lipoprotein (HDL), triglycerides, diastolic blood pressure, immunoglobulin E, lipoprotein (a), and BMI[16].

Estimates of the contribution of epistasis to phenotypic variation are difficult to obtain because of correlations with other components of genetic similarity[17], and are subject to large sampling error[18].

Causal loci whose genotypes are correlated either through physical linkage or linkage disequilibrium also contribute non-additively to phenotypic variation[6,8]. Such correlations can arise from non-random mating. Eaves and colleagues found evidence for a small contribution of assortative mating to variation in BMI[15].

Another important aspect of genetic architecture is the contribution of individual genes or sequence variants to multiple traits, or pleiotropy[19]. Pleiotropy causes compromises among adaptations of different traits when a mutation that is advantageous through one trait may be harmful through another[19]. Understanding pleiotropy is important in artificial selection programs[20], as well as in drug development programs since targeting one gene may have side effects. Those may be beneficial - as illustrated by the effects of statins on the cardiovascular function[21], or negative. The extent to which two traits share their genetic basis is measured by modelling jointly the two traits and estimating their genetic correlation[6]. Mitchell *et al.* found correlations of 0.486, -0.356, 0.274 between insulin levels and BMI, HDL levels, and waist-hip ratio respectively[22].

The genetic and environmental components of phenotypic variation are not always independent. For example, the normal practice of dairy husbandry is to feed cows according to their yield. Therefore the better genetic makeup will be given a more favourable environment[6]. In humans, a similar situation occurs where parents who are genetically gifted for verbal communication, in addition to passing on to their offspring some or all of those alleles favourable to verbal abilities, typically provide their offspring with an environment favourable to the development of verbal abilities[23]. Similarly to genotypic correlations, gene-environment correlations contribute more than their additive main effects to phenotypic variance. Lyons *et al.* found a genetic component to volunteering for service in Vietnam, serving in Southeast Asia, and self-reported combat experiences[24], providing an example of active genotype-environment correlation, while the milk yield and verbal abilities examples above are of a passive nature[25]. Kendler and Baker also report pervasive genetic contribution to environmental measures relevant to psychiatry/psychology, with modest to moderate impact[26]. If unaccounted for, gene-environment correlations will lead to misestimating other components of phenotypic variation.

Gene-environment correlations may be difficult to distinguish from gene-environment interactions in wild populations[25]. In the laboratory, inbred strains can be faced with multiple environments to investigate gene-environment interactions. For example, strains of the paradise fish *Macropodus opercularis* were exposed to four environments that differed in their level of novelty and threat, and their behaviour was recorded[27]. Significant gene-environment interactions were observed, with the ordering of the strains (from most to least anxious) varying widely across the different environments.

Phenotypic differences between sexes are ubiquitous but usually moderate. Valdar *et al.* found extensive evidence for a main effect of sex in laboratory outbred mice (71 out of 88 phenotypes), but in more than half of the cases the effect was <5%[28]. Significant gene-by-sex interactions were also detected for 53 phenotypes, and explained on average 22% of the phenotypic variance (in addition to the main effect of sex)[28]. An analysis of twin data found a significant contribution of gene-by-sex interactions for only 4% of 122 phenotypes[29].

In summary, quantitative genetics allows us to determine which of the possible genetic architectures that might exist actually applies to a given trait. It also provides important information on the overall effect of genetic loci. However it does not indicate which individual loci act. For those data we need to turn to molecular approaches.

1.2.2 Insights from genetic mapping

The progression from overall genetic contribution to individual loci requires genetic mapping experiments - linkage studies, candidate gene studies, and genome-wide association studies (GWAS). Perhaps surprisingly, we have most information about genetic architecture from studies of our own species. Genetic mapping has been carried out on a large number of phenotypes in humans, and has revealed a spectrum of architectures. On one side of this spectrum are Mendelian traits - traits controlled by a single locus; on the other side are polygenic (complex) traits, where large numbers of loci each make small contributions. Mendelian dominant conditions are very rare, recessive conditions less so. There are a small number of common Mendelian conditions, primarily those that have been subject to selection (e.g. alpha thalassemia in regions of endemic malaria may reach frequencies of 80%[30]), but in general the contribution to phenotypic variation within a population is rare[31]. However the molecular dissection of Mendelian phenotypes provides some important information about how genetic variants may act and shape genetic architecture.

Mendelian traits arise from mutations in single genes. However, each gene typically harbours many disease-causing alleles which are rare in the population[31]. The 1,942 mutations identified in the *CFTR* gene responsible for cystic fibrosis (CF) show how important allelic heterogeneity can be[32]. Often allelic heterogeneity leads to phenotypic heterogeneity: with CF, age at diagnosis, pancreatic sufficiency, and sweat chloride levels are highly correlated with *CFTR* genotype[33]. Disease alleles have non uniform frequencies in the population: for example, one *CFTR* allele (p.Phe508del) accounts for approximately 70% of CF alleles in Caucasian patients, while a group of 15-20 alleles accounts for an additional 15% of Caucasian cases[34]. Different alleles may even have opposite effects on the phenotype: missense alleles in the gene responsible for autosomal dominant hypercholesterolemia (*PCSK9*) have been associated with the disease, whereas nonsense mutations are associated with low plasma levels of LDL[34].

Phenotypic variation in Mendelian traits not only arises from allelic heterogeneity but also from the existence of modifier genes, which may influence age of onset (e.g. *CASP8* influences age of onset of breast cancers due to *BRCA* mutations[35]), and severity (e.g. 19q13 locus and cystic fibrosis[36]).

Environmental effects may also modify the phenotype. For example, a phenylalanine-restricted diet prevents phenylketonuria, a Mendelian disease associated with mutations at the PAH locus[37].

Finally, some studies have contributed to blurring the line between Mendelian and complex traits: first, diseases exist where multiple genes can individually cause the disease in a Mendelian fashion (e.g. *BRCA1, BRCA2* and breast and ovarian cancers[38]); second, a subset of the cases may have a Mendelian etiology and another subset a complex one (e.g. breast cancer[39]); third, diseases have been shown to arise from mutations in just two interacting genes (e.g. retinitis pigmentosa[40], and Bardet-Biedl syndrome[41]).

By contrast with our detailed information about the molecular basis of Mendelian genetic effects, molecular genetic studies of complex traits still provide only rudimentary insights into genetic architecture. In fact, as we will see below, most of the discussion still centres around the number of loci, their minor allele frequency and effect size, with almost nothing known about the importance of interactions and even less about the molecular basis of these variants.

To appreciate what GWAS tell us, we need to recognize they were designed under the ‘common disease–common variant’ hypothesis[42,43], which posits that common genetic variants (variants with minor allele frequency (MAF) >1%[10] or >5%[44]) cause common diseases. By testing all common variants, one could pinpoint key genes and shed light on underlying mechanisms.

Dense genotyping studies have shown that nearby variants formed a block-like structure due to lack of recombination within each block[45]. It was therefore proposed that only a subset of polymorphisms needed to be genotyped to tag each block. The International HapMap Consortium identified 500,000 SNPs that provide excellent power to test >90% of common SNP variation in out-of-Africa populations, with twice that number required in African populations[44].

GWAS have identified hundreds of genomic loci contributing to phenotypic variation[46] despite strict significance thresholds imposed to control the false positive rate when testing so many SNPs. In 2009, a review of 237 GWAS found that 151 of them had identified at least one trait-associated SNP, with a total of 531 trait-associated SNPs identified[47]. Therefore on average 2.2 loci were identified for each trait. The vast majority of associated variants were common with relatively small effects: the median risk allele frequency was 36% (interquartile range (IQR) 21%−53%) and the median odds ratio (for discrete traits) 1.33 (IQR 1.20-1.61)[47]. The median total sample size for these studies (initial and replication steps) was 7,858 individuals[47].

This raised the issue of 'missing heritability'[48]: it soon became clear that the variants identified in this first round of GWAS only explained a small fraction of narrow-sense heritability, as determined by quantitative genetics studies[48]. Missing heritability was apparent even in very large GWAS, such as that of height: with 63,000 participants, 54 associated loci were identified[49] but together they only explained about 5% of the heritable phenotypic variation (estimated to be 80%)[48]. A number of hypotheses were made to explain this observation, including a large number of common variants with tiny effects, rare variants with larger effects, structural variation not tagged by the polymorphisms assayed on genotyping chips, gene-gene interactions, and inadequate account for shared environment among relatives[48].

Visscher and colleagues provided an important element of response to the issue of missing heritability. They showed that a large proportion (40%) of the variation in height could be explained by all the (common) SNPs assayed on genotyping chips[50], while only 5% was explained by the loci passing the stringent significance thresholds of GWAS. This suggested that a large number of common variants each with a tiny effect contribute to variation in height. This hypothesis has prompted larger sample sizes in more recent GWAS[51,52]. Visscher and colleagues also showed that the remaining missing heritability in height was accounted for by incomplete linkage between genotyped markers and causal variants, which could arise if the causal variants were mostly at lower frequency than the genotyped markers (e.g. MAF 1%) - though not necessarily rare[50].

To investigate the hypothesis that low frequency variants (0.5% < MAF < 5%) might explain much of the variation in complex traits in humans, one needs a complete catalogue of low-frequency variants. This is now available, thanks to the 1000 Genomes Project[53], which identified most common and low-frequency variants segregating in 1,092 human genomes from 14 different populations[53]. The variants identified in the 1000 Genomes Project can be imputed in GWAS to test common and low-frequency variants more extensively[53]. Low-frequency variants have been associated with variation in many phenotypes[54]. They have not however explained much of the missing heritability, which remains an open question[13].

Finally, a contribution of rare variants to phenotypic variation is supported by theoretical considerations[55]. Because rare variants have small effect sizes across the population, no matter what their odds ratio is, they cannot be detected in the traditional GWAS design. Evidence for their contribution to phenotypic variation has come from testing for enrichment in rare variants in cases (burden tests), as illustrated by studies of schizophrenia[56], autism[57], HDL levels[58], and obesity[59]. Rare variants constitute the majority of SNPs segregating in the human population, consistent with recent explosive population growth[60].

While the number, allele frequency spectrum, and effect size distribution of causal variants has been extensively researched in human studies, other features of the genetic architecture of traits, such as interactions, are much less extensively described. There are three reasons to be interested in gene-gene, gene-environment, and gene-by-sex interactions. First, some loci might not be detected without taking them into account[61-67]. Second, they might lead to overestimating narrow-sense heritability and overestimating missing heritability[13,63,68]. Finally, they may provide insights into the etiology of the trait[61-63].

Gene-gene interactions, defined as statistical deviation from the additive combination of two or more loci in their effects on a phenotype[62], are difficult to detect genome-wide because of the burden of testing a very large number of combinations[61]. Consequently, most GWAS report only main effects. Epistatic interactions are sometimes tested between Quantitative Trait Loci (QTLs) with significant main effects[69,70], or in studies of candidate loci (e.g. Tsai *et al.* tested polymorphisms in genes of the renin-angiotensin system for epistatic effects on coronary artery disease[71]). Epistatic effects have mostly been documented in autoimmune diseases in humans[64-66,70,72].

Identifying gene-environment interactions may allow giving individualized preventive treatment before and after disease has been diagnosed. One highly cited and debated example of gene-environment interaction is the effect on the severity of depression of a variant in the serotonin transporter gene (*5-HTT*) that is only, or primarily, manifest in people who have suffered stressful life events[73]. Another is the flushing response seen following alcohol ingestion in individuals with low-activity polymorphisms in the aldehyde dehydrogenase gene[74]. Gene-by-sex interactions, which are more readily testable, exist for loci contributing to a variety of phenotypes[75].

Finally, what have molecular studies taught us about the extent of pleiotropy? Sivakumaran *et al.* recently reviewed pleiotropic action of SNPs reported to be associated with complex traits, based on the National Human Genome Research Institute's (NHGRI) Catalogue of Published GWAS[46]. They found abundant evidence of pleiotropy, with 16.9% of genes and 4.6% of SNPs in the catalogue showing pleiotropic effects[76].

Genetic architecture varies between African and out-of-Africa human populations mainly as a result of the genetic bottleneck that occurred during the exit from Africa[77]. Does genetic architecture also vary between species? No other wild population of animals has been used for mapping but laboratory populations of model organisms can shed some light on the question. The Drosophila Genetic Reference Panel (DGRP) consists of 192 inbred strains derived from a single outbred population[78]. Genetic mapping of three quantitative traits identified 203, 90, and 235 QTLs for each of the three measures, supporting a polygenic model of phenotypic variation. Most QTLs are at frequencies between 5 and 10%, and they explain most (65-90%) of the genetic phenotypic variation[78]. This is in stark contrast with human studies where most of the heritable variation remains unexplained even when imputation is used to genotype variants with frequencies down to 1%[53]. The explanation may lie in differences in the distribution of minor allele frequency in the DGRP and human populations: in humans, the majority of variants are rare while frequency is bounded at 0.5% (1/192) in the DGRP.

Other laboratory populations in which genetic mapping has been performed usually use fewer founders, typically two (e.g. F2 intercrosses and recombinant inbred panels) or eight (e.g. Heterogeneous Stocks[2,79], Collaborative Cross[80]). Therefore, minor allele frequencies are artificially constrained and most variants are common[81-83]. As a result, QTLs usually explain a large proportion of the heritable variation[2].

Of particular relevance to this thesis is the genetic architecture of the Northport mouse HS[2]. This published study indicates what the genetic architecture of complex traits might be in the rat HS, and allows comparison of genetic architecture between species. The mouse HS is an outbred population descended from eight *Mus musculus domesticus* inbred strains (A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL6/6J, DBA/2J, and LP/J) through more than fifty generations of outbreeding. Using 2,000 HS mice genotyped at 12,226 SNPs and phenotyped for 100 measures, Valdar *et al.* detected on average 8.7 QTLs per trait at a false discovery rate (FDR) of 30%[2]. The median of the proportion of phenotypic variance explained by the QTLs was between 2.5% (lower bound, accounting for the fact that relatedness inflates these estimates) and 13.7% (upper bound). The QTLs accounted for about three quarters of the heritable phenotypic variation.

1.3 Contribution of sequence information to the identification of causal variants and genes

There is an important dichotomy between the way in which sequence information helps identify the variants responsible for variation in Mendelian and complex traits. In the former, the causal variants are usually easily identified because they are mutations (i) known to be deleterious (ii) in annotated genes. The same is true for the causal variants detected in random mutagenesis screens in model organisms[84]. In both cases, the annotation of genes in the genome, the elucidation of the genetic code, and the prediction of damaging missense mutations[85] are instrumental in identifying the causal variants. For example, the mutation causing cystic fibrosis in Caucasians was singled out among other variants at the cloned locus because it was a 3-bp deletion that causes the loss of a phenylalanine residue in the predicted protein[86].

Today, advances in sequencing technologies make it possible to identify the mutation underlying a Mendelian disorder without going through the long process of mapping the causal locus to the genome, based solely on sequence variation. Thus Ng *et al.* identified the gene causing Miller syndrome, a rare multiple malformation disorder hypothesized to be an autosomal recessive disorder[87], by sequencing the exomes of four affected individuals from three independent kindred. They filtered the exonic variants using public SNP databases and HapMap exomes (which represent non affected individuals), and focused on deleterious variants (non synonymous variants, splice acceptor or donor site variants or coding indels). *DHODH* was identified as the causal gene since it is the only gene with two variants passing all filters in all four exomes[87]. *DHODH* encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway, which had not previously been involved with malformation disorders[87].

Initially, it was hoped that genetic variation contributing to complex traits would also be coding[88,89] - though variants were expected to have more subtle effects - so that sequence annotation would guide the identification of causal variants in complex traits too. In fact, most studies still focus on the interpretation of coding variants or other SNPs in transcribed regions, as illustrated by the large number of programs developed to interpret such variation[4,85,90,91]. This is the case despite abundant evidence to the contrary: 88% of SNPs associated with traits and reported in the NHGRI catalogue[47] are either intronic or intergenic. If GWAS are to realise their full potential, it is necessary to understand how genetic variation acts outside of coding regions.

The Encyclopedia of DNA elements (ENCODE) Project[92] was initiated to tackle this issue. Its aim was to identify all functional elements in the human genome sequence[93,94], particularly outside of known coding sequences. Based on experimental data, ENCODE identified regions of transcription, transcription factor association, chromatin structure and histone modification, which altogether cover 80% of the genome[94]. A more conservative estimate of what ENCODE has achieved comes from considering bases covered by a transcription factor (TF) binding site motif (4.6%) or a Deoxyribonuclease I Hypersensitive (DHS) footprint (5.7%) (8.5% of bases are covered by either)[94]. DHS sites reflect regions of open chromatin associated with active regions of regulatory DNA[95]. Transcription start sites (TSSs) are DHS sites when they are active. However, this is not where DHS annotation is most critical since TSSs and promoters can be relatively well predicted[96]. 95% of DHS sites are located distally to the TSS (intergenic or intronic)[97] and thereby constitute an important sequence annotation.

High-quality binding site motifs are available for only a minority of the >1,400 human TFs with predicted sequence-specific DNA binding domains[98], limiting the possibility to predict TF binding sites from sequence. Therefore, the TF binding sites identified by ENCODE have augmented the catalogue greatly. Enrichment of ENCODE regions in trait-associated SNPs (as reported in the NHGRI catalogue[47]) shows that this annotation of sequence is relevant to the interpretation of GWAS results: 12% of trait-associated SNPs but only 6% of 1000 Genomes SNPs overlap transcription-factor-occupied regions, and 34% of trait-associated SNPs but only 23% of 1000 Genomes SNPs overlap DHS sites[94].

However, the interpretability of TF binding sites and DHS sites remains limited: DHS sites do not tell what transcription factors bind to the element and therefore provide limited information to follow up experimentally GWAS hits. What is more, ENCODE showed that distal DHS sites and combinations of TF binding sites[99] are highly cell-type specific[97], suggesting that they are tissue specific *in vivo*. Their specificity to certain developmental stages has also been documented[100]. Therefore, function cannot be unequivocally attributed to a non-coding fragment based solely on the catalogue of ENCODE elements.

When a non-coding variant is associated with a complex trait and lies within an annotated feature, its function can be explored using experiments such as DNA electrophoretic mobility shift assays (EMSAs) or DNA pull-down assays with a particular transcription factor, or Chip-Seq to identify binding factors. For example, a variant associated with susceptibility to psoriasis was identified because it lay in a predicted RUNX1 binding site. EMSAs and reporter assays showed that the variant disrupted binding of RUNX1 and changed expression of a reporter gene under the control of the RUNX1 binding site[101], thereby supporting implication of this variant in phenotypic variation.

Complications arise in mapping populations where linkage disequilibrium decays slowly and loci are large, as is the case in most laboratory populations. Indeed, when a locus encompasses multiple functional elements (e.g. genes or regulatory elements), sequence annotation is not sufficient to identify the causal genetic factor. If sequence variation is available, every single variant in the associated region can be tested for association and the most highly associated ones prioritized for investigation(see Fig. 3C in[102]).

For this approach, it is critical to test all the variants that segregate in the population and lie in the associated region, and not only a subset of markers. While imputation using the 1,000 Genomes panel can genotype the majority of common and low-frequency variants[53], it does not give access to rare variants. Genotypes at all the variants in a region, including the rare ones, can be obtained by sequencing the associated region in the entire mapping population. However, sequencing costs remain prohibitive, so that usually only a subset of individuals are sequenced - commonly those in the tails of the phenotypic distribution[102]. Alternatively, a few individuals are sequenced to identify all varying positions, then a subset of variants deemed more likely to be causal based on prior knowledge are genotyped in the mapping population[102].

In laboratory populations descended from known founders, the genotypes at all segregating variants can be imputed by combining a reconstruction of the chromosomes of the outbred animals as a mosaic of the founder haplotypes with the sequences of these founders. Testing all imputed sequence variants identifies the candidates most likely to cause phenotypic variation. In crosses where few variants segregate (such as crosses between substrains[103]), this may point to a small number of variants, which may sometimes lie in a single functional element annotated in the region.

In populations descended from more than two founders, such as HS, in addition to testing each variant for association with the phenotype, one can test whether the variant captures the QTL effect as well as the founder haplotypes do. This approach, further explained in Figure 1-1 and Chapter 6, is called merge analysis[104]. Considering only those variants that capture the QTL effect as well as the founder haplotypes ('candidate' variants) allows fine mapping of the QTL, and may point to functional elements[104,105].

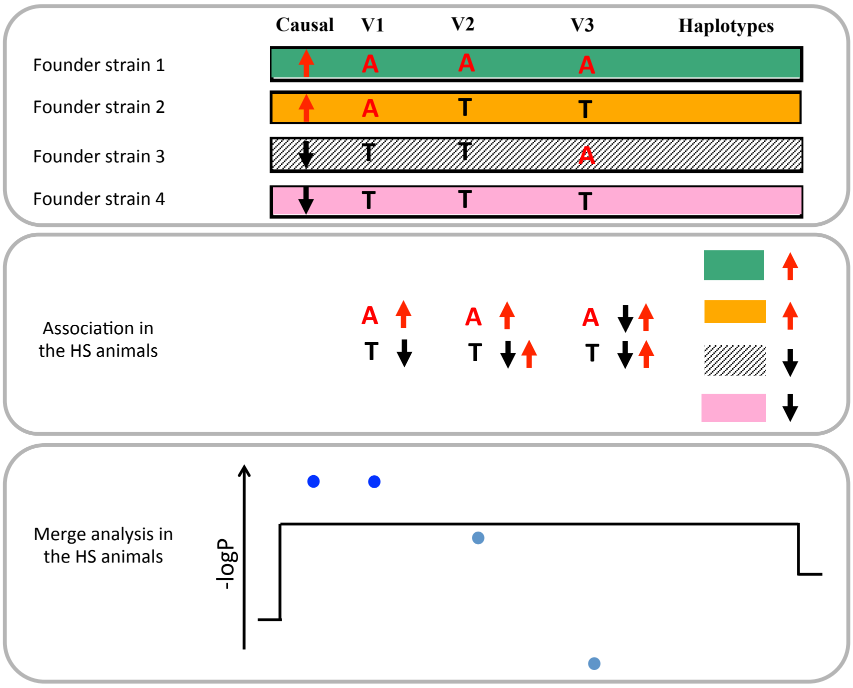


Figure 1-1 Merge analysis identifies 'candidate' variants. For simplicity a population with only four founder strains is shown. The upper panel shows a segment of the genomes of the founders, and four SNPs, one of which is causal for the QTL. The panel in the middle shows whether the SNP alleles and founder haplotypes cosegregate with the causal alleles in a mapping population descended from the four founders (HS animals). The lower panel shows the significance of the association between phenotype and founder haplotypes (black line), and between phenotype and each of the four variants (blue dots). The founder haplotypes always capture the QTL effect because the causal alleles segregate with the haplotypes they are on. The strain distribution pattern (SDP) of V1 is the same as that of the causal SNP (one allele in founders 1 and 2, the other in founders 3 and 4). Therefore, V1 and the causal SNP are perfectly correlated in the HS, genotypic variation at V1 explains phenotypic variation as well as haplotypic variation does, which leads to the association of V1 with the phenotype being more significant than that of the haplotypes (see Chapter 6 for mathematical justification). Contrastingly, the SDP of V3 is uncorrelated to that of the causal variants, V3 and the causal variant are uncorrelated in the HS animals, and V3 is not associated with the phenotype. The SDP of V2 is imperfectly correlated with that of the causal variant, so that V2 and the causal variant are imperfectly correlated in the HS. As a consequence, genotypic variation at V2 does not capture the QTL effect as well as the haplotypic variation does.

1.4 Conservation of causal variants and genes across species

One of the reasons for mapping complex traits in model organisms is that they should provide clues to the underlying variants in our own species. We know from knockout experiments that large effect mutations in homologous genes have similar, if not identical, phenotypes in different species (this assumption is a prime reason for carrying out mutagenesis experiments in the mouse as a way of understanding function in humans). We also know that loci underlying complex phenotypes include genes in which Mendelian mutations occur that give rise to the same phenotype. For example, *CYP17A1* was among the genes identified in a GWAS of blood pressure and hypertension[106]. A missense mutation and a 9-bp deletion in this gene are known causes of hypertension[107]. Based on these two observations, and the larger literature indicating conservation of physiological processes between species, it seems reasonable to expect that small effect mutations, the sort of genetic variant underlying complex traits, might exist at homologous loci and have similar phenotypic consequences in different species. The same sequence variants at exactly syntenic locations need not be involved, but functional variation should occur at the same locus (for example we might find in one species a variant in a gene promoter alters expression, while in another there is a variant in a 3’ control region of the same gene). The assumption is that variation might be constrained to occur in the same genes, reflecting the action of mutation and selection to generate similarly advantageous variants in the same genes. Thus identifying susceptibility loci in an animal model of a complex disease might indicate homologous loci that would contribute to disease in our own species.

Limited evidence supports this hypothesis, from studies where the gene underlying an association with disease has been identified in the mouse, and a significant difference in disease susceptibility or severity has been found between humans carrying one allele of the gene or another. Examples include a study of liver fibrosis involving the *Hc* gene[108], and a study of atherosclerosis involving *Tnfsf4*[109]. However, in these studies, the homologous locus in humans was not detected by a genome-wide test of association. Rare examples exist where a gene underlying an association with disease has been identified in the mouse and is also found at a locus identified in GWAS in humans (e.g. *Ctla4* for type I diabetes in humans and mice[110]).

There are also claims that "concordant" QTLs (homologous loci contributing to disease and detected by GWAS)[111] are common. For example, in a survey of 7 loci identified for atherosclerosis in mouse, Paigen and colleagues found that 5 overlapped with human GWAS loci[111]. Concordant QTLs have also been reported for kidney disease (rats, mice, and humans)[112], plasma levels of HDL, LDL, and triglycerides (mice and humans)[113,114], hypertension (rats, mice, and humans)[115,116], and bone density (mice and humans)[117]. While these studies give support to the idea that concordant QTLs are common, a more critical examination reveals problems. Given the size of the mouse and rat QTLs used in these studies (typically tens of megabases), overlap between many loci is expected by chance. In fact, only two[114,115] out of seven studies investigated the significance of the overlap, which they found was significant. However, these studies did not answer the question of whether QTLs detected in mice or rats can predict regions associated with disease in humans. Indeed, rather than testing whether QTLs mapped in mice or rats can predict human regions associated with disease, they showed that the human set of associated regions was significantly enriched in loci homologous with mouse/rat QTLs. Therefore, this question remains largely open.

In fact there are a number of reasons why we might expect concordant QTLs to be rare. First, when homologous loci exist in two different species, sampling fluctuations may mean that one species has no segregating variation. GWAS in model organisms typically use populations descended from a few founder inbred strains[83], and the variants that segregate in the laboratory population are only a subset of those in the species.

Second, when causal variants are sampled at homologous loci in two mapping populations, one locus may fail to be genome-wide significant. This could result from one of the studies having limited power to detect associations (e.g. because of a small sample size), or because the locus has a small effect size relative to the other loci segregating in the population.

Third, conservation may operate at the level of pathways rather than genes. There is evidence for conservation of pathways across species (e.g. Notch[118] and cell-death[119] pathways, and at a larger scale ageing[120] and response to cold[121] pathways). If pathways are conserved and selection happens at the pathway level, variants affecting phenotype could lie in different genes of the same pathway in different species.

Finally, some pathways and physiological processes are species-specific. For example, mouse models typically develop estrogen receptor (ER) negative tumours while rat models and humans develop ER-positive tumours[122]; Transgenic mice expressing either the human insulin receptor or a chimeric human/drosophila receptor have shown that differences existed between mammalian and insect signalling pathways downstream of the insulin receptor[123]; Differences between mouse and human pathways involved in embryonic stem cell self-renewal have also been identified[124].

1.5 Outline of thesis

Chapters 2 and 3 present an overview of genetic diversity in the HS founders and HS rats. Chapter 4 describes the phenotypes collected for this study and their characteristics. Chapters 2 to 4 report analyses carried out in preparation for QTL mapping and to investigate the genetic architecture of traits in the HS. Chapter 5 presents the results of QTL mapping, and Chapter 6 the integration of sequence information with the mapping results to identify the causal variants. Finally, Chapter 7 investigates conservation of causal genes and pathways across species. I discuss the main results of the thesis in Chapter 8.

Chapter 2

Genetic diversity among the eight founders of the rat HS

In this chapter I summarise the sequencing of the eight founders of the National Institute of Health (NIH) rat HS (presented in detail in Chapter 3),, and characterise the genetic variation that exists between them. These data give a near complete catalogue of variation segregating in the rat HS, which is essential for the identification of the causal variants, as explained in Chapter 6. I compare this variation to that in the founders of the mouse HS[2], and infer differences in the phylogenetic history of the rat and mouse laboratory strains. Victor Guryev performed the analyses presented in this chapter. I participated in writing up the characteristics of the founders, plotted the figures, and propose an interpretation of the results in terms of phylogenetic history of laboratory rats and mice.

2.1 Methods

Library construction, sequencing, alignment of the reads, and variant calling were performed by Edwin Cuppen's group at the Hubrecht Institute, Utrecht, Holland. Therefore, the methods are not presented here but are available in Baud *et al.*[125].

The variants existing between the founders of the mouse HS and used here for comparison with the rat are from Keane *et al.*[105], and are dated May 2011.

2.2 Results

We identified 7.2 million SNPs, 633,000 indels (<10 bp, with the majority consisting of 1-bp (79.3%) or 2-bp (12.3%) changes) and 44,000 structural variants in the accessible genome (defined in a similar way as for the mouse genomes[105]). Comparison of SOLiD and capillary sequencing of a bacterial artificial chromosome (BAC) library showed that 2.7% of SNPs, 2.2% of indels and 16.7% of structural variants were false positive calls. False negative rates were much higher: 17.2% for SNPs, 41.4% for indels and 65% for structural variants. Most false negative SNPs and indels are next to repeats (77.9% and 80.8%, respectively).

The variation present in each strain is summarized in Table 2.1. Excluding BN/SsN (which is a substrain of the reference and consequently has far fewer differences than the other strains), the average number of SNPs per strain was 2.8 million.

Strain	Gb of mapped data	Coverage	% of genome inaccessible	SNPs	Private SNPs	Indels	Private indels	Structural variants	Private structural variants
ACI/N	65.9	26.3	12.6	2,883,405	228,468	166,425	12,646	19,499	756
BN/SsN	54.4	21.7	9.4	71,038	563,308	0	14,839	27	4,203
BUF/N	62.3	24.9	12.7	2,748,633	125,202	172,934	7,195	22,176	1,002
F344/N	77.9	31.1	11.8	2,831,144	97,951	157,522	5,007	25,257	1,003
M520/N	72.5	28.9	12.3	2,836,898	89,277	170,031	5,008	24,090	915
MR/N	62.4	24.9	12.3	2,664,124	223,514	151,099	12,005	18,306	1,004
WKY/N	63.4	25.3	12.1	3,088,953	496,327	164,634	23,979	28,270	3,357
WN/N	62.3	24.9	12.2	2,698,493	249,563	154,769	13,541	18,563	700

Table 2.1 Sequence variation in the eight progenitor strains of the HS. Shown for each strain is the amount of sequence mapped to the reference, the coverage, the percent of the genome deemed inaccessible and the counts of the three classes of variants compared to the reference strain. Private variants are variants that distinguish a specified strain from all others; most of the alleles private to Bn/SsN are reference alleles.

We examined sequence diversity among the HS founders, identifying the following characteristics of this diversity. First, diversity between any pair of rat strains is approximately equal, so that there are no extremely sequence divergent strains (Figure 2-1a). This result differs from that previously obtained by the STAR consortium after genotyping 167 laboratory rat strains at 20,283 sites[126], which showed that BN was particularly divergent. The likely explanation is that the result obtained by the STAR consortium, based on genotyping data, was biased by SNP ascertainment. Second, regions of low diversity were small (median of 400 kb), with no blocks over 35 Mb in length (Figure 2-1b). Third, within divergent regions, there was a median of 151 differences per 100 kb (Figure 2-1c), and an average of 3.5 haplotypes (Figure 2-1e).

Genome-wide, the eight inbred strains that founded the mouse HS[105] are also approximately equally divergent (Figure 2-1d). However, they are more diverse than the rat founders (10.2 million SNPs), and mouse diversity is less homogeneous across the genome: long tracts of identical haplotypes alternate with segments with many SNPs (Figure 2-1b and Figure 2-1c). In these divergent segments, the average number of haplotypes is lower than that in the rat founders (2.7 in the mouse compared to 3.5 in the rat, Figure 2-1e).

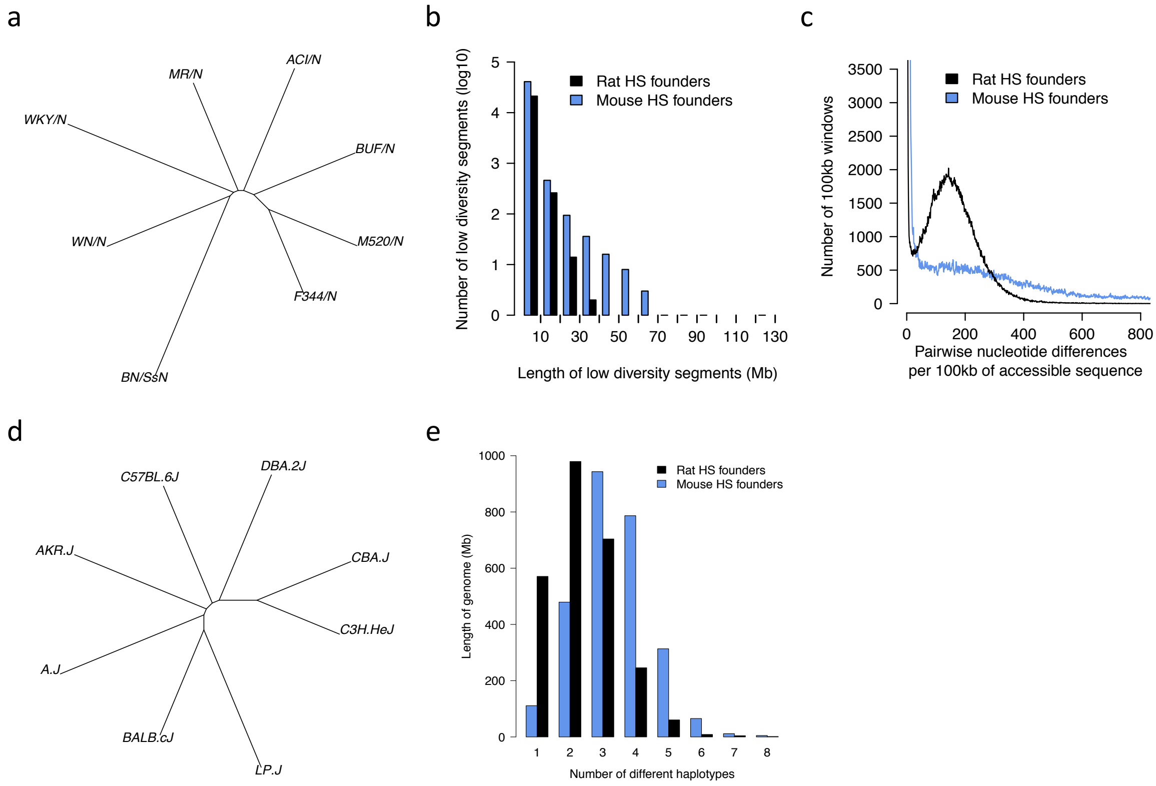


Figure 2-1 Sequence diversity among the progenitors of the rat and mouse HS. (a) Divergence between the rat HS founders. Neighbour-joining tree based on pairwise sequence distances (SNPs/kb of accessible sequence). (b) Length distribution of regions of low diversity in rat and mouse HS founders. The x axis shows the length of genomic regions with little sequence divergence (less than 13 SNPs/100kb). The y axis shows the numbers of segments observed in the progenitors. (c) Local sequence divergence in the rat and mouse founders. The x axis shows the number of pairwise sequence differences observed in 100kb windows. The y axis gives the number of observations. (d) Divergence between the mouse HS founders. Neighbour-joining tree based on pairwise sequence distances (SNPs/kb of accessible sequence). (e) Genome-wide distribution of the number of different haplotypes in the rat and mouse founders. The x axis shows the number of locally different haplotypes between the founders, and the y axis shows the length of genome (Mb) with this number of haplotypes.

2.3 Discussion

In this chapter we characterised genetic diversity in the eight founders of the rat HS, and compared with diversity in the founders of the mouse HS. I will discuss in Chapter 8 the relationship between genetic diversity in the founders, QTL mapping results, and the identification of the causal variants. In this section I propose an interpretation of sequence diversity in terms of phylogeny of the rat and mouse founder strains.

The differences between the founders of the rat and mouse HS populations shown in Figure 2-1 suggest differences in their phylogenies. Classical inbred strains of mice (that include all the HS mouse founders) were derived during the twentieth century from a small population of fancy mice, which showed substantial inbreeding[127,128]. Therefore, they show large multimegabase segments that are identical by descent, and limited haplotype diversity in those segments that are diverse. But because multiple subspecies contributed to the genetic variation present in mouse strains[105,127], those segments that are diverse harbour a large number of SNPs (but those only distinguish between a few haplotypes).

In contrast, laboratory strains of rats are descended from outbred stocks bred in the laboratory as early as 1850[129], most likely from wild caught animals of *Rattus norvegicus* origin[130]. Based on the larger haplotypic diversity found in the rat HS strains compared to the mouse strains, I suggest that these rat strains had a larger number of ancestors than did the mouse strains, but those ancestors were less divergent, as suggested by the smaller number of SNPs segregating in the rat founders.

Genetic diversity in the founders of the HS, together with the breeding scheme used to produce the outbred population, are the major determinants of genetic diversity in the outbred population, which I describe in Chapter 3.

Chapter 3

Genetic diversity in HS rats

In this chapter I characterise genetic variation in the HS rats, and analyse the HS genotype data in preparation for QTL mapping.

3.1 Background

3.1.1 Genotyping in the laboratory rat

The rat HS is descended from eight inbred strains through more than fifty generations of outbreeding. As a result, a large number of genetic variants is expected to segregate in the cross, most of them being of high frequency because the lowest-frequency variants segregating in the founders will have a minor allele frequency (MAF) of 1/8. Linkage disequilibrium (LD) between neighbouring polymorphisms is expected to decay within a few megabases[79]. In order to capture the genetic variation that segregates in the HS, it is necessary to genotype a large number of polymorphisms representative of the variation existing between the eight founders.

Historically, genotyping in the laboratory rat was carried out using simple sequence length polymorphisms (SSLPs) and restriction fragment length polymorphisms (RFLPs). The first complete genetic map obtained in a single population was published in 1995 for an SHR x BN F2 intercross and used 432 SSLPs[129]. It was soon followed by a genetic map for the

SHR x BN (HxB) and BN x SHR (BxH) recombinant panel (454 SSLPs[131]). The map was subsequently updated to include more markers and physical locations[132] following the release of the draft reference sequence for the rat in 2004[133]. More recent studies have used single nucleotide polymorphisms (SNPs)[134], making use of the SNP catalogue and physical and genetic maps made available by the STAR consortium[126]. The STAR consortium generated sequence data for 13 rat inbred strains and identified 3 million SNPs mapped to the reference sequence. They also genotyped 167 laboratory strains and two panels of recombinant inbred strains (the HxB-BxH panel - 31 strains - and the other commonly used F344 x LE - Le x F344 panel - 33 strains) and 89 rats of a BN x GK F2 intercross using an Illumina BeadLab station (10,752 SNPs) and a 9,691-SNP Affymetrix array. Since the strains used by the STAR consortium are among the founders of most of the populations used by the rat community, the SNPs they have identified have replaced SSLPs in the most recent genetic studies[79,134,135].

The Illumina and Affymetrix chips used by the STAR consortium are no longer commercially available. Instead, recent studies have used the Sequenom MassArray platform to genotype up to a few hundred markers[79,135]. Previous work on the mouse HS[2] and a pilot project carried out with 800 HS rats[79] suggested that at least 10,000 markers were necessary, which is confirmed *a posteriori* (see section 3.3.1). The development of the high-density Affymetrix Rat Diversity (RATDIV) array (803,485 SNPs) by our collaborators at the Max Delbruck Center (MDC), Berlin was critical for the success of this study. This array is commercially available for genotyping other rat populations.

3.1.2 Haplotype mapping

Even with such a high density of markers, there are risks of missing a genetic association if a causal variant is not well tagged by the markers on the array. This can happen if the strain distribution pattern (SDP) of the causal variant in the eight founders is uncorrelated with those of all the nearby genotyped SNPs (e.g. V3 in Figure 1-1). This is one reason why ancestral haplotype mapping is generally more powerful than single point analysis in the HS: because the stock is descended from eight known inbred strains, genetic variation in the cross segregates in the form of ancestral haplotype blocks (ancestral here means of the eight founders of the cross), with the exception of a very limited number of *de novo* mutations discussed at the end of this section. Testing for association with the eight ancestral alleles after reconstructing the chromosomes of the outbreds as mosaics of the ancestral genomes overcomes this problem, since the alleles of the QTL segregate together with the ancestral haplotypes (Figure 1-1).

The reconstruction of the chromosomes of the HS rats as mosaics of the eight founder genomes is carried out using a dynamic programming algorithm implemented in the R package HAPPY[136]. It is important to note that HAPPY does not rely on pedigree information, which is only available for the final eight generations of the HS. HAPPY reconstructs the chromosomes of the HS animals as a probabilistic mosaic of eight ancestral haplotypes. It outputs the probability $P_{L,i}(s,t)$ that the chromosomal segment at locus L (defined as the interval between two successive genotyped markers) of animal i is descended from founders (s,t). Thus, $Q(s) = \Sigma_t P_{L,i}(s,t)$ is the expected number of haplotypes of type s carried by rat i at locus L. The sum of these expectations for a diploid given animal at a given locus is always 2. Under the assumption that the phenotypic effect of the pair of haplotypes is the sum of the effects of the members of the pair (i.e. no interaction between the two alleles, "additive model"), we can work with $Q(s)$ instead of $P_{L,i}(s,t)$ which simplifies calculations. Examples of the reconstruction of the HS chromosomes as mosaics of the eight founder haplotypes are shown as heat maps in Figure 3-5.

The certainty with which a chromosomal segment can be attributed to a particular founder depends on how many markers have been genotyped at and around the locus, but also on how similar the founder genomes are at the locus: the more similar they are, the more markers are needed to distinguish between them.

Once a haplotype has been recognised, one can be confident that any association with it will be detected (provided it is strong enough) without needing to have markers that cover all possible SDPs. For example, if there were four instead of eight founders, identifying a haplotype would require at most three markers with different SDPs, while covering all possible SDPs would require fifteen (Figure 3-1).

Founder strain 1	+	-	-	-	+	+	+	-	-	-	+	+	+	-	+
Founder strain 2	-	+	-	-	+	-	-	+	+	-	+	+	-	+	+
Founder strain 3	-	-	+	-	-	+	-	+	-	+	+	-	+	+	+
Founder strain 4	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+

Figure 3-1 Number of SNPs necessary to tag all possible SDPs and to identify a founder haplotype. If there were 4 founders, it would take 15 SNPs to tag them all, but it only takes 2 or 3 to identify a founder haplotype.

The same combinatorial result extends to the eight-way cross and means that fewer markers are required to capture the genetic variation that segregates in the cross in terms of ancestral haplotypes than in terms of genotypes at markers. Finally, because the probabilities of descent at a given locus are computed using information at the locus and around, missing genotypes and errors can be overcome using information from neighbouring genotypes. In conclusion, haplotype reconstruction and haplotype mapping overcome a number of limitations of single-marker analysis.

Some *de novo* mutations certainly have arisen in the 50 to 60 generations over which the cross was bred, and some might be responsible for additional phenotypic variation. Since not all animals with the same haplotype carry the mutation mapping phenotypic variation to the ancestral haplotypes would result in lower power to detect any associations due to a *de novo* variant. This would also be the case with single point mapping, since markers on the array are those present in the founder strains. Although associations with *de novo* mutations are difficult to detect, they should not explain much of the phenotypic variation in the HS, because only a limited number could have arisen in the 50-60 generations of breeding of the HS.

3.2 Methods

3.2.1 Animals

For this project, a colony of HS rats[137] was established at the University Autonomous of Barcelona, Spain starting with forty breeding pairs of the 52nd generation. The present study started with rats of the 62nd generation (rats of the 52nd - 61st generations were used for the pilot project). Animals included in this study were bred over eight generations: at each generation and from each sibship, one male and one female were kept to breed the next generation while the remaining rats were phenotyped and genotyped (Figure 3-7). Animals were usually housed two in a cage (males) or three in a cage (females), with siblings sharing cages. Animals were maintained with food and water *ad libitum*, under conditions of controlled light and temperature.

The pedigree over the last eight generations was recorded on paper and the information uploaded into a database together with the phenotype data. The pedigree in Figure 3-7 was plotted using the R package “kinship” and its function “pedigree”. For this function to work, the pedigree has to pass a number of sanity checks. Errors in the records were thereby uncovered and later corrected by going back to the paper notes kept in Barcelona.

3.2.2 Design of an array to genotype the rat Heterogeneous Stock

An Affymetrix custom genotyping array was designed to genotype the rat HS and other populations (described in detail in Baud *et al.*[125] - Supplemental Information Development Rat Array). Polymorphisms assayed by the array were selected based on (partial) sequence information available from the STAR consortium[126] and other sources for fourteen rat strains, three of which are amongst the HS founder strains (Brown Norway BN, Fischer F344, and Wistar Kyoto WKY). Work by the STAR consortium had shown that variation segregating between these 14 strains was likely to be representative of all the laboratory rat strains, so it was expected that the markers assayed by the array should allow us to distinguish between the eight ancestral haplotypes. 803,485 SNPs are assayed with the array. Their position is given on the Rnor3.4 reference genome assembly[133].

3.2.3 Genotyping

Because of budget limitations, a subset of only 1,407 HS rats were selected for genotyping out of the 2,006 rats that had been phenotyped. Animals were not genotyped if they died before the end of the protocol (because fewer phenotypes were available), or if they were part of non-nuclear families because they created an additional level of structure in the population. Rats with only one kidney at harvesting were always included in an effort to map the genetic basis of renal agenesis. One hundred pairs of parents were also selected for genotyping, based on the number of their genotyped offspring. The eight founders (the original individuals from which the cross is derived) were also genotyped but the genotyping of the Brown Norway (BN) founder failed due to poor DNA quality. No tissue was available from the original HS founder to re-extract DNA, so the reference sequence (also BN) was used in place of the genotypes from the BN HS founder. DNA was extracted from liver samples, and genotyping was carried out at two centres (the Centre National de Genotypage (CNG) in Paris, France and the MDC in Berlin). A few animals were genotyped in both centres for quality control purposes.

3.2.4 Genotype calling and quality control

Genotypes were called from the CEL files using Affymetrix Power Tools (APT) version 1.14.2, and its BRLMM-P algorithm[138]. This algorithm uses information from all the CEL files genotyped in one run as well as priors on the position and scatter of the clusters used to call individual genotypes. The genotypes of the HS rats were called together with a number of other inbred and outbred rats so as to improve the quality of the calls.

A subset of high quality markers were selected for subsequent analysis. This was motivated by the need for high-quality founder genotypes in order to ensure accurate haplotype reconstruction. DNA sequences of the seven non-reference founders (i.e. all founders except for BN due to poor DNA quality) were used to identify markers whose array-called genotypes were concordant with the sequence-called genotypes. We thereby discarded polymorphisms not detected in the sequence data or with missing data in some founders.

We only retained those markers for which BRLMM-P called high-quality genotypes: Affymetrix genotyping relies on differential hybridization of the sample DNA to two types of probes, those complementary to the reference sequence and those complementary to the sequence with a SNP. Based on these two hybridization signals (called A and B), at each SNP the samples are clustered in a maximum of three classes (two homozygote and one heterozygote classes), and their genotypes called (Figure 3-2). Cross hybridization of the DNA to both probes or the presence of unknown SNPs in the DNA can result in poor clustering and unreliable calls (Figure 3-2). Thus markers that failed to show three large, well resolved clusters (with at least nine samples in each homozygote class, three in the heterozygote class and Fisher linear discriminant greater than 4) were filtered out. We also set thresholds on the "Heterozygotes Offset" value output by BRLMM-P for each marker (it had to be greater than -0.5) and on the call rate (which had to be higher than 99% for autosomal markers, and over 95% for markers on the X chromosome - there were no markers on the Y chromosome).

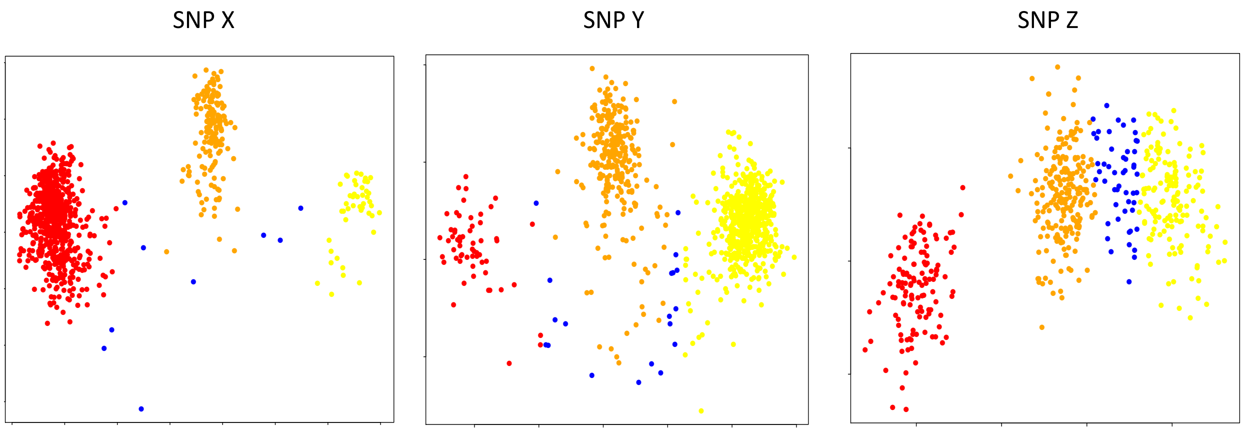


Figure 3-2 Transformed hybridization signals used by BRLMM-P to call genotypes for three representative RATDIV SNPs. The two axes are transformed hybridization signals (A-B and (A+B)/2). Each dot corresponds to a sample. The genotypes are called based on the clustering of the samples. Good quality markers show 3 large, tight, and well-defined clusters (e.g. SNP X).

Next, we used the genotypes of the 100 pairs of parents that were available to exclude any marker with over four Mendelian errors (inconsistencies between parental and offspring genotypes). This threshold was quite relaxed so as not to penalize a genotyping error in one of the parents that would lead to genotype inconsistencies with each of the offspring. Finally, we discarded SNPs for which any of the seven non-reference founders was heterozygote, as well as any SNPs that were monomorphic across the seven founders. The latter filtering should have been done across all eight founders because it resulted in the exclusion of all SNPs private to the reference strain Brown Norway. As shown in the Results section of this chapter, this did not prevent the reconstruction of BN haplotypes. This filtering yielded a subset of 265,551 high quality markers for subsequent analysis.

3.2.5 Linkage disequilibrium analysis

LD between neighbouring SNPs was calculated based on the genotypes of the 265K autosomal markers and using PLINK[139].

3.2.6 Allele frequencies

490K markers polymorphic between the founders were used out of the full set of 803K markers to calculate minor allele frequencies because the 265K markers of the clean set were selected to have three "common" genotypes.

3.2.7 Calculation of the pairwise genome-wide genetic similarity

Pairwise genome-wide Identity By State (IBS) was calculated using the R package EMMA[140] based on the genotypes at the 265K selected markers. Various options to deal with missing genotypes exist but all led to similar results because of the very high call rate. The default option whereby the missing alleles are imputed based on the minor allele frequencies was used. IBS was estimated under an additive model where AA/AA has IBS=2, AA/AB IBS=1, and AA/BB IBS=0. For comparison purposes, an IBS matrix of haplotype similarities was calculated by averaging the correlation (or distance) between the descent probabilities of rats i and j at locus L across all the loci in the genome.

3.2.8 Ancestral haplotype reconstruction

Haplotype descent probabilities were initially calculated by HAPPY for 265K intervals (where an interval is bounded by two adjacent genotyped markers) but were then averaged over 90kb windows to yield 24,196 probability matrices. We show in section 3.3.1 that the averaging made the QTL mapping analyses much faster but did not markedly reduce the sensitivity or resolution of QTL detection.

3.3 Results

3.3.1 Decay of linkage disequilibrium

As expected from the large number of generations of breeding, LD decays quickly in the HS, falling below 0.5 within 0.5 Mb (Figure 3-3a). However, we observed some correlation between distant markers, which has the potential to create spurious associations. Thus R^2 is greater than 0.2 at 0.6% of the marker pairs on different chromosomes as well as at 7.3% of the marker pairs that are on the same chromosome but more than 10 Mb apart. This result highlights the need to correct for relatedness when mapping in the HS. Additionally, four pairs of loci show very high interchromosomal LD. Using UCSC liftover tool, these regions were shown to map in the latest rat reference genome assembly (RGSC 5.0) to the regions with which they were in high LD in the current assembly (Rnor3.4). Therefore, they correspond to errors in the Rnor3.4 assembly. These loci were excluded from the analysis (APPENDIX A).

The observed decay of LD allowed us to check whether the 265,551 selected markers capture most of the genetic variation segregating in the HS and that averaging the descent probabilities over 90kb regions does not lead to loss of information. While the distribution of the length of the haplotype blocks would be an ideal measure to use for this purpose, this distribution is not directly available from the HAPPY probability matrices because the haplotype reconstruction is probabilistic and genetic similarity between the founders creates uncertainty. So we used information about LD instead.

The 265K SNPs are spaced on average about 10kb apart in the 2.7Gb rat genome. The median LD between markers separated by 10kb is greater than 0.995. Therefore, additional markers would be redundant. LD is still greater than 0.85 for markers 90kb apart, indicating that haplotype blocks are usually larger than 90kb and that averaging does not lead to loss of power, nor compromises mapping resolution. It does make association testing computationally less demanding however. All the values above are average or median values. Long stretches of genome without variants segregating in the founders exist (see Chapter 2), and mapping resolution in these regions will be poor (but equally, we do not expect many QTL to segregate in them).

Figure 3-3a shows that the decay of LD in the rat and mouse HS is very similar, but much slower than that in humans (Figure 3-3b, reproduced from Ke *at al.*[141]).

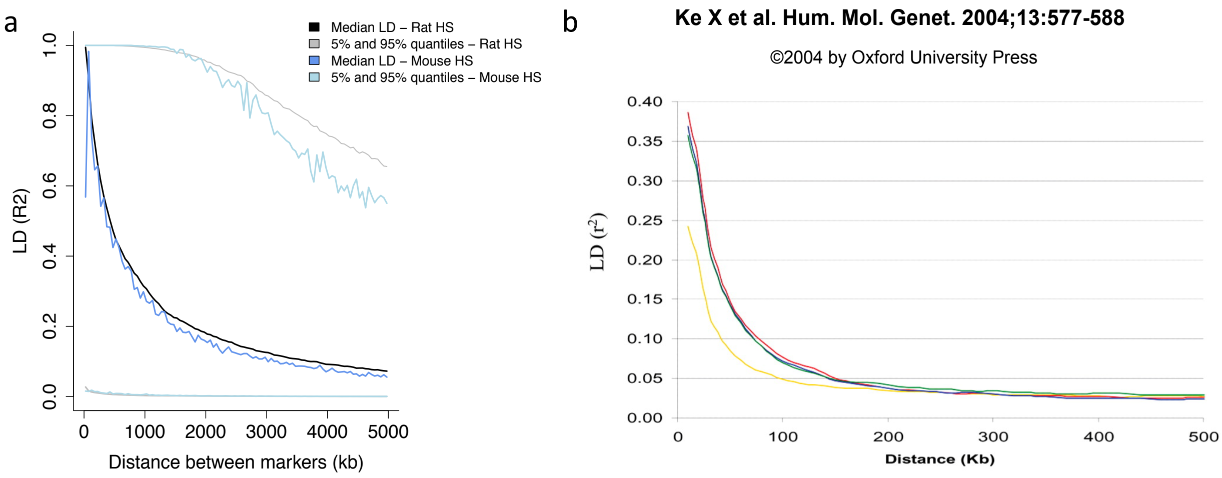


Figure 3-3 Decay of linkage disequilibrium (LD). LD was measured as R^2 between pairs of SNPs separated by distance plotted on the x-axis. (a) Decay of LD in the rat and mouse HS. (b) **Figure taken from Ke *et al.*[141]** Decay of LD in humans. Note the use of different scales on both axes between the panels.

3.3.2 Minor allele frequencies

Figure 3-4a shows that the vast majority of alleles are common in the rat and mouse heterogeneous stocks, as would be expected given their origins. Figure 3-4b shows the relationship between MAFs in the founders and in the HS rats. 4.4% of the markers polymorphic across the rat HS founders are fixed in the HS (where fixation is defined as MAF < 0.01).

The distribution of allele frequencies in the rat HS is similar to that in the mouse HS but differs radically from that observed in the (European) human population (as assessed from the 1000 Genomes Project[53], Figure 3-4a). While most alleles are common in both HS, genetic variation in humans is dominated by rare variants.

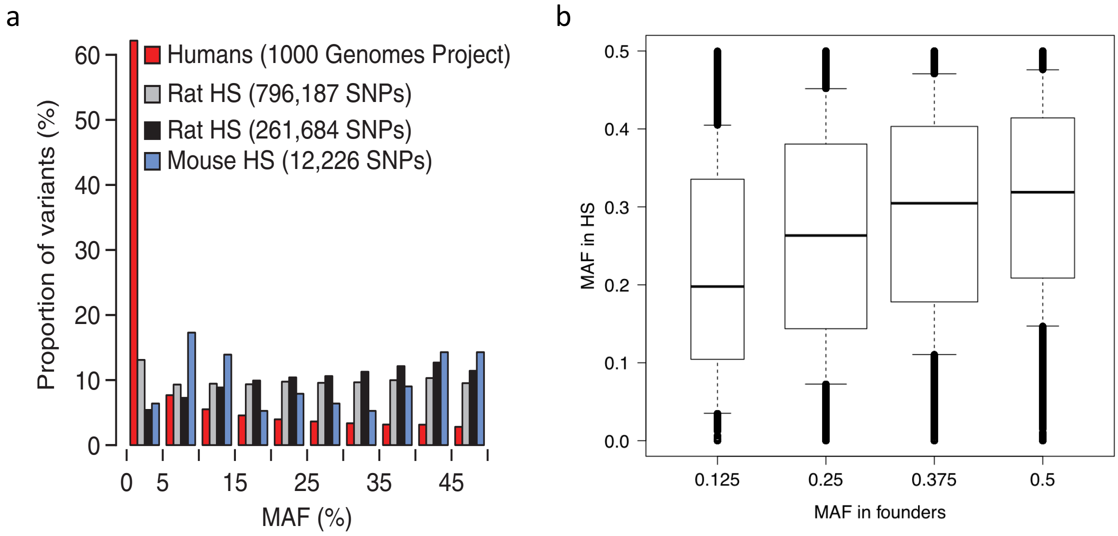


Figure 3-4 Minor allele frequencies (MAFs). (a) Minor allele frequencies in the rat and mouse HS and in the European human population. (b) MAF in the rat HS as a function of the MAF in the eight founders.

3.3.3 Ancestral haplotype reconstruction in the HS rats

Figure 3-5 shows the probabilistic reconstruction of 15 Mb of chromosome 1 for four HS rats into the eight ancestral haplotypes under an additive model (i.e. dosages of each of the founder haplotypes are shown). It illustrates the wide variation in haplotype block lengths within an individual, as well as the uncertainty that can arise in the reconstruction when the founders are locally very similar. The equally high probabilities observed for BUF and ACI between 55 and 60 Mb for three of the rats result from the fact the haplotypes of ACI and BUF are indistinguishable in this region, as confirmed with the genotype data (Figure 3-6).

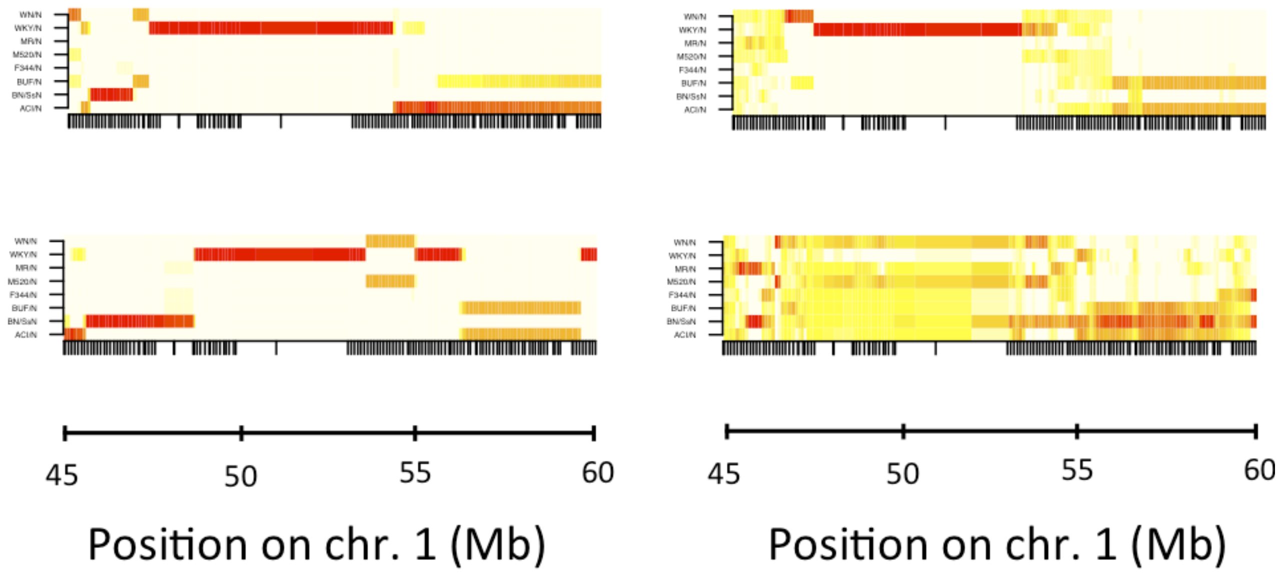


Figure 3-5 Reconstruction of the chromosomes of the HS rats as probabilistic mosaics of the eight founder haplotypes. Four 16 Mb long chromosomal segments corresponding to four HS rats are represented. For each animal, the dosage (probability of descent) of each of the eight founders (vertical axis) along the chromosome (horizontal axis) is represented as a colour (white = 0, red = 2, yellow ~ 1).

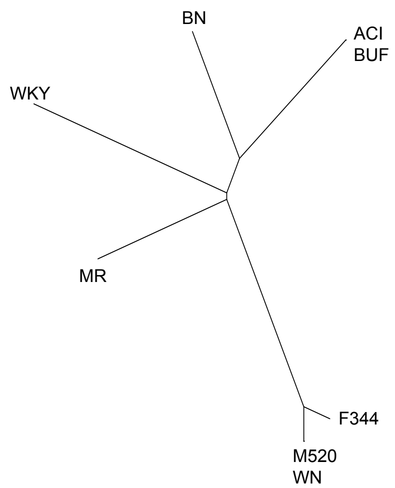


Figure 3-6 Local genetic similarity between the founders of the rat HS. A 100kb window encompassing 14 markers was chosen to illustrate how the founders locally cluster in groups of greater similarity.

3.3.4 Pairwise genome-wide genetic similarity in the HS rats

The pairwise genome-wide genetic similarity between all pairs of HS rats is represented as a colour-coded matrix in Figure 3-7. The rats are ordered along both axes by generation and family, so that siblings are adjacent. As expected, the figure shows that siblings are more genetically similar to each other than they are to unrelated HS rats. It also makes the final eight generations of breeding visible as off-diagonal lines parallel to the main diagonal. The existence of various levels of genome-wide genetic similarity in the HS implies that genotypic correlations exist between a large number of sites. Such correlations can create spurious associations if one of the loci is responsible for phenotypic variation. Accounting for pairwise genome-wide genetic similarity with a mixed model helps to control such associations (Chapter 5).

However, pairs of rats with little genome-wide genetic similarity can still be similar at specific loci, following partial fixation of pairs of haplotype blocks within subsets of the population during the breeding[142]. This translates into long-range correlations that are not well described by the IBS matrix and therefore poorly accounted for by mixed models. As we will see in Chapter 5, one way to avoid spurious associations arising from such correlations is to let the SNPs compete in a multilocus model for phenotypic variation[142].

The comparison with the IBS matrix calculated for the HS mice indicates clearly that a different breeding scheme was used: a limited number of breeding pairs were repeatedly mated to breed the HS mice that were genotyped and phenotyped, resulting in larger clusters of full-sibs. Interestingly, in the HS mice, clear patches of high-similarity can be observed outside the main diagonal, highlighting the presence of additional levels of genetic similarity.

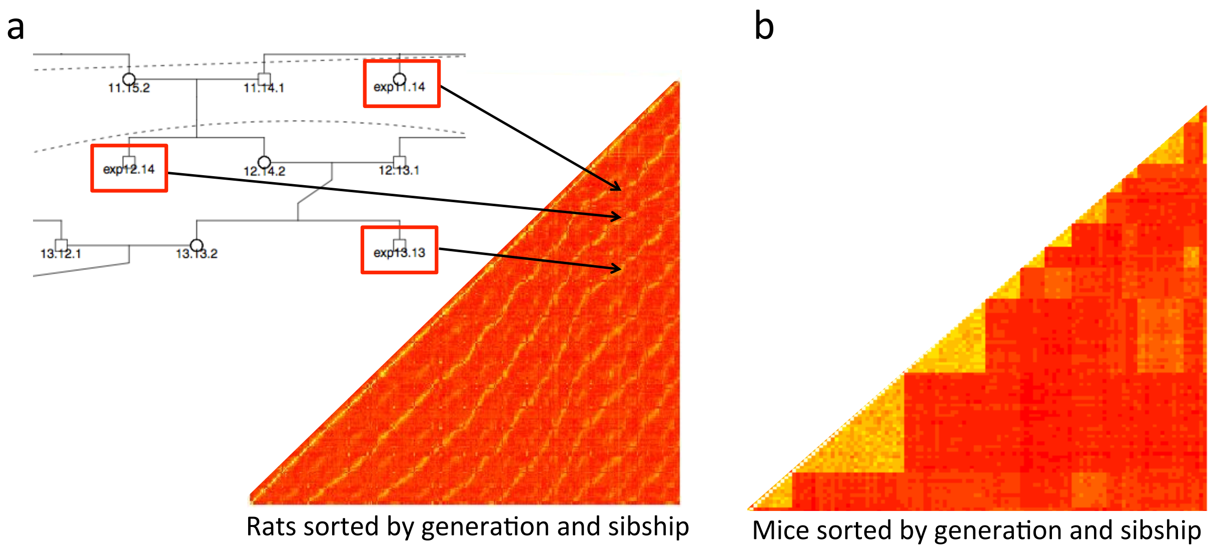


Figure 3-7 Pairwise genome-wide genetic similarity in the rat and mouse HS. Genetic similarity was calculated using Identity by state (IBS) Colour code: genetic similarity increases from red to yellow. The animals are sorted by generation and sibship along both axes. Only the lower triangle of the symmetric matrix of genetic similarities is shown. (a) Genetic similarity in the rat HS. An extract from the pedigree is shown. The animals are identified based on their generation, family, and sex. For example "11.15.2" indicates a breeding female of family 15 in generation 11. "exp" refers to the group of siblings that were not used to breed the next generation but genotyped and phenotyped. (b) Genetic similarity in the mouse HS.

3.4 Discussion

In this chapter I characterised genetic diversity in the HS rats. Four important features emerge from this analysis of the HS rat genotypes: (i) LD decays quickly in the HS; (ii) allele frequencies are high and are correlated with MAFs in the founders; (iii) genetic variation segregates in the HS as blocks of the founder haplotypes, and the certainty with which we can reconstruct each HS chromosome as a mosaic of the founder genomes depends on local genetic similarity in the founders; (iv) different degrees of genetic similarity exist in the HS and reflect different levels of relatedness. I will discuss the implications for QTL mapping and the identification of the causal variants in Chapter 8.

Chapter 4

Phenotypes of the HS rats

4.1 Background

In this chapter I analyse the phenotypes of the HS rats collected by the EURATRANS consortium. These analyses provided a necessary starting point for QTL mapping. I start by introducing the phenotyping pipeline, and discuss the identification of covariates, and normalization of the data for genetic mapping. I then describe the estimation of heritability.

4.1.1 Phenotyping pipeline

Inbred strains have a fixed genetic make-up, so that phenotypic measures can be collected on as many animals as necessary while genotyping is carried out only once. In contrast the outbred HS rats are genetically unique, so that all the rats that are phenotyped also need to be genotyped to test for genotype/phenotype associations. A phenotyping pipeline was set up so as to maximize phenotypic information whilst amortising genotyping costs. This study was carried out in collaboration with eight groups (seven European and one American) whose research covers a broad range of biology and a number of models of disease. The collaborating groups agreed on a set of measures to be collected on each genotyped individual (Table 4.1). The phenotyping pipeline was carefully designed so as to minimize the effect of collecting one measure on others collected later[79]. Another element that directed the choice of the phenotyping protocols was the need to phenotype more than 2,000 rats. For example, cardiac function is best investigated using cannulation and telemetry but these methods take too much time, equipment, and require advanced skills in surgery, so they were not used in this study. Blood pressure was measured using a tail cuff instead[143].

4.1.2 Covariates

Because rats were phenotyped in eight batches many months apart (each batch corresponding to a generation of breeding), most measures were collected by multiple experimenters and always over many days. Variation in the experimental protocols was thus expected. Moreover, the rats varied by sex, date of birth (and therefore age on the days the measures were collected), body weight, etc. Together with experimental variation, this physiological variation is likely to contribute to phenotypic variation, and thereby decrease the relative contribution of genetic factors, and make genotype/phenotype associations harder to detect. Some of this difficulty can be alleviated by accounting for these factors, by including covariates in the analysis. However, not all the experimental and physiological covariates are measured, and environmental effects might vary despite conditions in the animal house being controlled. Any unaccounted for non-genetic variation will obscure genotype/phenotype associations.

4.1.3 Heritability

It is useful to estimate the fraction of phenotypic variation due to additive genetic effects (narrow-sense heritability) prior to mapping any association. Narrow-sense heritability constitutes an upper bound to the proportion of phenotypic variation that the QTLs can explain[2,144,145] (also see Chapter 1).

Estimation of heritability is traditionally based on known familial relationships. The expected proportion of alleles shared identical by descent (IBD) corresponds to the degree of relatedness (coefficient of kinship)[6,7]. For example, full siblings are expected to share half of their alleles, while first cousins only share 12.5% on average. Narrow-sense heritability, the proportion of phenotypic variance attributable to additive genetic effects, is therefore estimated by regressing phenotypic resemblance on the expected level of allele sharing, or equivalently the degree of relatedness.

More recently, high-density genotyping has made it possible to estimate genetic relationships from genome-wide genotypic data, based on the proportion of alleles shared IBS[146]. One advantage of this approach is that genetic similarity can be inferred when pedigree data are not available, incomplete, or contain errors. The other advantage is that genome-wide genotypic data can provide better estimates of the level of genetic similarity between relatives than known relationships: genetic similarity between pairs of individuals with a given degree of relatedness fluctuates around its mean (except for monozygotic twins and parent-offspring pairs)[146]. By providing better estimates of genetic similarity between relatives, genotypic data allow better estimates of heritability, as proven by selection experiments in cattle[147].

4.2 Methods

4.2.1 Phenotyping pipeline

Each batch of rats was phenotyped over twelve weeks (Table 4.1), according to the protocol published by Johannesson *et al.*[79], which followed Solberg *et al.*[148].

Phenotype	Disease model	Number of measures	Week
Coat colour		4	7
Wound healing		1	7 and 17
Fear related behaviours	Anxiety	10	8 to 10
Glucose tolerance	Type II diabetes	6	11
Cardiovascular function	Hypertension	2	12
Body weight	Obesity	1	13
Basal hematology		26	13
Basal immunology		34	13
Induced neuroinflammation	Multiple sclerosis	11	13 to 17
Bone mass and strength	Osteoporosis	43	17
Arterial elastic lamina ruptures		6	17
Serum biochemistry		15	17
Renal agenesis		1	17

Table 4.1 Summary of phenotypes collected.

The rats were 7 weeks of age when the phenotyping began. Their **coat colour** was recorded. One of their ears was punched, making a 2-mm hole in the centre of the cartilaginous part of one ear. At dissection ten weeks later (week 17), the ear was placed in formalin so that the size of the partially healed hole could be later measured. The **healing rate** was calculated.

Next, measures relevant to **anxiety** were collected in three behavioural tests. First, the animals were placed in an elevated annular platform ("elevated zero maze") that comprises two sections closed by walls on both sides, and two open sections. The open sections are anxiogenic to rats so that anxious rats tend to stay in the closed sections. Measures indicating the propensity of rats to enter the open sections were collected. Second, the rats were placed in a cage similar to their home cage. Strong illumination contributed to making this experience frightening. Their activity during the first five minutes was recorded as a measure of anxiety (anxious rats tend to explore less in the first five minutes) while their activity twenty-five minutes later served as a basal measure of locomotor activity. Third, the rats were subjected to fear conditioning: in a box with two compartments, a conditioned stimulus (sound and light) was paired with an unconditioned stimulus (foot shock). Crossing of the animal from one compartment to the other interrupted/prevented foot shock administration. Rats are expected to learn in this paradigm that crossing to the other compartment prevents shock administration. Anxious rats tend not to learn as well as less anxious rats. Measures relevant to the crossings were collected, as well as the time each rat spent freezing.

At week 11, glucose was injected intraperitoneally in the rats in post absorptive state, and glycemia measured 0, 30, 60, and 120 minutes after injection. Elevated fasting glucose and/or elevated glucose tolerance (slow clearance of glucose from the blood after glucose administration) are indicative of a pre-diabetic state in humans. This test was therefore used here to investigate the genetic basis of **glucose homeostasis** with applications to type 2 diabetes.

At week 12, blood pressure was measured using a tail cuff. This measure investigates the **cardiovascular function** and was completed by weighing the hearts of the rats after they were sacrificed by exsanguination (week 17).

At week 13 a blood sample was collected. It was aliquoted to carry out two analyses: **complete blood count** (i.e. quantification and various measures for red blood cells, leucocytes, and platelets), and detailed quantification of the different populations of lymphocytes (using fluorescent assay cell sorting). Blood was collected prior to any infection and is therefore relevant to **basal immunology**.

Immediately after blood sample collection, the rats were immunized with myelin oligodendrocyte protein emulsified in adjuvant, to induce **experimental autoimmune encephalomyelitis** (EAE). EAE is a highly reproducible model of multiple sclerosis with a robust clinical score scale. The scale is as follows: 0 = healthy; 1 = tail weakness or tail paralysis; 2 = hind leg paresis or hemiparesis; 3 = hind leg paralysis or hemiparalysis; 4 = tetraplegy, urinary, and/or fecal incontinence; and 5 = death. The rats were scored daily during 28 days after immunization, at which point they were sacrificed.

At sacrifice (week 17), a number of **organs were dissected** and stored appropriately: heart, portion of the abdominal aorta and of the left common iliac artery, thymus, punched ear, brain, portion of the spinal cord, liver, spleen, kidneys, adrenals, hind limbs, serum, tail). Any observation of renal agenesis (lack of a kidney) was recorded.

Lesions in the abdominal aorta and the left common iliac artery were counted following staining. Such lesions form a potential model for aneurysms.

The density, structure, and strength of the hind limbs were measured, in an attempt to understand the genetic basis of bone pathologies including osteoporosis.

Serum concentrations in lipids, ions, proteins and other small molecules were obtained using an automated analyser. Lipids and ions concentrations (especially sodium and chloride concentrations) are relevant to the study of the cardiovascular function and metabolism.

The rats were **weighed** five times (weeks 8, 9, 13, 14, 17).

In total, more than 200 measures were collected in 2006 rats. 160 traits were mapped.

4.2.2 Data collection and quality control

Most of the phenotypes and covariates were uploaded to a database (Integrated Genotyping System[149]) as they were collected. Quality checks for typographical errors were performed by inspecting the distribution of phenotypic values for each measure. Data censoring (e.g. hemalysis was scored but sensored at 6), use of 0s for missing data, and other common errors were also checked. Time series for body weight and EAE scores were plotted for each rat, and organised by similarity of profile so that data points falling out of pattern could be quickly seen by inspection. Further checks, such as comparing body weight or iron levels and recorded sex, were performed to try and detect sample mix during phenotyping. Sample mix-up affecting the genotype data (i.e. where DNA labels were incorrect) was investigated by comparing genome-wide genotypic similarity with the known pedigree, but was not very conclusive because all the HS rats are closely related. Recorded sex and genotypic sex were compared.

4.2.3 Identification of experimental covariates

For each trait, a set of relevant covariates was identified. The significance and effect size of each were estimated using either a linear or a generalized linear model, depending on the trait distribution. The effect of sex was evaluated first, then the effect of batch (generation) conditional on sex, then the effect of individual covariates conditional on sex and batch. Whenever the covariates were partly confounded with sibship, their significance and effect size were estimated in a mixed model where sibship was fitted as a random effect. Those covariates that were highly significant ($p\text{-value} < 0.05$) and explained more than 3% of the phenotypic variance were retained and included as fixed effects in the models used to test for association. However, covariates with too many missing values and those that were categorical with many categories (e.g. day of the experiment) were not fitted. The former would have resulted in reducing the sample size and losing power, and the latter would have needed to be fitted as multiple random effects but the implementation of mixed models used here only allowed for one random term (kinship, to account for relatedness).

4.2.4 Data transformation

Phenotypes whose distribution was approximately normal (i.e. unimodal) were normalized conditionally on the important covariates using a Box-Cox transformation (available in the R package MASS[150]). The phenotypes that were far from normally distributed (e.g. with a negative binomial distribution) were not transformed.

4.2.5 Phenotype correlations

All pairwise phenotypic correlations were computed on the residuals of (generalized) linear models in which the important covariates were fitted. Pearson and Spearman correlations were calculated.

4.2.6 Heritability estimation

Pairwise genome-wide genetic similarity was estimated from the genome-wide genotypic data using IBS and the R package EMMA[140] (function emma.kinship). Heritability was estimated as $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$ in the following mixed linear model:

$$y = X\beta + \upsilon + \varepsilon \qquad \text{(model 0)}$$

where y is the vector of phenotypic values, β the vector of regression coefficients and X the design matrix of the covariates. The covariates include a dummy intercept term. υ and ε are random effects, with covariance matrices $\sigma_g^2 K$ and $\sigma_e^2 I$ respectively, where σ_g^2 and σ_e^2 are the variance components of the model, and quantify the contribution of the genetic background (polygenic effects) and the residual noise respectively. They were estimated using the R package EMMA (function emma.REMLE)[140]. K is the genetic relationship matrix (the IBS matrix above), and I the identity matrix.

The variance components estimation carried out by EMMA assumes the trait is normally distributed (conditionally on the fixed effects, i.e. covariates). Therefore, estimates of heritability were obtained only for those measures that were approximately normally distributed.

4.3 Results

4.3.1 Important covariates, data distribution and transformation

APPENDIX B shows the covariates that were included when testing for genetic association, whether or not their distribution was normalized, and the number of non-missing data points per measure. Albino coat colour is an important covariate for behavioural measures because albino rats are more affected by bright environments than non albino rats, which contributes to their increased anxiety in the behavioural tests where there is a bright light (Albert Fernandez Teruel, personal communication).

Many covariates had significant and large effects when tested in a model including sex only. However, only those shown in APPENDIX B were included in the models used to test for association, because the others (cage, litter, day the experiment was carried out, etc.) were not expected to have large effects and were highly confounded with the genetic effects. For example, cage is a covariate that captures the microenvironment in which the rats lived, which could have significant effects on some phenotypes. However, since the conditions of light, temperature, food, water, and bedding were controlled in the animal house, we had no reason to believe cage should have an effect. Cage is a categorical variable with many levels (839 different cages) so that its effect is best accounted for by fitting a random term in a mixed model.

Because siblings were housed in the same cage (possibly in more than one cage if males and females were present, and when the sibship was large), cage effects and random genetic effects due to all the genes contributing to a phenotype are highly confounded. Therefore, since the models used to test for association included a term for random genetic effects, and because cage was not expected to have a large effect, cage was not included as a covariate. Day the experiment was carried out and experimenter were also confounded with the familial relationships, and since batch captured the effects of these covariates, they were not included. The rat study was not amenable to investigating gene by environment interactions because genetic effects and covariates were all confounded due to the production of rats over eight generations. This contrasts with the mouse HS study where the same forty breeding pairs were used to produce multiple litters, so that sibship (genetic effects) was less confounded with the covariates[2,148].

4.3.2 Phenotype correlations

Measures related to the same phenotype (i.e. novel cage activity, aortic lesions, etc.) were often highly correlated, as expected. No pair of measures related to different phenotypes had an absolute correlation greater than 0.2.

4.3.3 Heritability

Heritability was estimated based on genome-wide genotypic similarity in a mixed model[140], and is reported in APPENDIX B for those measures that were approximately normally distributed. Heritability ranges between 4 and 74% for these measures. Figure 4-1 shows that behavioural measures have the lowest heritabilities while measures of basal immunology and hematology, measures relevant to bone morphology, and body weight have the highest.

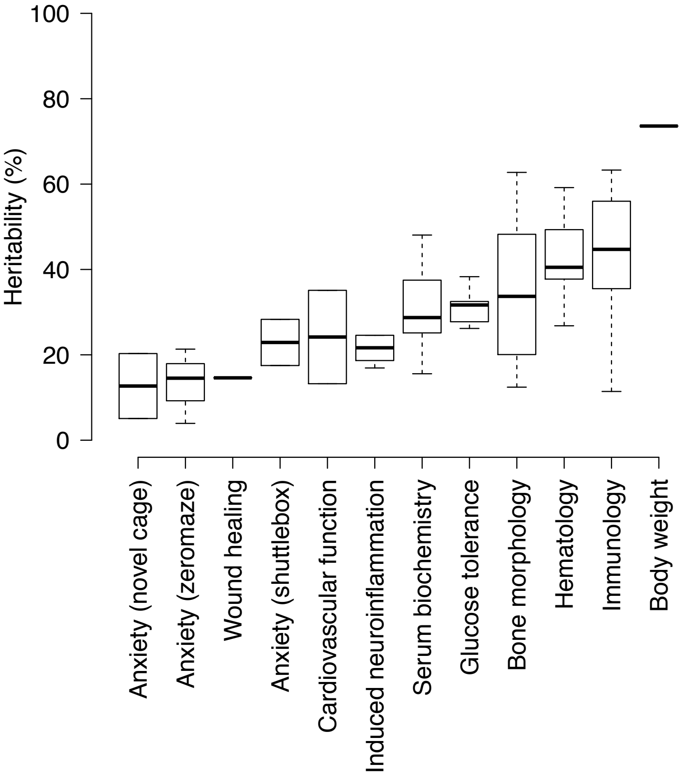


Figure 4-1 Heritability of the measures collected in the rat HS. The measures are grouped by phenotype.

Heritability estimates for 32 measures collected both in the mouse and in the rat HS were compared, and showed a high and significant correlation (Pearson's correlation 0.60, p-value 0.0002; Figure 4-2), with the greatest heritabilities being the most conserved.

For those measures in the mouse HS that were not exactly mirrored in the rat HS, measures of basal immunology and body weight have the highest heritabilities in the mouse HS as they do in the rat HS, and behavioural measures also tend to be the least heritable.

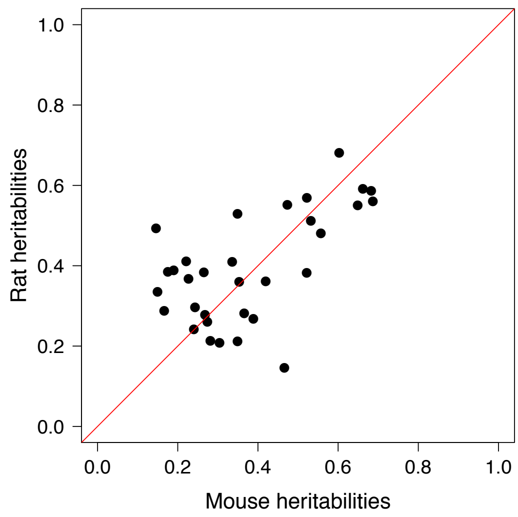


Figure 4-2 Comparison of heritability in the rat and mouse HS. Heritability was calculated for the same 32 measures in both HS.

4.4 Discussion

In this chapter I presented the measures that were collected on the HS rats by the EURATRANS consortium. I justified their transformation for QTL mapping and the selection of a subset of covariates that will be taken into account in subsequent analyses. I estimated heritability for each of them. This analysis revealed that heritability of behavioural measures was on average lower that that of physiological and morphological measures, and that heritabilities of homologous phenotypes in the rat and mouse HS are correlated.

Heritability is specific to a population because it depends on the genetic variation that segregates in that population, and the experimental/environmental variation to which the population is faced. However, consistent patterns of heritability estimates have been observed in many studies of natural outbred populations in various species, and are seen in this study as well. Namely, heritability tends to be lower in behavioural measures than in physiological and morphological traits[49,151].

The analyses presented here show that heritability in HS rats is consistent with that seen elsewhere. It is therefore likely that many genetic loci will be associated with the traits measures. In the next chapter, I present the results of QTL mapping, the first step towards identifying the genetic variants underlying heritability.

Chapter 5

QTL mapping

This chapter maps the genetic basis of 160 traits in 1,407 HS rats. I start by explaining why genetic mapping in an HS is challenging. I then describe the methods that are used to overcome those challenges. I present the results from genetic mapping, the validation of some of my findings, and describe relevant features of the genetic architecture of the HS.

5.1 Background

5.1.1 Population structure in the HS

The major difficulty for genetic mapping in an HS is that the stock consists of individuals with different degrees of relatedness. In Chapter 3 I described the pedigree of the HS rats: siblings, half sibs, cousins, uncles, great-uncles, etc. were included in the mapping population. These generate structure in the mapping population, as can readily be seen from the IBS matrix (Figure 3-7). Structure translates into linkage disequilibrium between distant markers, or long-range genotypic correlations. The limited population size during breeding of the stock can also create correlations between physically distant loci[142].

5.1.2 Genetic mapping in the presence of population structure

Association mapping methodologies often assume that all individuals are unrelated (as is the case in most human genetic studies) which is the same as assuming they are equally related (as happens in the analysis of classical inbred strain crosses, F2 and back cross designs). In these designs, there are no long-range genotypic correlations. In the HS, correlations at distant loci do exist and can cause spurious associations because any locus correlated with a causal locus will tend to be associated with the phenotype as well. This is the problem that genetic mapping in an HS has to address.

A number of different methods have been used to control for structure. For family based studies, transmission disequilibrium tests (TDTs)[152,153] can be used. The TDT is robust to population structure because they require linkage. TDTs require markers that unequivocally determine the parental origin, such as multiallelic loci. Such markers are rare in the genome so linkage with the causal variants may be imperfect. Moreover because TDTs needs genetic information from multiple family members, genotyping costs are high. In the case of the rat HS, the varying degrees of relatedness rule out the use of the TDT.

Alternative approaches have been developed that require neither family information nor highly polymorphic markers. Genomic control uses unlinked markers to estimate the level of inflation in the test statistics due to population structure and corrects the statistics accordingly. STRAT[154] proceeds by identifying subpopulations in the mapping population, and testing for association within subpopulations. More elegantly, the principal components of genotypic variation are fitted in the model used to test for association[155].

More recently, mixed models have become the tool of choice to account for population structure in association studies[93,106,140,156-158]. Mixed model methods account for differences in pairwise genome-wide genetic similarity in the mapping population, so that genetic loci are required to explain phenotypic variation independently of background polygenic effects to be deemed associated with the phenotype. In addition to these single-point methods (which test for the effect of one SNP or haplotype at a time), methods exist that allow SNPs to compete in multilocus models, so that long-range correlations do not give rise to spurious associations[142,159]. Such methods use various techniques to select the loci to include in the models, most often forward and backward selection, and are sometimes coupled with resampling for robustness.

The process by which spurious associations arise depends on the genetic architecture of the trait, on whether is it controlled by a few major loci or highly polygenic. At one extreme a spurious association could arise from correlation with a single causal locus of large effect, at the other a large number of correlated loci each of small individual effect could inflate each other's association.

Acknowledging this, I considered two methods to control for spurious associations in this study: mixed models, which is expected to work best for highly polygenic normally distributed traits[160], and a multilocus method developed by Valdar *et al.*[142], expected to work well on spurious associations arising from correlations with larger effect loci.

I compared both methods in terms of power to detect true associations and rate of false positive associations. This comparison was carried out on phenotypes simulated with genetic architecture resembling the phenotypes measured in the rat HS, and is presented in the first section of this chapter, before summarizing the results of genetic mapping in the HS.

5.1.3 Evaluation of the success of an association study

In this section I review the criteria that define a successful association study, where success measures the capacity the study has to offer clear leads on the genetic variants that cause phenotypic variation. I will later report the results of my study with these criteria in mind.

The number of QTLs identified at an acceptable significance threshold is a measure of the success of an association study, because each QTL offers a lead on the genetic basis of a complex trait. The large number of loci detected by human GWAS is often cited to support their success[161,162]. While the genome-wide significance level used in published human GWAS is almost invariably 5.10^{-8} (following a proposition by Risch and Merikangas[163] and equivalent to a p-value of 0.05 after a Bonferroni correction for 1 million independent tests), it varies in studies of laboratory populations where different breeding designs can be used and linkage disequilibrium extends further, and there is greater variability in the method used to calculate significance thresholds[164,165].

In addition to a strict control of the number of false positive associations reported by each study, replication in a second population is usually required in human association studies[166,167] to overcome the problem of false positive results arising from underpowered studies[168]. Replication has received less attention in studies of laboratory populations, although equally necessary[169,170], likely as a result of different investigators using different mapping populations. We attempted to find evidence of replication for the QTLs we identified in this study by comparing our results to those reported in the largest QTL database for the rat (the Rat Genome Database[171]) and by genotyping the QTLs in an independent sample of 752 HS animals.

QTLs of large effect size are preferred because they are more likely to be genuine and more tightly mapped, and because it makes the QTL more tractable in subsequent functional studies[83].

Finally, the precision of the mapping, i.e. the width of the confidence interval of each QTL, should be as small as possible to be able to identify the genetic variants or genes underlying each association[83,172]. Large QTLs indeed lead to intractably long lists of candidate variants and genes, and leave open the possibility that multiple linked causal variants in different genes in combination generate the detected QTL[172].

5.2 Methods

5.2.1 Mixed models

Mixed models use genome-wide genotypic similarity between individuals to model phenotypic correlations. QTLs are detected if they contribute to phenotypic variation independently of genetic background effects caused by many loci of small effects (polygenic effects). The three steps in mixed model association are the construction of the genetic relationship matrix, the estimation of the effect sizes of the polygenic effects and residual noise, and the testing of genetic loci for association with the phenotype. The last two steps can be decoupled, as first suggested by Kang *et al.*[157] and as performed in this study. I will discuss the assumptions underlying this strategy at this end of this section.

The IBS genetic relationship matrix is used here, and was constructed using the R package EMMA[140] (Chapter 3).

The contributions of polygenic effects and the residual noise to phenotypic variation were estimated under the null model (model 0, presented in matrix form in Chapter 4):

$y_i = \Sigma_c x_{i,c}\beta_c + \upsilon_i + \varepsilon_i$ (model 0)

The models used to test for association between the ancestral haplotypes segregating at a locus L and phenotypic variation were:

$y_i = \Sigma_c x_{i,c}\beta_c + \Sigma_L [\Sigma_s P_{L,i}(s)T_L(s)] + \upsilon_i + \varepsilon_i$ (model 1)

where $P_{L,i}(s)$ is the dosage of the founder haplotype s present at the locus L (defined as the interval between two successive genotyped markers) of animal i, as defined in the introduction of Chapter 3. $T_L(s)$ is the deviation in phenotypic value that results from carrying one copy of a haplotype from strain s at locus L. The variance components σ_g^2 and σ_e^2 are the estimates obtained in the null model (model 0). In matrix form, model 1 is:

$y = X\beta + P_LT_L + \upsilon + \varepsilon$ (model 1)

The presence of a QTL at L is tested by estimating the $T_L(s)$ and testing whether they are significantly different from 0. We pre-multiply the models by A^{-1}, where A is a matrix square root of the variance matrix $V=A^2$:

$(A^{-1}y) = (A^{-1}X)\beta + A^{-1}(\upsilon+\varepsilon)$ (model 2)

$(A^{-1}y) = (A^{-1}X)\beta + (A^{-1}P_L)T_L + A^{-1}(\upsilon+\varepsilon)$ (model 3)

so that the variance-covariance structure of the random term $A^{-1}(u + \varepsilon)$ is now proportional to a diagonal matrix and so can be fitted as a standard linear model. This follows from the eigenvalue decomposition of the covariance matrix V as EDE' where E is the matrix of eigenvectors and D is the diagonal matrix of eigenvalues. Because V is orthogonal, $E' = E^{-1}$. It follows that:

$$var(A^{-1}y) = A^{-1}\, var(y)\, A^{-1\,\prime} = A^{-1}\, V A^{-1\,\prime}$$

$$= (ED^{1/2}E')^{-1}\, EDE'\, (ED^{1/2}E')^{-1} = E'^{-1}D^{-1/2}E^{-1}EDE'E'^{-1}D^{-1/2}E^{-1} = E'^{-1}E^{-1} = I.$$

The transformed models (pre-multiplied by A^{-1}) were fitted using the R function lm, for each of the (24,196) loci, and analysis of variance (ANOVA) was used to estimate significance (p-value) and effect size (the ratio between the fitted sum of squares and the total sum of squares) of each locus.

The decoupling of estimating the variance components from association testing is carried out by estimating the variance components once only under the null model, and using them to calculate V and transforming the models used to test for association. It assumes that the variance components do not change much when local genetic variation is fitted in the model. Kang *et al.*[157] showed this was likely to be true when the locus explains a small fraction of phenotypic variation.

5.2.2 Multilocus resample model averaging

The resampling method used here is that developed for the analysis of the mouse HS and implemented in BAGPHENOTYPE[142,148]. BAGPHENOTYPE fits multilocus models, which model phenotypic variation as a function of variation at multiple genetic loci simultaneously (indexed in the equations below by the variable L). The loci to be included in the model are chosen using analysis of variance and forward selection. Analysis of variance compares the following two models:

$$y_i = \Sigma_c\, x_{i,c}\beta_c + \varepsilon_i \qquad \text{(null model)}$$

$$y_i = \Sigma_c\, x_{i,c}\beta_c + \Sigma_L\, [\Sigma_s\, P_{L,i}(s)T_L(s)\,] + \varepsilon_i \qquad \text{(alternative model)}$$

where y_i , β_c, x_{ic}, $T_L(s)$, and $P_{L,i}(s)$ are defined as above. ε_i is a residual for animal i, the residuals for different animals being assumed uncorrelated.

In forward selection, the locus that explains phenotypic variation best is included in the model in the first step, the remaining loci are tested again in models that include that best locus and the best one is included in the model in the second step and so on. A causal locus with greatest effect is likely to be included first, so that loci correlated with it because of population structure but with no effect independently of the causal locus will not be subsequently selected in the model. However, forward selection is prone to overfitting the data, so resampling is used to increase the stability of the models: a large number of subsamples of the mapping population are drawn and a multilocus model fitted with each, and the number of times a genetic locus is selected is recorded to give a "posterior inclusion probability" that is a compound measure of the strength of an association and how robust the association is (i.e. whether the locus would still be associated if the mapping population was slightly different).

Generalized linear models can be fitted in BAGPHENOTYPE when the distribution of the phenotype is not Gaussian and is closer to one of a subset of classical distributions (e.g. binary and negative binomial distributions). However, fitting can be slow and can fail for generalized linear models due to numerical stability issues. In this study, a locally adapted version of BAGPHENOTYPE was used, as the software is no longer maintained by its developer (see http://valdarlab.unc.edu/software.html for newer, related methods).

5.2.3 Comparison of multilocus and mixed models by simulation

I compared the performance of mixed models and BAGPHENOTYPE to determine which had more power to detect true associations while controlling the rate of false positive associations. To do so, it is most common to proceed in two steps: first, simulate a large number of phenotypes without any true association (null simulations), and, for each method, determine the threshold at which on average x% of genome scans show at least one association (i.e. a genome-wide significance level of $x/100$); second, simulate a large number of phenotypes with one true QTL, analyse with each method and determine the number of simulations where the true association is detected at the threshold previously determined (statistical power of the method). However, Table 2 in Valdar *et al.*[148] reported that BAGPHENOTYPE performs differently on null simulations and simulations with a number of true associations: the false discovery rate was higher in the latter case. Therefore, the number of false positives cannot be controlled by calculating a threshold using null simulations, and I compared the two methods based on their false discovery rate determined in simulations with multiple true QTLs.

Simulation of normally distributed phenotypes

I simulated complex traits arising from a combination of many loci of small effect (a polygenic model), in combination with a few loci with larger effects. The justification for a polygenic effect stems from the observation that most traits in all populations studied are heritable to some extent, but the combined effects of all the associations detected for each usually explain only a small fraction of the heritable phenotypic variance. This fraction ranges from 10% to 90% in the mouse HS[148] and is much smaller in humans[48]. Therefore, I simulated all phenotypes with a polygenic background in addition to true QTLs. Based on the number and effect sizes of the associations detected in the mouse HS[148], I simulated seven true QTLs, each explaining 5% of the phenotypic variance. In conclusion, I simulated 1,000 normally distributed traits, under the following model:

$$y_{sim} = \Sigma_{(L1,\ldots,L7)}\, a_L P_L T_L + \upsilon + \varepsilon \qquad \text{(model 4)}$$

Each phenotype arises from a combination of three effects: (i) a polygenic effect υ simulated by drawing from a multivariate normal distribution with mean 0, covariance matrix the pairwise genome-wide average IBS matrix scaled so that the polygenic component explained 20% of the phenotypic variation; (ii) seven effects representing true QTLs with seven loci chosen at random. Haplotypic effects rather than genotypic effects were simulated, using the HAPPY probability matrices P_L, with a scaling factor a_L chosen so that each locus explained 5% of the phenotypic variance; (iii) uncorrelated errors explaining the remaining 45% of phenotypic variation. Because the polygenic term and the QTLs are correlated, the final effect size of each component could differ substantially from the target.

Simulation of non-normally distributed phenotypes

Not all the phenotypes measured in the rat HS were normally distributed, so I also compared the methods' performance on non-normal simulations. I simulated one set of 50 non-normally distributed phenotypes for each of the 19 non-normally distributed traits collected in the HS. To do so, I first simulated normally distributed phenotypes under the following model:

$$y_{sim} = X\beta + \Sigma_{(L1,\dots,L7)}\, a_L P_L T_L + \upsilon + \varepsilon \qquad \text{(model 5)}$$

which differs form model 4 by including covariates' effects. Then, I transformed y_{sim} by quantile normalization so that its distribution matched the distribution of the HS phenotype.

In a preliminary set of null simulations where the polygenic effect υ explained 20% of the phenotypic variance in the normally distributed simulated phenotypes (model 5), I noticed that transforming y_{sim} to change its distribution led to the polygenic effect being badly estimated (the polygenic effect estimated from y_{sim} was on average lower than 0.2 in this preliminary set). Therefore, for all my simulations of non-normally distributed traits, the effect size of the polygene υ in the normally distributed simulations (model 5) was adjusted so that the effect size of υ estimated from the transformed simulations matched that from the real phenotype. I also matched covariates' effects and major QTLs (considered as covariates).

Evaluation of the performance of the methods

The 1,000 normally distributed simulations and the 19 sets of 50 non-normally distributed simulations were mapped by both mixed models and BAGPHENOTYPE. A QTL is here defined as the 4Mb window centred on the marker interval with the local maximum negative logarithm 10 of the p value ("negative logarithm 10 of the p value" hereafter "logP") / inclusion probability. A true association is defined as a QTL overlapping a simulated association.

For each simulation the detected QTLs were ranked in decreasing order of logP or inclusion probability, and the number of false associations that ranked above the k^{th} strongest true association was recorded, with k taking values in [1,…,7]. The simulations were pooled to estimate the smallest number of false positive associations found when i true associations have been detected. Since some of the simulated QTLs were completely missed by BAGPHENOTYPE, I defined detectable simulated QTLs as those with a non zero inclusion probability, and compare the false discovery rate (FDR) of both methods at different proportions of detectable QTLs detected.

5.2.4 Calculation of thresholds to call QTLs at a FDR of 10%

Inclusion probability thresholds for resample model averaging

I determined thresholds corresponding to a 10% FDR (ratio between number of spurious associations and total number of associations), from simulations for each of the 19 non-normally distributed phenotypes. I called QTLs using inclusion probability thresholds successively between 0.05 and 1 (in steps of 0.05), and calculated the FDR at each threshold. The inclusion probability threshold that led to the FDR closest to 10% was used to call QTLs for the real phenotype.

Significance thresholds for mixed models

I determined a logP threshold corresponding to a FDR of 10% for each of the normally distributed phenotypes. To do so, for each normalized phenotype, 1,000 phenotypes were simulated using the observed effects of covariates significant for the trait, correlated genetic random effects and uncorrelated random errors with effects matching those for the normalized phenotype. I mapped these null simulations (no QTLs simulated) using mixed models and obtained the distribution of the genome-wide maximum negative log p-value of association under the null hypothesis of no association. I fitted an extreme value distribution (EVD) to these maxima (fitting an EVD produces more accurate estimates of tail probabilities)[2].

I estimated the logP threshold necessary to achieve a FDR of 10% across each normally distributed trait by applying the following procedure for every significance threshold between the 5^{th} and 95^{th} percentile of the EVD: (i) call QTLs in the simulated data exceeding the specified threshold, (ii) calculate FDR as the ratio of the number of false positive associations to the number of detected QTLs, where the number of false positive associations is determined by the significance threshold (e.g. if the significance threshold is the x^{th} percentile of the extreme value distribution, there will be a false positive association in (100-x)% of null genome scans, or (100-x)/100 false associations per scan). I found that using the x=65^{th} percentile of the extreme value distribution for each phenotype ensured a FDR of 10% across all phenotypes.

5.2.5 Calculation of confidence intervals

Confidence intervals for QTLs mapped using mixed models were calculated by simulating a large number of phenotypes each arising from a single QTL in addition to correlated genetic random effects and uncorrelated errors. The QTLs were simulated over a range of effect sizes, were then binned according to the logP of the detected QTLs, to obtain confidence intervals for different logP bins[173]. To do so, for each simulation within a bin, the distance between the simulated QTL and the highest interval in the 20Mb window around the QTL was recorded. The distribution of distances obtained this way was used to calculate the 90% confidence interval of the QTL. For those phenotypes mapped using resample model averaging, we report QTLs as 4Mb windows centred on intervals with inclusion probabilities greater than that required to achieve a FDR of 10%.

5.2.6 Calculation of QTL effect sizes and comparison with heritability

QTL effect sizes were defined as the ratio between the fitted sum of squares and the total sum of squares after removing covariates in a linear model (no random genetic component). This is an upper bound on the true QTL effect size, as it attributes all possible variation to the QTL, but note that including the random genetic component would result in underestimation of most of the effect sizes because part of the variance would have been attributed to it.

Joint effect sizes were defined as the ratio between the fitted sum of squares and the total sum of squares in a model without a random genetic component, including covariates and all the QTLs called for a given phenotype. This is an upper bound. Thus, the QTL effect sizes reported are probably overestimates.

Heritability was defined as the ratio of the genetic variance component to the sum of the variance components estimated in the null mixed model (covariates but no QTL, see Chapter 4). It was calculated for those traits that were normally distributed. The proportion of heritable variance explained by the QTLs was defined as the ratio between the joint effect size and heritability of the trait.

5.2.7 Replication of behavioural QTLs in an independent set of HS rats

I attempted to replicate 28 behavioural QTLs using an independent sample of 768 HS rats of the generations before those used for the main study, which were phenotyped in the same laboratory and using the same protocols[79]. The 28 QTLs were chosen based on the logP of association with genotypic variation at the genotyped markers, obtained by a mixed model. To avoid trying to replicate associations due to genotyping errors, only those QTLs with more than one marker with a high logP were chosen. 55 markers were selected for genotyping (on average 2 per QTL). Genotyping was done using the Sequenom MassArray platform according to the manufacturer's instructions (www.sequenom.com).

Association in the replication sample was tested using fixed effects linear models because genetic relationship between the 768 rats used for the replication experiment cannot be reliably estimated from only 55 genotyped markers.

In order to be able to compare significance of the loci in the HS and independent sample, association between the 55 markers and the 1,407 HS rats was also calculated using fixed effects linear models.

5.2.8 Overlap between rat HS QTLs and QTLs catalogued in RGD

The QTLs catalogued in the Rat Genome Database (RGD[171]) were retrieved on August 9th, 2012. For each measure mapped in the HS, I manually identified all the RGD QTLs associated with the exact same measure (although the names could be slightly different). For the analysis, I only considered RGD QTLs smaller than 50Mb, as defined by the start and stop coordinates in the Rnor3.4 rat genome assembly. For each measure, I calculated the number of protein coding and miRNA genes overlapping both the QTLs mapped in the HS and the QTLs catalogued in RGD. To calculate the significance of this overlap, I sampled 1,000 sets of genomic intervals at random so that each set had as many intervals as there are RGD QTLs associated with the measure, and each interval had as many genes as the corresponding RGD QTL. I then calculated the number of genes overlapping both an interval in the random set and an HS QTL. I thereby obtained a distribution of the number of overlapping genes under the null hypothesis of no shared genetic basis between RGD and HS QTLs. The P-value of the overlap between HS and RGD QTLs for a given phenotype was obtained by comparing the number of genes overlapping both RGD and HS QTLs to this null distribution.

5.3 Results

5.3.1 Comparison of mixed models and BAGPHENOTYPE

QTL mapping results for all normally distributed phenotypes and for two examples of non-normally distributed phenotypes are presented in Figure 5-1. The mixed models performed slightly better (smaller FDR) than resampling for normally distributed phenotypes, while the latter method performed always at least as well and often better than mixed models for the 19 non normally distributed phenotypes.

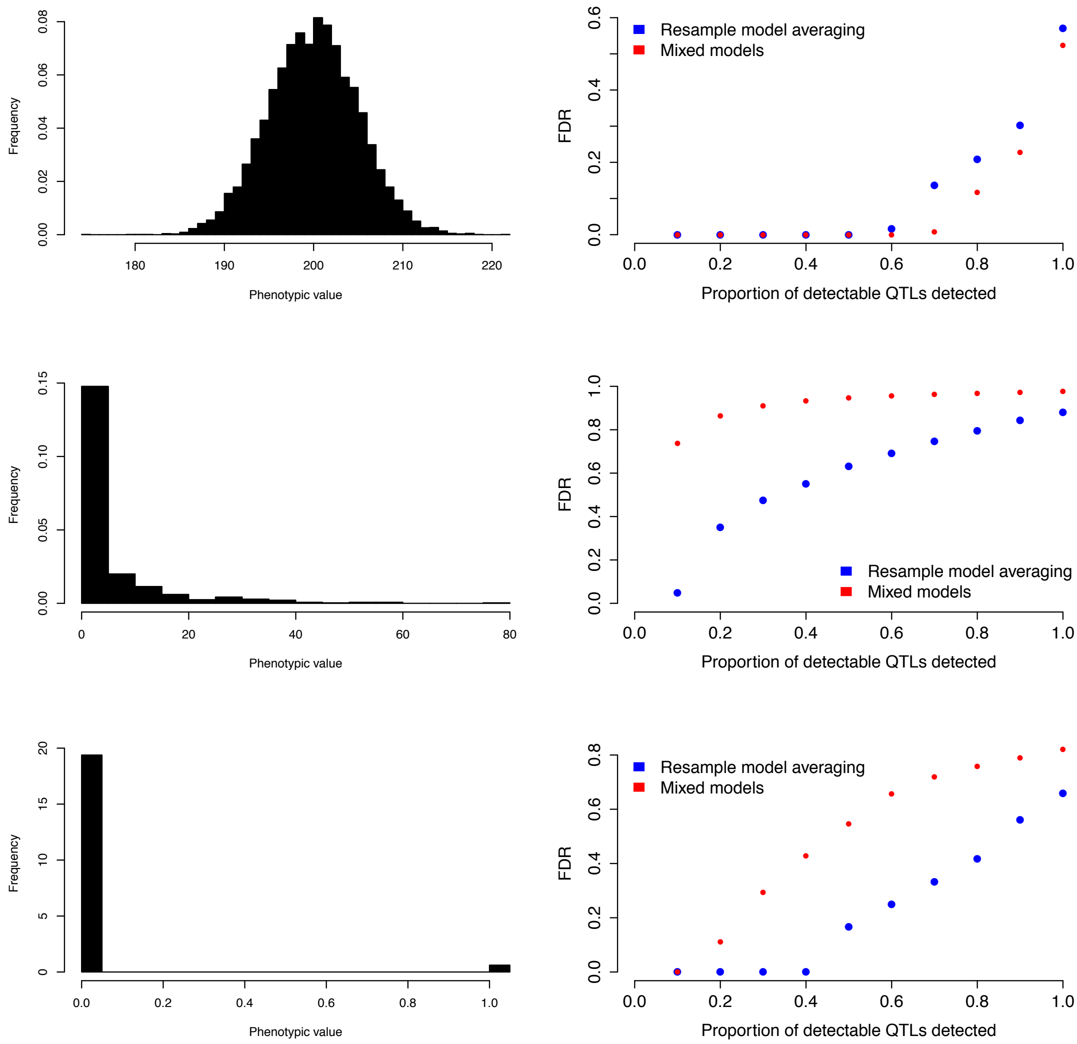


Figure 5-1 Comparison of the performance of BAGPHENOTYPE and mixed models. Three phenotypic distributions (Gaussian, negative binomial, and binary) were considered, and are shown in the left-hand panels. The performance of the methods was evaluated in terms of their FDR (vertical axes in the right-hand panels).

Because these methods have different advantages, I mapped all traits with both, but only report those QTLs detected at FDR of 10% by the method that performed best for each trait - mixed models if it is normally distributed (which is the case of the majority of the phenotypes), BAGPHENOTYPE otherwise (APPENDIX B). At the end of this chapter, I look at the support from the resampling method for those QTLs identified with mixed models.

5.3.2 Number of QTLs detected

At 10% FDR, 355 QTLs were detected for 122 out of 160 phenotypes (APPENDIX C). The number of QTLs detected per measure ranges between 0 and 10 (mean 2.9) for the 141 normally distributed traits and between 0 and 2 for the 19 non-normally distributed traits (Figure 5-2).

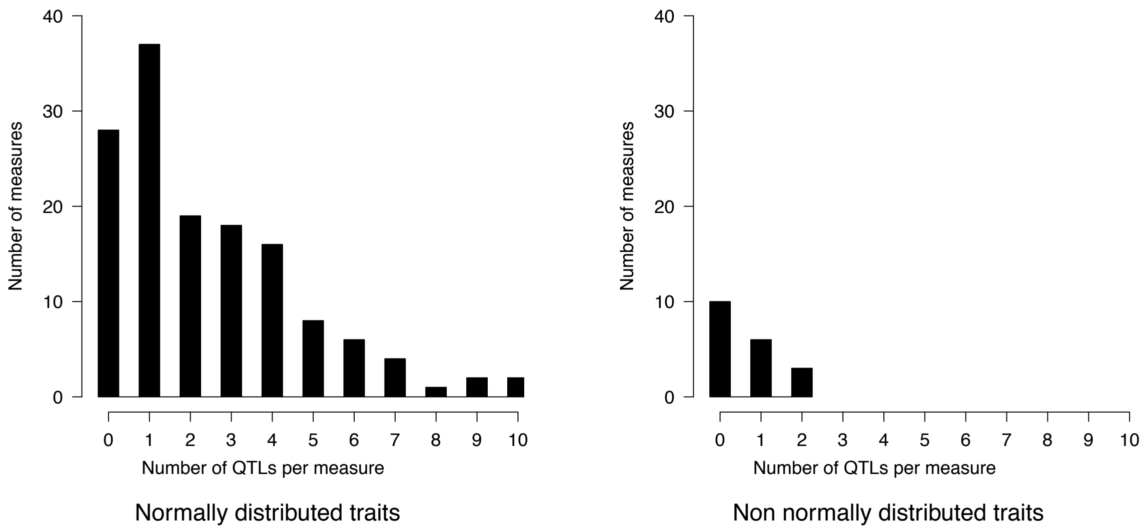


Figure 5-2 Number of QTLs per measure.

Variation in the number of QTLs detected for the normally distributed traits (which constitute the vast majority of the traits studied here) likely reflects a different genetic architecture: the more QTLs were detected, the more genetic loci contributed a sufficiently large amount to phenotypic variation. The total number of loci contributing to phenotypic variation cannot be inferred just from the number of QTLs detected: there could be many loci of too small effect to be detected, or alternatively no other loci than those detected. The next section looks at how much of the genetic variance the detected QTLs explain.

Because of the problems of comparing distributions and the different mapping methodologies used, it may not be meaningful to compare the effect sizes of their QTLs and the proportion of genetic variance they explain, for normally vs non-normally distributed traits. However, it is still striking that many fewer QTLs are detected for the non-normally distributed traits compared to the normally distributed ones. Could it be that non-normally distributed traits have a simpler genetic architecture than normally distributed ones? We know for example that highly heritable traits controlled by single loci have binary distributions (e.g. the simple characteristics of the peas studied by Mendel), and as a consequence of the central limit theorem, the accumulated effects of many loci will tend to be normally distributed.

A simpler genetic architecture certainly accounts for the detection of only a few QTLs for at least 4 out of the 19 non-normally distributed traits: those related to coat colour. The coat colour related phenotypes studied here are indeed known to be mostly under genetic control and determined by one or a few loci, depending on the trait. For example, whether or not the rat is albino (i.e. white with red eyes) is determined by a single gene (*Tyr*, tyrosinase), with no known environmental influence[125]. I mapped one QTL for this trait, and its confidence interval encompasses *Tyr*. I mapped two QTLs for the spotted phenotype (i.e. presence of both white and brown patches). One QTL is the *Tyr* locus, because all the rats with a defective *Tyr* gene will be albino and will not be spotted, while rats with a functional *Tyr* gene may be spotted or not. The other QTL maps to the gene *Kit*, a gene associated with the hooded phenotype[125]. Two shades of brown were recognized: dark (yes/no phenotype) and light (binary as well). Both traits map to the same two QTLs, one of which is the *Tyr* locus (albino rats will never be dark brown while non albino rats may be dark brown or not, and the same applies to light brown). The second QTL is the agouti locus, responsible for the agouti coat colour[159], here called light brown.

The identification of only one or two QTLs per coat colour phenotype reflects a simple genetic architecture of these traits but I was able to say this only based on additional evidence from the literature. In order to investigate the genetic architecture of the traits mapped here and better understand why different number of QTLs were detected for different traits, it is important to consider QTL effect sizes.

5.3.3 QTL effect sizes and proportion of heritable variance explained

Because effect sizes are not strictly comparable between traits with different distributions, even more so when mapped by different methods, I consider separately the QTLs identified for the normally and non-normally distributed traits.

The effect sizes of the QTLs detected for the normally distributed traits have a markedly skewed distribution (Figure 5-3a), with a median effect size of 5% (mean effect size of 6.5%). Large-effect QTLs were rare: only 22 QTLs explained more than 15% of the variance, and none more than 40%.

The distribution of the joint effect size of the QTLs detected for each phenotype is also skewed (Figure 5-3b), with a maximum of 47% of phenotypic variation explained by the QTLs.

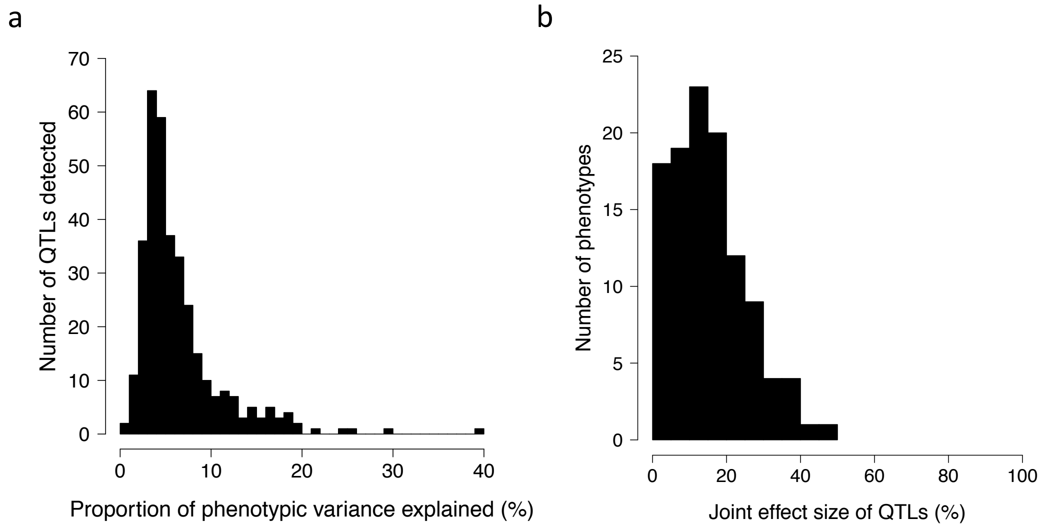


Figure 5-3 Effect sizes of the QTLs mapped for the normally distributed phenotypes. (a) Individual effect sizes. (b) Joint effect sizes.

Figure 5-4 shows the differences that can exist between sum of the effects and joint effect. Specifically, the joint effect size can be much smaller than the sum of the effects. This is due to the genotypes at the QTLs being correlated and the detection method, which is single point so that the loci do not compete to explain phenotypic variation. Correlated QTLs may both be true (causal variants present at each locus), or one may be true and the other spurious. I discuss this issue in more detail later. Because QTLs may be correlated, it is important to use their joint effect when comparing the proportion of phenotypic variance that is explained by the QTLs to the proportion that we know is genetic (heritability).

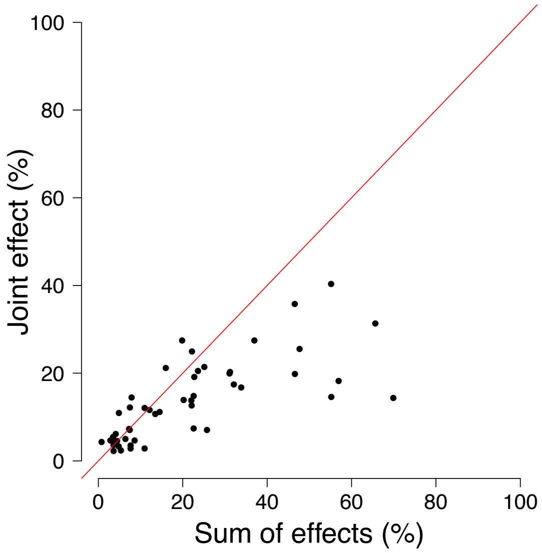


Figure 5-4 Comparison between the sum of the individual effect sizes and the joint effect sizes. Each dot represents a normally distributed phenotype.

This comparison is shown in Figure 5-5. On average, the QTLs jointly explain 42% of the heritable phenotypic variation. This is much higher than the corresponding percentage from human GWAS, where typically the QTLs detected explain less than 10% of the heritable phenotypic variation. The figure also shows that the pattern is similar to the mouse HS.

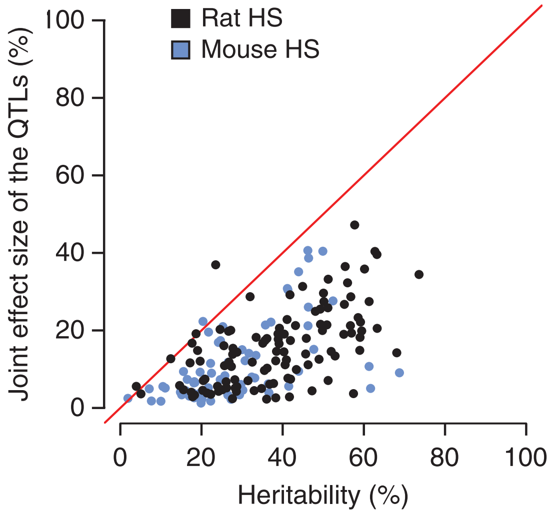


Figure 5-5 Comparison between heritability of the phenotypes and joint effect of the QTLs in rat and mouse HS. Each dot represents a phenotype.

In non-normally distributed traits, major QTLs explaining between 19% and 96% of the phenotypic variance were detected for the coat colour phenotypes as expected (APEENDIX C). In addition, one large QTL explaining close to 10% of the phenotypic variation was identified for aortic lesions, and a very small QTL was identified for a measure of anxiety.

Because heritability is estimated only in the mixed model framework, which assumes normally distributed phenotypes, I did not measure heritability for the non normally distributed traits. Therefore, I was unable to investigate why no QTL, or only one QTL were detected for the non-normally distributed traits that were not related to coat colour. I could not tell whether only a few QTLs explained all of the heritability, or many more QTLs existed, but each with an effect too small to be detected.

5.3.4 Mapping resolution

Figure 5-6 shows the median size of the 90% confidence intervals as a function of logP. The median size of the 90% confidence interval was 4.5 Mb, containing between 2 and 394 genes, and on average more than 40 genes.

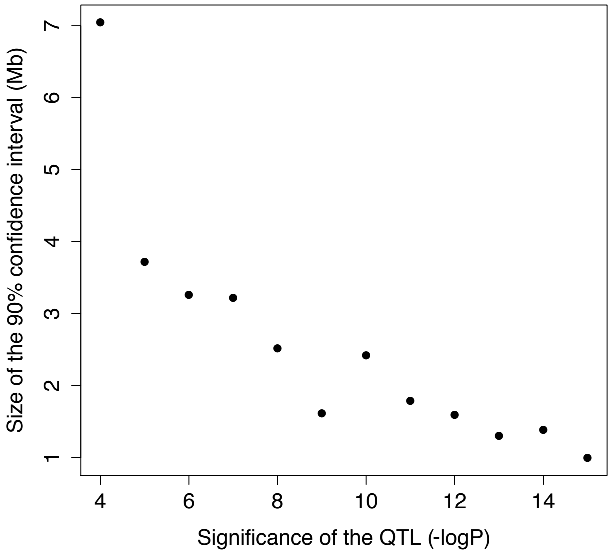


Figure 5-6 Size of the 90% confidence interval as a function of the significance of the QTL.

5.3.5 Replication experiment

46 QTLs detected by mixed models were selected for replication in an independent sample of HS rats. The logP calculated for these QTLs in fixed effects models ranged from 4.2 to 22.2. The corresponding logPs were calculated in the replication sample and ranged from 0.1 to 2.9. To estimate the significance of these values, we calculated the association value between every behavioural measure and every marker genotyped in the replication sample, and compared the values of the pairs that were selected for replication (strong association in the HS) to those for all the other pairs. The results, presented in Figure 5-7, show that the association values for the associations that we were attempting to replicate are not

significantly higher than a set of marker/measure pairs that were not associated in the HS (p-value of one-sided t-test with unequal variance: 0.2234).

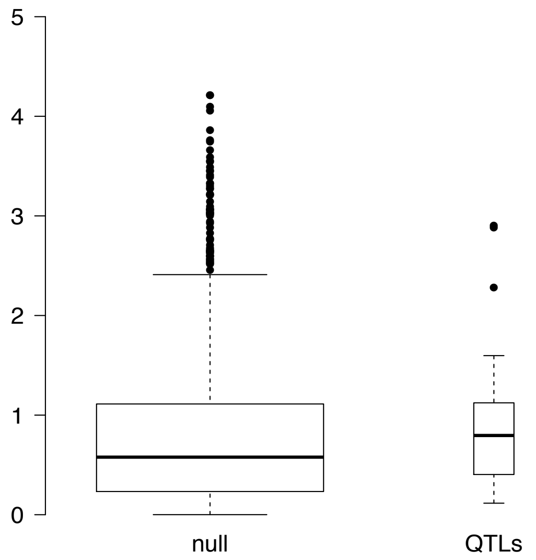


Figure 5-7 Comparison of the significance of the association between loci taken for replication in an independent sample of HS rats and random loci. The significance (logP) of the association, calculated in a fixed effect model, is plotted on the vertical axis. 46 QTLs detected using the mixed model method in the HS were tested for association with the same phenotype in the independent sample of HS rats.

5.3.6 Comparison with QTLs reported in the Rat Genome Database

I compared the QTLs detected in the rat HS with QTLs listed in the Rat Genome Database (RGD[171]). Care was taken to only consider the precise same measures in both cases. Whenever QTLs mapped in the HS and QTLs reported in RGD overlapped, I calculated a significance value to establish whether this was due to chance only or a true shared genetic basis. It is important to note that the confidence intervals for the QTLs reported in RGD are very large, of the order of dozens of Mb, because they were mapped in intercrosses and backcrosses. To account for this in the calculation of significance, I selected matching random QTLs. I found a significant overlap for the number of arterial elastic lamina ruptures, total cholesterol levels and heart weight (at a nominal P value of 0.05; APPENDIX D). Significant overlap suggests that the phenotype under consideration has a common genetic basis in the rat HS and the population reported in RGD. Lack of overlap on the other hand might reflect spurious associations in either population. However, it likely often reflects true differences in causal genetic variation since no population in RGD shares all eight founders with the rat HS: they at best share two founders with it, leaving the possibility that the six other founders contribute genetic variation influencing the trait in the HS, which could create differences in QTLs.

5.4 Discussion

In this chapter I presented simulations that guided my choice of using either mixed models or resample model averaging to map each phenotype. I showed that mixed models are generally preferable for normally distributed phenotypes, while resample model averaging has a better FDR for non-normally distributed traits. I then reported the mapping of 355 QTLs for 122 measures, which on average individually explain 6.5% of phenotypic variation, and jointly explain 42% of heritability on average. I will discuss these results in Chapter 8.

We used average sizes to define the confidence interval of each QTL. Doing so ignores differences in local linkage disequilibrium (LD) and therefore is not optimal. To determine a confidence interval more precisely, LD blocks can be identified from simulations of QTLs around the position of the detected QTL (Figure 5.8).

Finally, I showed that the loci that contribute to behaviour in this study do not replicate in an independent sample of HS rats. A number of reasons may explain this lack of replication. First, the significance of the QTLs in each HS sample (the sample used throughout this study and the independent sample) was estimated by fixed effects models that do not account for relatedness because it was not possible to reliably infer genetic relationship in the independent sample from only 55 markers at 22 loci. Accounting for relatedness using mixed models would modify the significance of the QTLs. Therefore, while we would expect the logPs calculated in suitable mixed models for each sample (if we knew the genetic relationship matrix of the independent sample) to be correlated, it is not entirely unexpected that the logPs calculated in models that do not account for relatedness are not correlated. Another possible explanation for the lack of replication is a change in the phenotyping pipeline between the two experiments. The HS rats used in the replication study were subjected to additional behavioural tests before and in between the tests reported here: they were tested in the black and white box first of all (additional test), then in the novel cage (test reported here), zeromaze (test reported here), fear potentiated startle test (additional test) and finally in the shuttlebox (test reported here). It is likely that these additional tests affected the anxiety levels of the rats used in the replication experiment. Finally, the frequencies of some causal alleles could have changed between the two generations, as a result of genetic drift or occasional changes to the original breeding scheme. These changes happened for example when no male or no female was born in a litter.

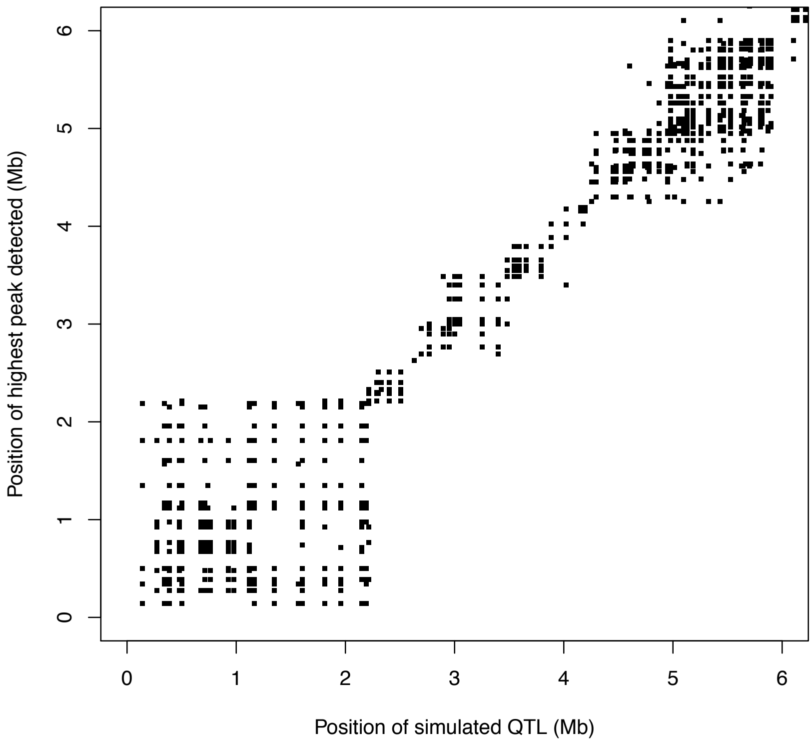


Figure 5-8 Investigation of local linkage disequilibrium (LD) to refine a QTL's confidence interval. Following detection of a QTL for number of aortic lesions on chromosome 5 at 1.6Mb, local LD was investigated by simulating QTLs in the first 6Mb of chromosome 5, and recording the position of the QTL then detected in that region. The lamina lesions QTL sits in a 2Mb-wide LD block at the beginning of chromosome 5. We refined its confidence interval by using for our calculation only on those simulations where the effect was simulated in the LD block encompassing the lamina lesions QTL.

Mapping the QTLs is the first step towards identifying the sequence variants and genes that contribute to phenotypic variation. That is the subject of the next chapter.

Chapter 6

Integration of QTLs with sequence information

Identifying the functional DNA elements that cause phenotypic variation is necessary in order to link with known pathways and processes, or to experimentally manipulate the gene or the protein it encodes to assess function. The rat HS offers unprecedentedly high mapping precision (90% confidence intervals on average 4.5Mb wide), but these segments still typically encompass about 40 genes. Except for a minority of cases where fewer genes are in the confidence interval, there are still too many genes at the QTL to allow efficient confirmation and follow up in the laboratory.

Fortunately, the descent of the HS from eight known founders permits a test, called merge analysis[104], of whether a sequence variant is responsible for phenotypic variation. In this chapter I describe how I integrated the sequence data from the founders with the mapping results to rule out, by merge analysis, the majority of sequence variants at each QTL, and thereby reduce the list of variants and genes that could be causal ("candidate variants" hereafter). With candidate genes and variants in hand, it was also possible to investigate pleiotropy in the HS.

6.1 Methods

6.1.1 Implementation of merge analysis in a mixed-model framework

To integrate sequence data with mapping results and identify potential causal variants, I used a test called "merge-analysis"[104], whose principle is illustrated in Figure 1-1. Merge analysis is a form of imputation appropriate to populations descended from more than two known founders, such as heterogeneous stocks. The genomes of the individuals in the population are mosaics of known founder haplotypes. Merge analysis asks two questions at each variant identified as segregating in the founders: is the variant associated with the phenotype? (a standard test of association), and is its association as significant as the association in the haplotype-based test in the locality of the variant? I implemented merge analysis in a mixed-model framework by comparing

$$y = X\beta + P_L T_L + \upsilon + \varepsilon \qquad \text{(model 1)}$$

and

$$y = X\beta + M_G U_G + \upsilon + \varepsilon \qquad \text{(model 7)}$$

where G is a sequence variant in interval L and M_G is the merge matrix for the variant, formed by summing those columns of P_L that carry the same allele at G (each column of P_L represents one founder strain). This can be computed efficiently by defining a matrix B_G that encodes the columns to be merged such that $M_G = P_G B_G$. This test is applied at every variable site in the catalogue of single-nucleotide variants that segregate between the eight HS founders (Chapter 2). From a statistical point of view, there is no difference between two variants with the same strain distribution pattern at a locus; they will give the same merge analysis result.

Because models (1) and (7) are nested, the best possible fit (in terms of variance explained) is obtained with haplotype model (1). If the QTL arises from variation at a single variant *G*, the fit of merge model (7) for variant *G* will be as good as the fit of model (1), and its significance will be greater, owing to the fewer number of degrees of freedom (for a diallelic variant, there is 1 degree of freedom for the merge model compared to up to 7 degrees of freedom for the haplotype model). Therefore, under the assumption that the QTL arises from a single causal variant, any variant with a merge logP smaller than the haplotype logP is unlikely to be causal.

We called variants with merge logPs greater than the maximum haplotype logP at the QTL “candidate" variants because they may be causal. The merge model is fitted as detailed in Chapter 5 by multiplying by A^{-1}. To measure whether a single variant explained a QTL, I calculated the difference $d = \text{logP}_{\text{merge}} - \text{logP}_{\text{haplotype}}$, where $\text{logP}_{\text{haplotype}}$ is the maximum logP of the haplotype test of no association and $\text{logP}_{\text{merge}}$ is the maximum of all merge logP of variants included within the QTL. Any imputed variant with a merge logP that exceeded the maximum haplotype log10 P-value was termed a candidate variant. If $d < 0$, then no candidate variants existed at the QTL.

6.1.2 Simulating all possible strain distribution patterns at a QTL

For each QTL lacking candidate variants, I looked for unobserved causal variants that might not have been sequenced. We simulated variants with every possible SDP (127 possible SDPs for diallelic variants and 1,094 possible SDPs when allowing for 3 alleles). Simulated variants were generated within each marker interval of the QTL.

6.1.3 Simulating different QTL architectures.

I investigated the hypothesis that an inability to detect candidate variants by merge analysis (where no sequence variants with merge logP exceeding the haplotype logP exist) reflected complex architecture of the QTLs. To do this, I simulated QTLs arising from a single causal variant, QTLs arising from multiple causal variants within the same locus and/or multiple causal variants at linked loci, and QTLs arising from haplotypic effects not reducible to individual variants. In all cases, the phenotypes were simulated from three components: a genetic random effect explaining 20% of phenotypic variation, uncorrelated errors explaining 75% of phenotypic variation and a single QTL explaining 5% of phenotypic variation. When multiple causal variants were simulated, each explained the same proportion of phenotypic variation (5% divided by the number of causal variants). The effect sizes calculated *a posteriori* could be quite different from their target values owing to correlations between the different components of the simulated phenotypes. For the simulations reported in Figure 6-3, either a single causal variant was simulated or nine causal variants were simulated in three linked loci (with each locus within 2 Mb of the central locus and separated by at least 200 kb from each other locus). Alternatively, the P_L probabilities were used to simulate irreducible QTLs. I analysed each simulation by merge analysis, and, when $\text{logP}_{\text{haplotype}}$ was between 4 and 6 (to have a similar distribution of logP values to that of the rat QTLs), I calculated $d = \max \text{logP}_{\text{merge}} - \max \text{logP}_{\text{haplotype}}$. I compared the distributions of d from the different simulation sets to determine the probable genetic architecture of the QTLs.

6.1.4 eQTL mapping and merge analysis in the mouse HS

I investigated the ability of merge analysis to identify causal variants at expression QTLs (eQTLs) by mapping expression levels in the hippocampus of 460 heterogeneous stock mice[59] using mixed models similar to those used to map the phenotypes. QTLs were called in the same way as for the phenotypic QTLs but using a confidence interval of 8 Mb and a significance threshold of 4. *Cis*-eQTLs were defined as being within 2 Mb of the beginning of the probe, and *trans*-eQTLs as being on a different chromosome than that of the probe, or more than 10 Mb away from it on the same chromosome. Merge analysis was carried out at each eQTL, and the difference between the maximum merge log P value and the maximum haplotype log P value was calculated.

6.1.5 Protein structure modelling

Modelling of the structure of the rat proteins by homology with published structures and prediction of the consequences of sequence variation were carried out by Tomas Malinauskas, and is described in Baud et *al.*[174].

6.1.6 Number of genes mapping to a QTL

The number of genes mapping to each QTL confidence interval was calculated using Ensembl protein-coding genes and genes coding for microRNAs (downloaded from BioMart48).

6.1.7 Overlap between QTLs mapped for different measures in the HS

For each pair of phenotypes mapped in the HS, I identified those protein coding and miRNA genes common to both sets of QTLs. If the set was non-empty, I calculated the p-value of the overlap by sampling for each measure 2,000 sets of intervals at random as described above, and calculated the number of genes in common for each pair of random sets, and thereby obtained a null distribution. The p-value for the overlap was obtained from this null distribution. For the pairs of measures that overlapped by at least one gene I identified shared merge-defined candidate variants and the genes they are in, as well as any gene for which both measures had distinct candidate variants.

6.1.8 Significance thresholds for the merge and SNP-based mapping analyses

200 of the 1,000 simulations used to get significance thresholds for the haplotype-based method were used to get significance thresholds for the SNP- based method and merge analysis. The 65^{th} percentile of the extreme value distribution was used, as for the haplotype-based analysis. Significance thresholds for the SNP-based and merge analyses were thereby obtained for 12 and 20 measures respectively, and we extrapolated significance thresholds for all the other measures from a regression of these thresholds on the thresholds for the haplotype-based method. The R^2 of the correlation between the haplotype-based thresholds and the merge analysis thresholds was 0.96 ($p<0.05$) and that for the SNP-based thresholds 0.73 ($p<0.05$). QTLs were then called in the same way as for the haplotype-based analysis.

6.2 Results

6.2.1 Identification of causal variants and genes by merge analysis

I investigated the extent to which our near-complete catalogue of segregating sequence variants in the eight HS founders would identify genes and causative mutations, using merge analysis[104]. At 131 out of 343 QTLs (38%) I identified at least 1 candidate variant (APPENDIX C). Focusing on these candidate variants rules out a causal role for the great majority of sequence variants (usually over 90%) within most QTLs (Figure 6-1a), and thereby helps identify causal genes at a QTL.

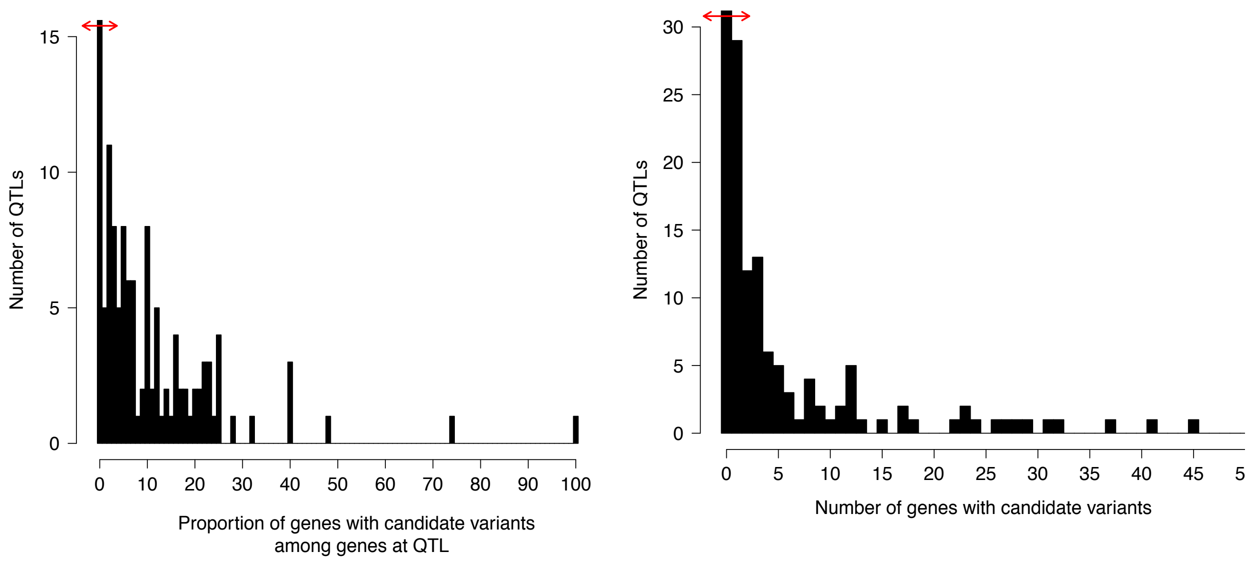


Figure 6-1 Genes harbouring candidate variants. (a) Proportion of the genes at each QTL that have candidate variants. The y axis is truncated. (b) Number of genes at each QTL that have candidate variants. The x and y axes are truncated.

I found 28 QTLs at which only a single gene contained candidate variants in its transcribed sequence or in the upstream or downstream sequences 5kb from it (Figure 6-1b, Table 6.1). One example was *Ctnnd2* (encoding catenin δ2) at a QTL for a measure of conditioned fear (Figure 6-2a). CTNND2 is a protein found in complexes with cadherin cell adhesion molecules at neuronal synapses and *Ctnnd2* KO mice have been shown to be impaired fear conditioning[175]. Another example involved a locus influencing heart weight, where, out of 82 coding genes within the QTL, only *Shank2* contained candidate SNPs (Figure 6-2b). *Shank2* encodes a synaptic protein[176] not previously associated with cardiovascular physiology.

At other loci, multiple genes harboured candidate variants (Figure 6-1b) but merge analysis nonetheless reduced the number of genes likely to be causal by about 90% (Figure 6-1a). This approach confirmed a well-established relationship between a cluster of apolipoprotein genes at a QTL on chromosome 1 and cholesterol biosynthesis (HDL, LDL, and total cholesterol). It also identified a locus influencing platelet aggregation on chromosome 4 that harbours the von Willebrand factor gene (*Vwf*), encoding a key glycoprotein involved in blood coagulation.

Finally, merge analysis contributed to an understanding of the pathogenesis of EAE, an autoimmune neuroinflammatory disease with clinical and pathological similarities to multiple sclerosis[177]. Indeed, the major histocompatibility complex (MHC) class II region on chromosome 20 (*Eae1*) is known to influence EAE susceptibility, but attempts to identify the responsible gene have had limited success. In this study, the two variants most likely to underlie the QTL effect on chromosome 20 (with the highest merge log10 P-value) were a variant in an intron of *Btnl2* and a variant 274 bp upstream of *RT1-Db1*, both in the MHC class II region. The human ortholog of *RT1-Db1*, *HLA-DRB1*, is associated with multiple sclerosis, with risk allele *HLA-DRB1**15:01[64].

When merge analysis identified candidate variants in coding regions, I considered those predicted to affect protein structure more likely to be causal. First, at a QTL for expression of CD45RC at the surface of T-cells that encompassed the gene encoding CD45RC (*Ptprc*, Figure 6-2c), I postulated that a non-synonymous coding variant was highly likely to be causal because it changed an amino acid (p.Arg114His) in the protein domain of CD45RC that binds the antibody mediating the quantification of CD45RC. Second, I found 13 other candidate variants that were predicted to affect protein structure and/or function based on homology modelling: out of 91 non-synonymous candidate variants identified, the structural consequences of the mutations on protein structure could be predicted for 43 of them (for a further 48 candidate variants, there were no homologies with known protein structures), and 13 variants in nine genes and for 8 measures contained candidate variants for which structural evidence suggested that protein structure or interactions might be altered (Table 6.1).

An example is shown (Figure 6-2d) for the protein TBX21, encoded by a gene within a QTL on chromosome 10 influencing the proportion of CD4+ cells with high expression of CD25. Here the candidate variant changed glycine to arginine (p.Gly175Arg), and the substitution with arginine could alter the DNA-binding characteristics of this protein. *Tbx21* has been implicated in the genetic control of regulatory T cells[178], a subset of T cells with high surface expression of CD25.

A second example is that of a candidate variant (p.Thr233Met) found in *Abcb10*, a gene located at a QTL for mean red blood cell volume. The crystal structure of human mitochondrial transporter ABCB10 is shown in Figure 6-2e. p.Thr233Met maps to a position in the protein structure where the side chain of the residue points to the centre of the transporter channel (Figure 6-2e). Threonine has a polar, uncharged side chain, whereas methionine has a hydrophobic side chain. Therefore the p.Thr233Met alteration positions a larger, bulkier residue in a region of ABCB10 that is tightly packed in the open-outward conformation of ABC transporters, potentially interfering with the conformational changes that are essential for transport of the substrate. Evidence from mouse knockouts indicates that *Abcb10* gene is essential for erythropoiesis[179,180].

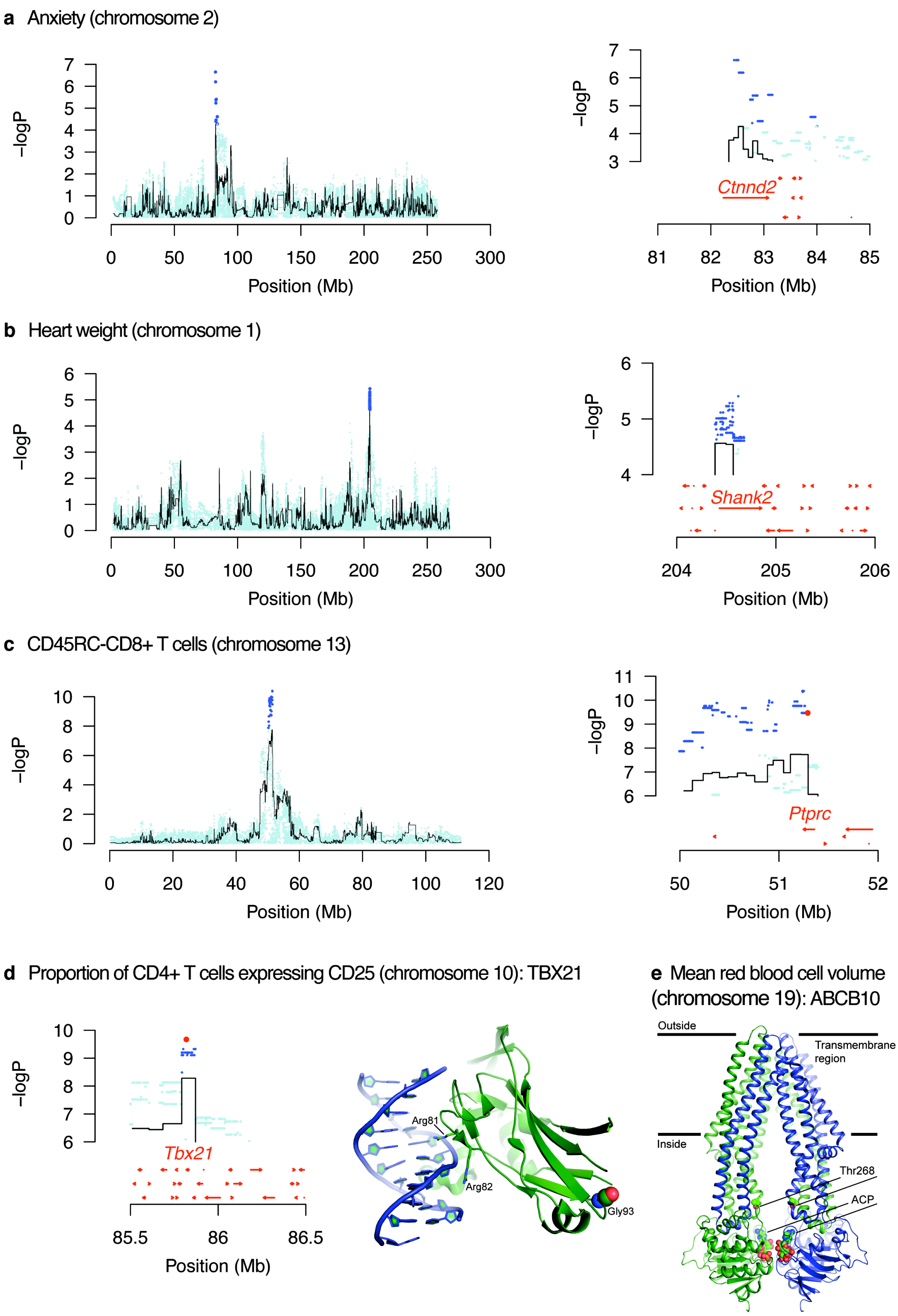


Figure 6-2 Merge analysis to identify causative genes and sequence variants. The top three panels (a – c) show, on the left, the scans for a whole chromosome, with the name of the phenotype. The black lines represent the haplotype analysis and the blue dots are the merge analysis results of testing for association with all sequence variants identified in the

progenitor strains. On the right is an enlargement of the highest peak showing the location of candidate variants and genes. Candidate variants are those whose significance exceeds that of the haplotype analysis (i.e. blue dots are above the highest value of the black line). Genes are shown by red arrows. (d) Candidate variants on chromosome 10 for the proportion of CD4+ cells with high expression of CD25. The highest variant lies within the TBX21 protein. The crystal structure of human TBX5-DNA complex (PDB code 2X6V) maps the location of the rat TBX21 mutation Gly175Arg to the DNA binding domain. The structure of TBX5 (green) complexed with DNA (blue) is shown in ribbon representation. Gly93 is shown as spheres (C atoms in green, O atoms in red N atoms in blue). Gly93 and corresponding Gly175 (rat) are conserved. Side chains of two arginines that mediate interactions with DNA are shown as sticks. (e) A candidate variant in the Abcb10 gene on chromosome 19 for a locus influencing mean red cell volume. The structure of the homodimeric ABCB10 (PDB code 4AYT) is shown in ribbon representation, with the monomers in blue and green. Two ATP analogues (ACP) and side chains of Thr268 are shown as spheres (C atoms in green, O atoms in red N atoms in blue and P in orange). The rat ABCB10 mutation Thr233Met lies in the central cavity of the translocation pathway. Amino acid sequence identity between rat and human ABCB10 is 84% (587 aligned residues); Thr268 in the human protein corresponds to conserved Thr233 in the rat.

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Mean response latency	2	80.23 - 84.83	*Ctnnd2*	Catenin delta-2	yes	none	-
Femur neck width	1	156.27 - 160.9	*Fchsd2*	FCH and double SH3 domains protein 2	yes	none	-
Distal femur total density	2	152.74 - 157.22	*Kcnab1*	Voltage-gated potassium channel subunit beta-1	yes	none	-
Femoral neck total density	5	4.03 - 8.22	*Eya1*	Eyes absent homolog 1	yes	none	-
Femur midshaft cortical density	6	38.24 - 41.52	*Lpin1*	phosphatidate phosphatase LPIN1	yes	none	-
Femur midshaft total area	2	43.96 - 48.57	*Ndufs4*	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial	yes	none	-
Femur work to failure	8	21.57 - 26.17	*Dpy19l1*	protein dpy-19 homolog 1	yes	none	-

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Lumbar trabecular area	20	21.1 - 25.75	*F1LW02_RAT*	Uncharacterized protein	yes	none	-
Heart weight	1	202.15 - 206.63	*Shank2*	SH3 and multiple ankyrin repeat domains protein 2	yes	none	-
Area under glycemia curve over baseline	2	80.5 - 85.11	*Ctnnd2*	Catenin delta-2	yes	none	-
Hemoglobin concentration	12	1.62 - 5.77	*Insr*	Insulin receptorInsulin receptor subunit alphaInsulin receptor subunit beta	yes	none	-
Mean platelet mass	1	193.98 - 197.88	*Dock1*	dedicator of cytokinesis protein 1	yes	none	-

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Mean platelet mass	9	52.53 - 88.11	*ErbB4*	Receptor tyrosine-protein kinase erbB-4ERBB4 intracellular domain	yes	none	-
Platelet clumps	8	100.57 - 104.81	*Clstn2*	Calsyntenin-2	yes	none	-
Platelet count	11	14.47 - 18.54	*Hspa8*	Heat shock 70kDa protein 8	yes	none	-
Absolute CD25+CD4+ cells	19	50.71 - 54.96	*Galnt2*	polypeptide N-acetylgalactosaminyltransferase 2	yes	none	-
Absolute CD8+ T cells	20	1.00 - 8.90	*RT1-Db2*	RT1 class II, locus Db2	yes	none	-
Proportion of B cells in white blood cells	10	27.1 - 31.59	*D3ZTU5_RAT*	Uncharacterized protein	yes	none	-
Proportion of B cells in white blood cells	20	1.00 - 2.66	*Olr1687*	olfactory receptor Olr1687	yes	none	-

Measure	**Chr**	**QTL location (Mb)**	**Gene symbol**	**Gene description**	**Only gene with candidate variants in QTL**	**Amino acid change with potential effect**	**Location of the residue, potential effect**
Proportion of CD4+ cells expressing CD45RC	13	36.86 - 62.54	*Ptprc*	Receptor-type tyrosine-protein phosphatase C	yes	none	-
Proportion of CD4+ cells in T cells	20	14.83 - 19.43	*RGD1559903*	Uncharacterized protein	yes	none	-
Proportion of CD8+ cells expressing of CD45RC	13	50.49 - 55.97	*Ptprc*	Receptor-type tyrosine-protein phosphatase C	yes	none	-
Proportion of CD8+ cells with high expression of CD25	19	52.29 - 56.8	*Sipa1l2*	Signal-induced proliferation-associated 1-like protein 2	yes	none	-
Lowest weight	3	121.45 - 126.25	*Pak7*	serine/threonine-protein kinase PAK 7	yes	none	-
Weight loss compared to day 0	2	169.79 - 174.4	*Fam198b*	Protein FAM198B	yes	none	-

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Serum alkaline phosphatase	3	18.49 - 23.11	*Lrp1b*	low density lipoprotein-related protein 1B (deleted in tumors)	yes	none	-
Serum chloride concentration	9	32.72 - 36.5	*Uggt1*	UDP-glucose:glycoprotein glucosyltransferase 1	yes	none	-
Serum triglycerides	4	74.8 - 79.28	*Dfna5*	Deafness, autosomal dominant 5	yes	none	-
Weight loss compared to day 0	20	2.48 - 7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	no	p.Thr182Ala	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48 - 7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	no	p.Thr182Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48 - 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.His200Arg	Surface exposed, disturbed intermolecular interactions

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Weight loss compared to day 0	20	2.48 - 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.Thr165Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48 - 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.Gln162Arg	Surface exposed, disturbed intermolecular interactions
Expression on RT1B on B cells	17	26.63 - 27.55	*Tbc1d7*	TBC1 domain family member 7	no	p.Ser116Leu	Surface exposed, disturbed intermolecular interactions
Proportion of B cells in white blood cells	1	182.36 - 186.67	*Itgal*	Integrin alpha L	no	p.Asn891Ser	Abolish glycosylation
Proportion of CD4+ cells with high expression of CD25	10	84.27 - 87.32	*Tbx21*	T-box transcription factor TBX21	no	p.Gly175Arg	Surface exposed, additional DNA interactions
Ratio of T cells to B cells	1	183.58 - 187.41	*Rabep2*	Rab GTPase-binding effector protein 2	no	p.Ile336Thr	Partially buried, disturbed oligomerization

Measure	Chr	QTL location (Mb)	Gene symbol	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Ratio of T cells to B cells	1	183.58 - 187.41	*Itgal*	Integrin alpha L	no	p.Leu806Ser	Surface exposed,disturbed intermolecular interactions
Mean corpuscular red cell volume	19	53.11 - 55.80	*Abcb10*	ATP-binding cassette, sub-family B (MDR/TAP), member 10	no	p.Thr233Met	Transport channel-exposed, altered transport
Platelet count	12	1.00 - 7.47	*Rfc3*	Replication factor C (Activator 1)	no	p.Pro173Ala	Surface exposed, alteration of the alpha helix
Proportion of monocytes in white blood cells	1	250.37 - 254.00	*Pdcd11*	Protein RRP5 homolog	no	p.Glu160Gly	Surface exposed

Table 6.1 Summary of genes identified at QTLs and potential functional variants. The table shows the phenotype measured, the chromosome (Chr.), the start and stop coordinates of the QTL (in megabases Mb), gene symbol and description, whether the gene is the only one at a QTL with candidate variants, whether a variant alters an amino acid and if so the residue changed and potential consequences. The two variants p.Thr182Ala and p.Thr182Met in *RT1-Da* correspond to two successive nucleotide positions.

6.2.2 Single variants rarely account for HS QTL genetic effects

Unexpectedly, 212 out of 343 QTLs mapped with mixed models (62%) had no candidate variant (APPENDIX C, Figure 6-3a). We considered four explanations for this observation: (i) causative variants were missing from the sequence catalogue; (ii) haplotype mapping was biased toward QTLs without candidate variants; (iii) the merge analysis underestimated statistical significance compared to single-marker association analysis; and (iv) there were multiple causal variants at a single QTL.

First, causal variants might have been missed because our sequence data were incomplete. Despite linkage disequilibrium extending over a few megabases, not all variants were tagged by a nearby variant with identical SDP in the founders. For example, only 50% of the structural variants were tagged by a SNP lying within 1 Mb of the variation. However, because only a limited set of possible SDPs exist in the heterogeneous stock, we can test whether missing genotypes are responsible for the inability to detect candidate variants. We generated SDPs for all possible diallelic and triallelic variants at every locus within the 212 QTLs and tested each by merge analysis to determine how many would have been candidate variants. Only 44 out of 212 QTLs had candidate diallelic variants, and 165 had diallelic or triallelic variants. Thus, if the effect for each QTL were attributable to a single diallelic variant that we had not sequenced, there would still be 168 QTLs (49%) without a candidate variant. If the effect were attributable to a diallelic or triallelic variant, the fraction would be reduced to 14%. However, triallelic SNPs are very uncommon and are therefore unlikely to explain the large number of QTLs without candidate variants – in fact these hypothetical triallelic variants may well simply be tagging haplotype effects.

Second, haplotype mapping might simply not be powerful enough to detect QTLs that have candidate variants, or it might be biased towards QTLs without candidate variants. I addressed the first possibility by simulation. I report the distribution of the *d* values (differences between maximum $logP_{merge}$ and $logP_{haplotype}$ values), where, for QTLs where candidate variants exist, $d > 0$ (Figure 6-3a). When simulated QTLs arose from single causal variants, merge analysis did indeed identify candidate variants at almost all QTLs placed in random regions of the genome as well as at QTLs simulated in the same locations as the detected QTLs.

I also considered the performance of the method at QTLs where it was highly probable that a single variant was the causal variant, namely at *cis*-acting expression QTLs (*cis*-eQTLs)[181,182]. I tested 1,398 *cis*-eQTLs detected in the hippocampus of heterogeneous stock mice[59], finding that the merge analysis identified variants with logP that exceeded those of the haplotype-based test at 97% of QTLs (Figure 6-3b). Notably, when I carried out the same analysis on *trans*-eQTLs, the distribution of *d* was similar to that seen for the rat phenotypic QTLs (Figure 6-3b). This difference between *cis*- and *trans*-eQTLs was true across all logP values, indicating it is not due to lower power to detect trans eQTLs.

Because mapping QTLs using haplotype analysis might bias results toward loci without candidates (a winner's curse is likely to operate), I used merge analysis to map QTLs across the genome. The two methods did not identify the same QTLs (152 were unique to the merge method), but the merge method identified 16% fewer QTLs than the haplotype method. Notably, only 9% of the merge-identified QTLs had no candidate variants (Figure 6-4). Consequently, I conclude that haplotype mapping overestimates the number of QTLs without a candidate variant, whereas merge analysis underestimates the number of these QTLs. Therefore, my best estimate of the proportion of QTLs without candidate variants is obtained from combining both methods. In the set of QTLs identified by either merge or haplotype mapping, I found that 44% of QTLs could not be explained by single causal variants (compared to 62% when only the haplotype-based QTLs were considered). Thus, although a winner's curse does operate in favour of the haplotype analysis, it cannot account for all QTLs without a candidate variant.

The third explanation was that the merge analysis underestimates statistical significance. I compared the performance of the merge analysis with that of single-marker association analysis at genotyped SNPs. Across all phenotypes, the slope of the regression of merge logPs on single-marker logPs was 0.9, with an R^2 correlation of 0.9, and agreement strongest for the most highly associated SNPs. This result indicates that merge analysis performs as well as SNP analysis.

Finally, I investigated the extent to which multiple variants at QTLs would account for these findings. I investigated the consequences of a variety of complex QTL architectures by simulation (Figure 6-3a). Simulating multiple causal variants on different haplotypes reduced the frequency at which any single variant exceeded the maximum haplotype log P-value, although this simulated complexity was still insufficient to mimic the observed frequency of QTLs without causal variants (Figure 6-3a). Simulating irreducible haplotypic effects arising from the reconstructed haplotype mosaics in the heterogeneous stock (rather than from a selection of sequence variants) also led to fewer QTLs with candidate variants (Figure 6-3a), although, again, the simulated proportion of QTLs without variants did not match that observed with the real QTL set. My simulations suggest that the presence of multiple causal variants at a locus accounts in part for the inability to identify candidate causal variants.

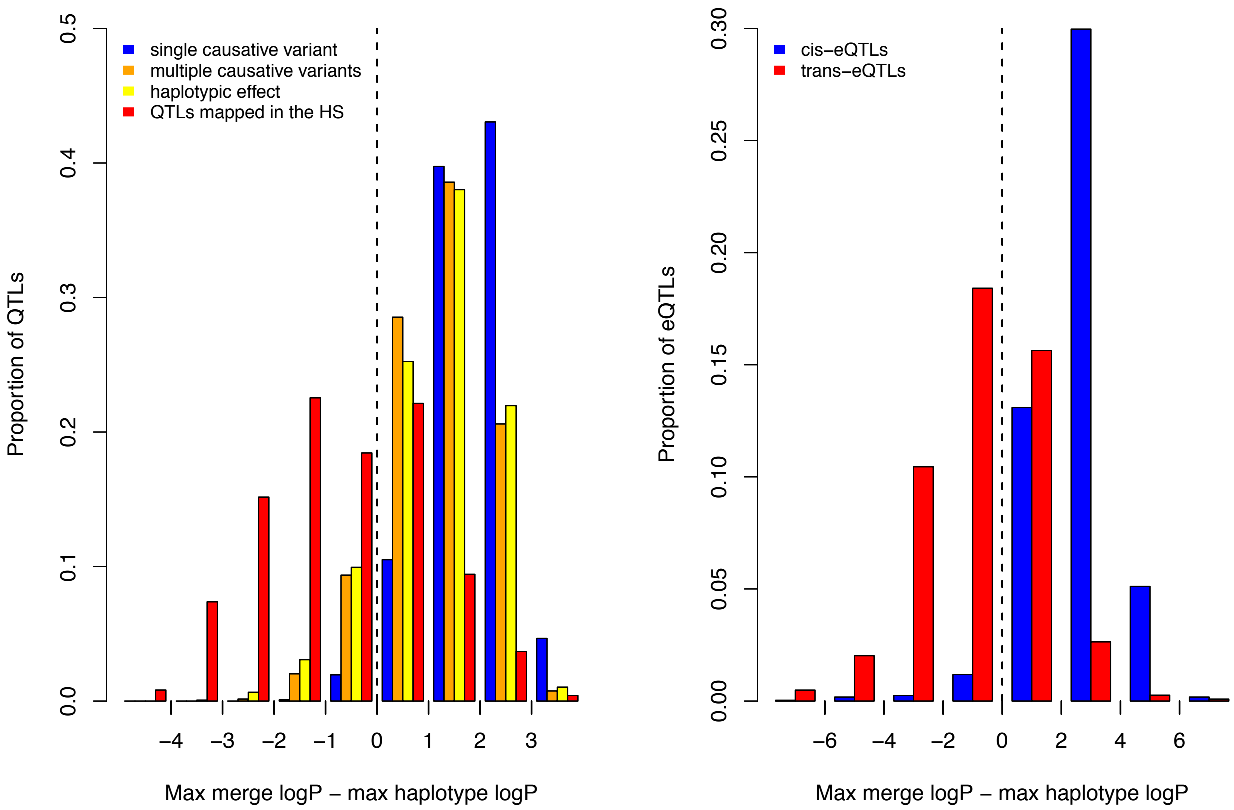


Figure 6-3 Merge analysis and simulations. The figure plots the difference between the negative logarithm 10 p-value of association (logP) of imputed variants and haplotype-based logP for the rat QTLs (a) and a set of 1,386 cis-acting and 7,464 trans-acting expression QTLs mapped in a mouse HS (b). In cases where there is a single causal variant at a QTL, the logP of some imputed variants will exceed that of the haplotypes, so that the mean of the distribution of the difference between these two logP values will be greater than zero. This is shown as a blue histogram on the plot a. The distribution observed for the phenotypic QTLs, shown in red in a, has a mean less than zero. The results of simulating haplotypic effects are shown in yellow, and in orange the consequence of simulating multiple causative variants. The distribution of the difference in logP for the mouse cis-eQTLs is shown in blue in b to highlight the resemblance with the results of simulating single causative variants. The distribution for the trans-eQTLs is shown in red in b and is most similar to that for the phenotypic QTLs.

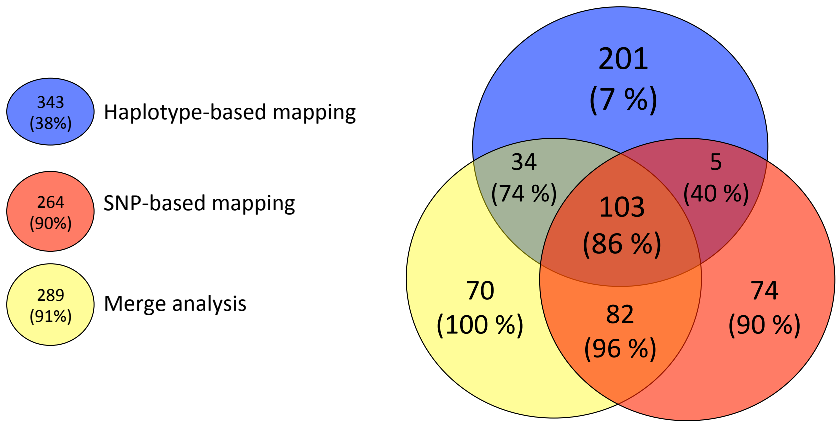


Figure 6-4 Concordance between the haplotype-based and SNP-based mapping methods, and the merge analysis. The number of QTLs detected by each method is indicated, as well as the proportion of QTLs with candidate variants (in brackets). The colour code and the figures for the full sets are indicated on the left side.

6.2.3 Pleiotropy

Fine-mapping of 160 phenotypes made it possible to investigate pleiotropic action. Table 6.2 lists 22 pairs of measures that either share candidate variants (15 pairs) or have different candidate variants but in the same gene (7 pairs). While some are for related measures (such as HDL and total cholesterol), 11 pairs are not expected and represent novel pleiotropies. For example, a set of common candidate variants in *Shank2* was found between heart weight and the proportion of T-cells in white blood cells, and different candidate variants but in the same gene (*Ctnnd2*) were found for a measure of glucose tolerance and a measure of anxiety.

For all 22 pairs, I calculated the significance of the overlap between the QTLs identified for the measures in the pair. This information is partly independent of the overlap of candidate variants because the overlap between two sets of QTLs can be significant when different candidate genes are identified, and inversely common candidate variants and genes do not ensure significant overlap of the QTLs but still provide evidence for a common genetic basis.

Of the 11 novel pleiotropies identified, 3 show significant overlap between the QTLs of the pair. One of these is the pleiotropy involving *Ctnnd2* and its contribution to glucose tolerance and anxiety.

Measure1	Measure2	Number of common candidate variants	Number of genes with common candidate variants	Number of common genes with different candidate variants	Number of overlapping QTLs	Number of overlapping genes	p-value of overlap
HDL	Total cholesterol	976	18	14	2	258	<1/2000
Body weight at day immunization	Lowest weight	153	6	0	2	121	0.025
Mean corpuscular hemoglobin	Mean corpuscular volume	211	5	0	2	101	0.0040
Plateletcrit	Platelet count	263	2	0	1	197	0.045
Expression on RT1B on B cells	**Weight loss compared to day 0**	3	1	4	1	129	0.11
Absolute T cells	**Weight loss compared to day 0**	1	1	4	1	189	0.15
Absolute CD25+CD8+ cells	**Weight loss compared to day 0**	23	1	1	1	171	0.066
Mean platelet mass	Platelet count	1	1	0	1	197	0.019
Mean platelet component	Platelet count	52	1	0	1	194	0.029
Mean platelet component	Mean platelet mass	1	1	0	1	253	0.047
Mean platelet mass	Plateletcrit	1	1	0	1	199	0.053
Mean platelet component	Plateletcrit	52	1	0	1	194	0.080

Measure1	Measure2	Number of common candidate variants	Number of genes with common candidate variants	Number of common genes with different candidate variants	Number of overlapping QTLs	Number of overlapping genes	p-value of overlap
Heart weight	**Proportion of T cells in WBC**	67	1	0	1	70	0.11
Hematocrit	Hemoglobin	4	1	0	2	59	<1/2000
Absolute B cells	WBC in the Baso channel	8	0	0	1	64	0.060
Hematocrit	**Plateletcrit**	0	0	3	1	51	0.069
Mean corpuscular hemoglobin	**Plateletcrit**	0	0	3	1	55	0.15
Area under glycemia curve over baseline	**Mean response latency**	0	0	1	1	10	0.013
Absolute CD25+CD4+ cells	**Mean corpuscular volume**	0	0	1	1	41	0.017
Absolute CD25+CD4+ cells	**Mean corpuscular hemoglobin**	0	0	1	1	68	0.046
Hemoglobin	**Plateletcrit**	0	0	1	1	55	0.061
Bone neck width	**Red blood cell count**	0	0	1	1	69	0.17

Table 6.2 Pleiotropy in the rat HS. Pairs of measures that either share candidate variants or have different candidate variants but in the same gene are shown. For each the number of overlapping QTLs, number of overlapping genes and the p-value of the overlap are indicated.

6.3 Discussion

In this chapter I showed how the QTLs can be fine-mapped using merge analysis, and in some cases causal variants and genes can be identified. I identified 35 candidate genes for 31 phenotypes, confirming known factors in some cases and identifying novel contributors in others. An unexpected result of my analysis was the observation that at the majority of QTLs the genetic effect cannot be accounted for by a single variant. I will discuss these findings and my results regarding pleiotropy further in Chapter 8.

Chapter 7

Concordance between species

In this chapter I investigate whether genetic loci contributing to variation in the same phenotype in different species are in orthologous genes and pathways (KEGG pathways, see Methods). This question is fundamental because the use of model organisms in biomedical research is based on the assumption that results obtained in them provide information on human biology. As was pointed out earlier, there is abundant evidence that gene knockouts produce similar Mendelian-type effects in different species. The question is whether natural variation segregating in different species also has similar phenotypic effects.

Data from the rat and mouse HS[2] allowed me to examine whether genes and/or pathways are conserved between rats and mice, and from there evaluate the likelihood they are conserved between rodents and humans. The mouse and rat HS studies are well suited because a large number of traits were collected in a similar manner in both studies, and QTLs are mapped at high resolution.

7.1 Methods

I compared the variants segregating in the rat and mouse HS by lifting over the coordinates of the variants that segregate in the mouse HS to the rat genome, and compared to the variants segregating in the rat HS.

I used the phenotypes and genotypes collected in the mouse HS to compare the extent to which regions associated with the same measure in the two populations are syntenic. 38 measures were collected in the two studies with the same protocols (APPENDIX E).

For consistency, we called QTLs for the mouse phenotypes using the identical methodology as used in the rat HS rather than use the QTLs reported in[2]. To do so, I mapped each measure with a mixed model, determined the significance threshold required to have a FDR of 10% across the mouse measures (50^{th} percentile of the extreme value distributions), and called QTLs using the corresponding logP thresholds and 4Mb wide confidence intervals.

I then lifted over the mouse QTLs to the rat genome using the UCSC liftover tool[183] (10kb minimum size for the query and target, minimum match of 0.1, and multiple output regions allowed), and calculated for each measure the number of protein coding and miRNA genes overlapping both a rat QTL and a lift-over mouse QTL. To estimate a P-value for each overlap, I then generated 1000 sets of intervals sampled at random on the mouse genome and such that there were as many intervals as mouse QTLs for the measure, and each interval had as many mouse genes as the corresponding QTL. I then lifted over these random intervals to the rat genome, and for each random set calculated the number of genes overlapping both a random interval and a rat QTL. I thereby obtained a null distribution from which I computed the p-value of the overlap between rat and lift-over mouse QTLs for a given phenotype. It is important to note that the same strategy, including the lift-over step, was applied to the QTLs and the random intervals. Thus we accounted for the fact that some mouse QTLs could not be lifted over to the rat genome with the criteria we used, and that others got split into multiple segments.

I performed a pathway analysis for the QTLs detected for each measure in the rat and mouse heterogeneous stocks. Kyoto Encyclopedia of Genes and Genomes (KEGG[184]) pathway terms were retrieved using the R KEGG.db package[185]. I used INRICH[186] to find enrichment of pathways in the mouse and rat QTLs. INRICH takes a set of independent genomic intervals (here the QTLs detected for each measure, which are unlinked loci following manual curation of the QTLs that were initially called by an R script) and asks whether there is enrichment in these intervals of particular sets of target features (here KEGG pathways). It reshuffles the intervals to obtain the null distribution of chance overlap between the test intervals and targets. INRICH controls approximately the SNP density within intervals, as well as their size and total number of genes. INRICH was run with the default option whereby a QTL will be counted only once even if multiple genes at the QTL are part of a certain KEGG pathway. I chose to do so because our QTLs usually encompass many genes and functionally related genes tend to cluster in the genome. QTLs for this analysis were defined by their 90% confidence interval, and were called at a low significance threshold (20th percentile of the extreme value distribution). I report the empirical p-values and the p-values corrected for all sets tested, both of which are output by the program.

The phylogeny of the MHC and PLCL2 regions was investigated using the UCSC Genome Browser (http://genome.ucsc.edu/) and its "Net" tracks.

7.2 Results

Because mice and rats diverged 12-25 million years ago[133,187-189] and have large effective population sizes[190], the majority of the variants that segregated in their most recent common ancestor are expected to have gone to fixation. Shared variants in the two populations could also have arisen from independent occurrences of the same mutation. I used the sequence data from the 8 founders of the rat HS and the same data from the 8 founders of the mouse

HS[105] to estimate the proportion of variants segregating in the mouse HS that are orthologous to variants segregating in the rat HS. I found that only 0.36% of the variants that segregate in the mouse HS and could be lifted over to the rat genome also segregate in the rat HS (70% of the variants segregating in the mouse genome could be lifted over to the rat genome).

The fact that different genetic variants segregate in the HS mice and rats leaves open two possibilities: different variants in orthologous genes contribute to the same phenotype in both species, or variants in non orthologous genes do. To evaluate these possibilities, I used 38 measures, listed in APPENDIX E, that were collected in both HS studies with the same protocols, mapped them with the same methods, and determined the overlap between the QTLs mapped in the mouse and rat HS.

Only one measure, the ratio of CD4+ to CD8+ T cells, showed overlap at a 10% FDR threshold and with QTLs defined by their 90% QTL confidence interval. This overlap was not significant however, even without accounting for multiple testing (empirical P value of 0.1). I repeated the analysis using QTLs called at a lower significance threshold (20th percentile of the extreme value distribution for each measure) and expanding the width of each QTL to 8 Mb. Overlaps for eight phenotypes, only two of which were significant at an empirical P value of 0.05 (serum urea concentration and the ratio of CD4+ to CD8+ T cells), were found (Table 7.1). Three QTLs contributed to the significant overlap between mice and rats QTLs for the ratio of CD4+ to CD8+ T cells. One of them, on rat chromosome 20 and mouse chromosome 17, is a notable example of locus acting similarly in different species: it lies within the MHC in humans[191] and mice[192] and we now know in rats as well. One of the other loci contributing to the overlap for the CD4+ to CD8+ ratio is situated next to the MHC locus in mice but on a different chromosome from the MHC locus in rats (chromosome 9). This result suggests that the QTL on chromosome 17 in the mouse HS actually corresponds to two linked QTLs, which are unlinked and contribute to phenotypic variation in the rat HS. This is further supported by the detection of two QTLs on mouse chromosome 17 by resample model averaging, one overlapping with the MHC and one overlapping with the region syntenic to the rat chromosome 9 QTL. In humans (and other mammals), the two rat QTLs are syntenic to two regions on different chromosomes, the human leukocyte antigen locus (HLA, the human equivalent of the mouse and rat MHC) on chromosome 6 and the region 3p24.3 on chromosome 3 (Figure 7-1 A).

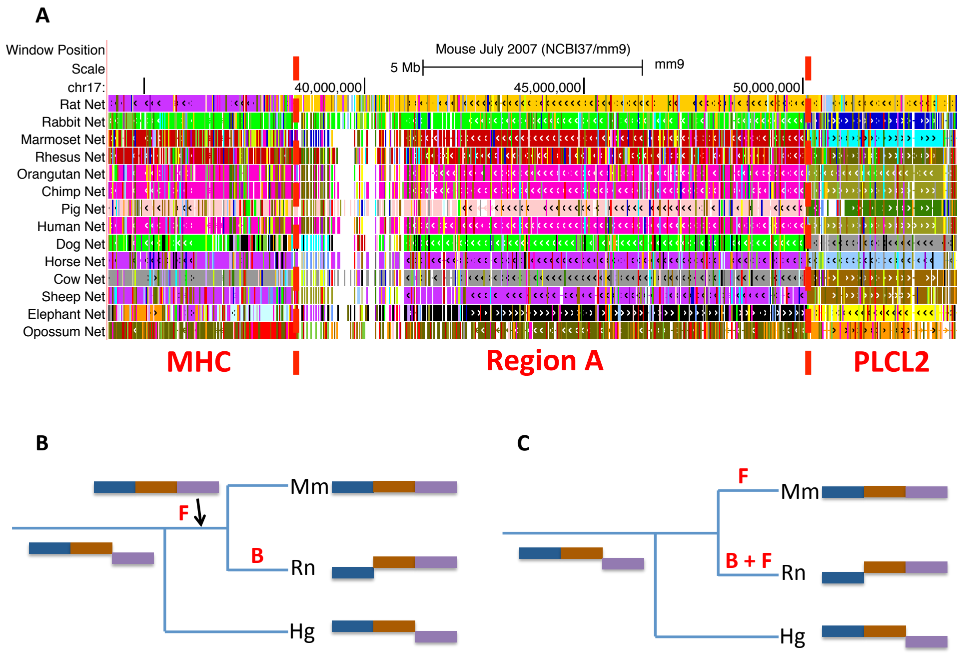


Figure 7-1 Synteny between three regions in rats, mice, and other mammals, and their phylogeny. Panel A shows an alignment of the rat genome (top row), the human genome (eigth row) and other mammal genomes to a 10Mb region of mouse chromosome 17. Three regions have been identified: one overlapping with the MHC locus, one overlapping with *Plcl2* (a candidate for the QTL on chromosome 9 in the HS rats), and a region “A”. Panels B and C show two hypothetical phylogenies for these three regions. Mm: mouse, Rn: rat, Hg:

human. Blue segment: MHC, brown: A, purple: PLCL2. F indicates a fusion between two segments, B a break.

Measure collected in both rat and mouse HS	Rat chromosome	Start rat QTL (bp)	Stop rat QTL (bp)	Mouse chromosome	Start mouse QTL (bp)	Stop mouse QTL (bp)	p-value of overlap
CD4/CD8 ratio	2	80,514,861	88,514,861	8	71,702,719	79,702,719	
CD4/CD8 ratio	20	1	21,130,034	17	29,765,847	49,461,862	0.009
CD4/CD8 ratio	9	163,535	8,163,535	17	29,765,847	58,223,414	
Serum urea	3	42,216,359	50,216,359	2	62,251,440	70,251,440	0.017
Serum calcium	12	32,823,911	40,823,911	5	122,622,730	130,622,730	0.082
White blood cells	10	57,687,472	71,772,794	11	64,922,467	72,922,467	0.115
White blood cells	20	47,407,891	55,242,732	10	40,739,571	48,739,571	
T/B cells ratio	13	76,732,108	84,732,108	1	169,631,425	177,631,425	0.149
T/B cells ratio	20	37,587,352	45,587,352	10	36,253,282	48,688,440	
Serum chloride	9	30,612,920	38,612,920	13	2,905,961	15,193,693	0.22
Monocytes	20	166,573	8,166,573	17	20,998,210	28,998,210	0.301
Serum total cholesterol	4	17,089,467	25,089,467	5	12,515,714	20,515,714	0.598

Table 7.1 Syntenic QTLs mapped in the rat and mouse HS for the same measure.

I next asked whether the genes that lie in the syntenic QTLs were in the same pathways. To tackle this question, I asked for each of the 38 measures whether the same KEGG pathways were enriched for QTL-associated genes in both mouse and rat heterogeneous stocks. QTLs for this analysis were called at a low significance threshold because loci that do not quite reach the high significance threshold associated with a FDR of 10% may still harbour true signal and significance of the enrichment is controlled independently of the significance of the QTLs. APPENDIX F and APPENDIX G present the results of the enrichment analysis carried out with INRICH. Only two measures, mean corpuscular volume (MCV) and proportion of B cells in WBC (pctB), have QTLs enriched (at a non corrected p-value threshold of 0.05) in genes part of the same pathways in both HS: "Dorso-ventral axis formation" for MCV, and "Jak-STAT signaling pathway" and "Hematopoietic cell lineage" for pctB. The mouse QTLs for pctB are enriched in "Hematopoietic cell lineage" even at a corrected p-value threshold of 5%. Five rat QTLs and seven mouse QTLs harbored genes annotated in the “Hematopoietic cell lineage” pathway, and three of these QTLs in each species overlap after lift over. This shows that the pathway analysis incorporates additional signal from those QTLs that do not overlap between the two stocks but harbor functionally related genes. Overall though, there is little evidence that the genes in the same pathways are associated with a phenotype in the mouse and rat HS.

7.3 Discussion

In this chapter I reported overlap between QTLs mapped in the rat and mouse HS for homologous phenotypes. I will discuss the apparent lack of conservation in Chapter 8. Here I discuss the small proportion of orthologous variants segregating in the mouse and rat HS, the genes underlying the QTLs for the CD4+ to CD8+ ratio on chromosomes 20 and 9 in the rat, as well as the processes that brought together two QTLs for that ratio in the mouse.

Orthologous variants segregating in the mouse and rat HS may either be ancestral and have been maintained by balancing selection since the mouse and rat lineages split, or they may be independent occurrences of the same mutation. The latter can happen anywhere in the genome by chance, but is more likely at mutation hotspots (such as the one in p53[193]). As for balancing selection, it has been documented for polymorphisms in the MHC[194] and ABO blood groups[195] loci, and Leffler *et al.*[196] recently identified several other cases. I did not investigate the origin of the orthologous variants that segregate in the mouse and rat HS.

The MHC/HLA class II gene *RT1-Da* has been identified as contributing to the CD4+ to CD8+ ratio in mice and humans[191,192], and might therefore underlie the QTL on chromosome 20. In rats, the molecular nature of the locus on rat chromosome 9 is unknown (Maja Jagodic, personal communication). The corresponding region in Humans, 3p24.3, has been associated with a number of immune diseases, including inflammatory bowel disease[197], multiple sclerosis[198], primary biliary cirrhosis[199], Crohn's disease[200], psoriasis[65], and HIV-1 control[201]. In these studies, the lead variant is either intergenic or in phospholipase C-like 2 (*PLCL2*)[198,199]. The protein PLCL2 lacks the catalytic activity that, in other phospholipases C, is responsible for the hydrolysis of phosphatidylinositol 4,5-biphosphate. It has nevertheless been shown to be involved in a number of processes in mice, among which the regulation of gonadotropin levels[202] and the trafficking of gamma-aminobutyric acid receptors to the cell surface[203]. Interestingly, using a genetically modified mouse model where the second exon of *Plcl2* was missing, Takenaka *et al.* showed that PLCL2 negatively regulates BCR signalling and immune response in mice[204]. However, they found no evidence that PLCL2 also affects T-cell development. This could result from the involvement of a different *Plcl2* isoform, one that lacks exon 2, in T-cell development.

However, no Expressed Sequence Tags (EST) supporting the existence of such an isoform could be found. *Satb1* is the best candidate gene in the vicinity of *Plcl2* because it is a known regulator of T cell development and differentiation in CD8 single positive cells [205,206]. However, *SATB1* is more than 1Mb away from the lead variants in those human GWAS that suggested the implication of *PLCL2*, making *SATB1* a less likely candidate for the association. What is more, we (Jonatan Tuncel and I) investigated the role of *Satb1* in the HS rats and found no correlation between the expression of *Satb1* in the thymus and the CD4+ to CD8+ ratio (unpublished data).

Functionally related genes sometimes cluster together on the genome (e.g. HOX genes, the genes in the MHC, olfactory receptors[207], and housekeeping genes[208]). This clustering can happen through multiple mechanisms: functionally related genes can originate from tandem gene duplications[209]; natural selection can generate and maintain the linkage of genes when genomic rearrangements create favorable combinations of alleles (epistasis, [209]); it is also suggested that rearrangements are more likely between regions that are euchromatic at the same time (C. Berthelot *et al.,* submitted), typically regions containing genes that are expressed at the same time. It is intriguing that two regions contributing to the same phenotype - the MHC region and the region of *Plcl2*, which both contribute to the ratio of CD4+ to CD8+ T cells - are next to each other in the mouse genome and not in the rat and other mammalian genomes (Figure 7-1 A). Could it be that a rearrangement that placed the MHC and PLCL2 regions next to each other in the mouse lineage was selected because it linked favourable combinations of alleles of the MHC locus and *Plcl2*? Statistical testing of the contribution of epistatic interactions to phenotypic variation did not support this hypothesis (p-values of 0.33 and 0.076 respectively in the rat and mouse HS, analysis of variance between a model with both main and epistatic effects and a model with main effects only). What is more, although it was not possible to determine the phylogeny of the regions on close examination of the genomes alignments, the two most likely scenarios, shown in Figure 7-1 B and C, require a break in the rat lineage, suggesting that the linked state was not maintained by natural selection.

Chapter 8

Discussion

My thesis set out to investigate the relationship between genetic and phenotypic variation. In this discussion I return to the main findings in each chapter and consider how those results address the questions I raised in my introduction. To reiterate, the three questions I investigated are: what is the genetic architecture of complex phenotypes in the rat HS? How does sequence information contribute to the identification of causal variants? Are the same genes involved in phenotypic variation of homologous traits in different species?

8.1 The genetic architecture of phenotypes in the rat HS

The genetic architecture of most traits in the rat HS is complex, consisting of multiple loci each contributing a small amount of phenotypic variation. This summary statement conceals great diversity in the number of QTLs identified and their effect sizes. For example, we found only one significant locus for measures of number and size of ruptures in the arterial elastic lamina, whereas we detected ten loci for body weight. Finally, we found little evidence for pleiotropy. I discuss these points in more detail below.

The observation that the genetic architecture of the rats HS is in general polygenic should not lead to the assumption that it is comparable to the genetic architecture of human populations. Considerable amounts of human GWAS data have shown that complex trait loci make very small contributions, with odds ratios often not much over 1.1, or explaining less than 0.1% of variation. Jointly, the loci rarely explain more than 10% of the heritable phenotypic variation. The rat data are quite clearly not like this: effect sizes are larger (mean 6.5%) and the loci explain more of heritability in the HS (mean 42%).

Differences in genetic architecture between the rat HS and human populations are the result of higher allele frequencies and fewer variants segregating in the HS. The striking difference in the distribution of MAFs between HS rats and humans (Figure 3-4) is a result of the breeding of the HS from eight inbred strains, which constrains allele frequencies to be initially equal to or greater than 1/8. QTLs have greater effect sizes as a result because the effect size v_q of a locus q is a function of the MAF p_q of the causal variant and the allelic effect a_q:

$$v_q = p_q(1-p_q)a_q^2 / 2$$

The fact that fewer variants segregate in the rat HS than in human populations also contributes to larger effect sizes. There are 5M SNPs and indels with MAF greater than 1/8 segregating in Europeans (data from the 1000 Genomes Project[53]) and many more rare variants[53,60], compared to a total of 7.8M SNPs and indels segregating in the HS. The number of genetic variants in the HS is also limited by the descent of the stock from eight inbred founders, which sampled only a fraction of the variation that was present in the wild population at the time the strains were established in the laboratory. Fewer variants segregating in the HS and contributing to phenotypic variation means that each QTL explains a greater proportion of heritability.

Greater effect sizes in the HS means that QTLs can be detected with much fewer rats than individuals are necessary in human GWAS. The HS is an intermediate population in terms of genetic architecture between wild populations, where a very large number of variants segregate and allele frequencies are low, and crosses between two inbred strains, where all

MAFs are 0.5 and a very limited amount of variation segregates. The right amount of genetic variation to include in a laboratory mapping population is debated, with some investigators seeking to maximise genetic variation in their laboratory population (e.g. breeding of the CC from wild-derived strains[80], or establishment of an isogenic panel of Medaka fish (*Oryzias latipes*) from wild caught animals[210]) to be able to map many QTLs, and others studying crosses between two inbred strains to maximize power to detect association[211].

The second issue is that not all phenotypes have the same architecture. The most striking examples are for arterial elastic lamina ruptures and body weight, where the only significant QTL identified for the former explains 9% of phenotypic variation while each of the ten loci identified for the latter explain only a few percent. Similarly in human studies there is notable heterogeneity in the distribution of effect sizes for different traits. For example, the Wellcome Trust Case Control Consortium analysis of seven common diseases showed that although the same number of individuals was used for each disease and a common set of controls, the number of loci varied across diseases (nine loci for Crohn's disease, only one for coronary artery disease)[155].

One possible explanation is that some phenotypes might be regulated by variants with small effect sizes while others might be regulated by variants with greater effect sizes.

Another explanation for differences in genetic architecture is that some phenotypes may be difficult to define and measure. Phenotypic categorization may not map to the genetic. In that case, an even larger number of loci will contribute to phenotypic variation, each with a small effect that is difficult to detect. This is thought to be one of the reasons why genetic studies of behaviour and psychiatric disorders have been particularly challenging[212]. Phenotypic heterogeneity also arises from phenocopies, whereby a phenotype is due to environmental factors in some individuals and to genetic effects in others. For example, high blood pressure in some individuals can be genetically driven, while it can arise from high salt comsumption in others. In the presence of phenocopies, the genetic effects explain a smaller proportion of phenotypic variation, and they are harder to detect. Phenotypic similarity may obscure mechanistic and genetic complexity.

Finally, I found little evidence for pleiotropy in the rat HS. Only 22 pairs of measures (0.08%) out of 25,600 possible pairs either share candidate variants or have different candidate variants but in the same gene. Because many measures undoubtedly have a common genetic basis (there are many related measures, for example among the measures relative to bone morphology) and we failed to detect pleiotropies for many of these related measures, we know that the analysis fails to identify a large number of pleiotropies. This is a result of limitations in power to detect QTLs, and an inability to identify causal variants by merge analysis when a QTL arises from multiple causal variants. Therefore, my analysis almost certainly underestimates the extent of pleiotropy.

11 of the 22 pleiotropies we did identify, those that are not in bold in Table 6.2, are between related measures and therefore provide little biological insight. Of the 11 others, some reflect known relationships (for example, between T cell function/repertoire and EAE, whose severity is reflected by weight loss 28 days following induction of the disease) while others uncover previously unknown relationships (e.g. heart weight and proportion of T cells in the the white blood cells). One of the novel pleiotropies stands out in terms of robustness and suggests an interesting biological mechanism, that between the area under the glycemia curve over baseline (also called ΔG), which is a measure of glucose tolerance, and the mean latency to go to the other compartment in the shuttlebox test, which is a measure of fear conditioning. The reasons why it stands out are the following: (i) it is supported by the most significant QTL overlap in the analysis (p-value 0.013); (ii) out of ten overlapping genes,

only one, *Ctnnd2*, harbours candidate variants both for the measure of glucose homeostasis and the measure of response to fear; in addition, *Ctnnd2* is the only gene with candidate variant for the behavioural measure; (iii) although different candidate variants in *Ctnnd2* contribute to variation in the two measures, one founder (ACI) bears the least frequent allele of each of the two candidate variants that are most significantly associated with each measure. The contribution of *Ctnnd2* to variation in both measures is supported by the literature: *Ctnnd2* knock-out mice show impaired fear conditioning[175]; *Ctnnd2* is one of the targets of the transcriptional repressor REST/NSRF in pancreatic cells[213], and the expression of REST targets (including *Ctnnd2*) in the beta cells of the pancreas is essential for insulin production[213]. This pleiotropy suggests that *Ctnnd2* may be responsible for coupling “flight” response to a perceived threat with increased glucose availability for the muscles.

Pleiotropies can reflect balancing selection, whereby one allele of the pleiotropic variant is selected through one phenotype and the other allele is selected through another phenotype[19]. One can speculate that balancing selection is responsible for the pleiotropy between T cell levels and EAE severity: the allele that contributes to increased T cell levels has probably been selected because it is advantageous to fight viral infections, but it also increases the risk of autoimmune disease, so that it has not been fixed.

8.2 Sequence information sometimes identifies causal variants

The main finding of my thesis is that half of the QTLs detected in the rat HS cannot be accounted for by variation at a single sequence variant. We investigated a number of technical issues that could have led to this result: (i) causative variants were missing from the sequence catalogue; (ii) haplotype mapping was biased toward QTLs without candidate variants; (iii) the merge analysis underestimated statistical significance compared to single-marker association analysis. First, our catalogue of sequence variants is incomplete: BAC sequencing showed that 17.2% of SNPs, 41.4% of indels, and 65% of structural variants are missing from it. In addition, *de novo* variants have arisen during the 62 generations between the founders of the cross (which were sequences) and the rats used for mapping. To give rise to a QTL that we have power to detect, a *de novo* variant would need to be involved in the biological processes underlying the phenotype and either have a strong effect or be present in a large fraction of HS rats (i.e. appear in the first few generations), which is quite unlikely. Nevertheless, we simulated a situation were causal variants would be missing from our catalogue and showed that they would be tagged by catalogued variants with the same SDP in the majority of cases. I addressed the second possibility by simulations and by analysing cis-eQTLs, and showed that haplotype mapping can detect QTLs with candidate variants. When both QTLs with and QTLs without candidate variants exist, as is the case with phenotypic QTLs, haplotype mapping is biased towards detecting those QTLs that do not have candidate variants. The best estimate of the proportion of QTLs without candidate variants is therefore obtained by considering QTLs identified with haplotype mapping or merge analysis, and is 44%. Finally, I verified that merge analysis did not underestimate significance by comparing the merge logPs and classical association logPs of genotyped markers.

The concern that our definition of candidate variants, based on a comparison of two analyses of variance (one comparing the haplotype model to the null model and one comparing the merge model to the null model) rather than one analysis of variance on the haplotype and merge models, might prevent us from finding candidate variants is addressed by the analysis of cis-eQTLs and simulations of single causal variants, which both show that candidate variants are detected when the QTLs are (suspected to be) due to single causal variants.

To explain why 44% of QTLs do not have candidate variants, I asked if they arose from the additive effect of multiple causal variants on different haplotypes (i.e. with different SDPs). These variants may be in the same gene (allelic heterogeneity) or in different genes (locus heterogeneity) - the mapping resolution in the HS does not allow us to say.

The hypothesis of multiple causal variants at the QTLs with no candidate variants is supported to some extent by simulations: simulating additive effects of multiple causal variants increases the proportion of QTLs without “candidate" variants. However, this proportion does not reach 0.5 - the proportion observed with the real QTLs. Therefore, the situation is likely to be even more complex at the real QTLs, not only involving multiple causal variants but also possibly interactions between them.

Instances of locus and allelic heterogeneity have been reported in a large number of studies, both for common and rare variation, Mendelian and complex traits. We saw in the introduction that allelic heterogeneity is a hallmark of Mendelian diseases. For example, 1,942 mutations have been identified in the *CFTR* gene responsible for cystic fibrosis. Examples of allelic heterogeneity in complex traits include rare alleles in *IFT140* associated with Jeune asphyxiating thoracic dystrophy (JATD, a ciliopathy)[214], multiple common variants in *PCKS9* were shown to be independently associated with coronary disease[215]. Causal variation in nearby genes is not uncommon either, with for example reports of independent effects contributing to susceptibility to multiple sclerosis in the HLA region[216].

This complex scenario will mean two things for the many laboratory populations that are descended from multiple known founders (mouse Collaborative Cross[80] and Diversity Outbred[217] - same eight founders for both, Drosophila Synthetic Population Resource[218] - 15 founders, Arabidopsis MAGIC population[173] - 19 founders, etc.). First, if interactions are at play, additive haplotypic effects will not be optimal to map QTLs, resulting in potential loss of power to detect associations.

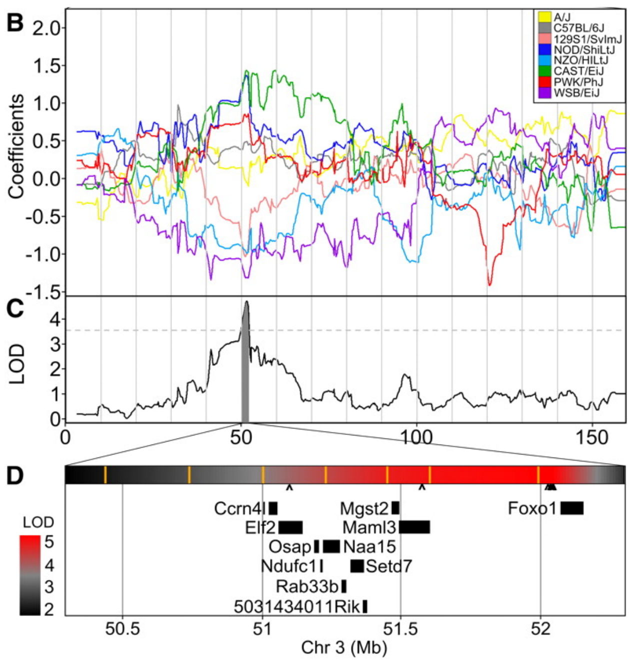
Second, sequence variation will be of limited use to identify the variants underlying the QTLs. Two strategies are commonly used amongst the groups studying laboratory populations descended from multiple founders. One is merge analysis, which identifies candidate variants under the assumption that the QTL arises from a single causal variant. In the presence of multiple causal variants, the interpretation of candidate variants as defined by merge analysis is unclear. The second strategy consists in comparing the SDP of the QTL effect with the SDPs of the variants present at the QTL, as illustrated in Figure 8-1 reproduced from Svenson *et al.*[217]. When multiple causal variants underlie a QTL, interpreting the QTL effect in terms of SDPs of the causal variants will be difficult, as illustrated in Figure 8-2. This will be even more true when the causal variants interact.

Ignoring allelic/locus heterogeneity and trying to assign the QTL effect to a single variant may lead to focus on variants that are not causal but capture the overall effect of the causal variants better than any of the causal variants (Figure 8-2). This possibility argues against trying to identify causal variants solely based on consistency between QTL and variant SDPs (equivalently, on the significance of the association).

Acknowledging possible heterogeneity would suggest investigating models with two, three or more variants. Since the number of causal variants at each QTL in unknown however, one would need to investigate a very large number of models at each QTL. What is more, the model selection is usually based on the principle of parsimony, so that models with fewer variants would be preferred over models with more variants, as long as they explain the QTL effect well enough. As illustrated in Figure 8-2, this might lead to focusing on non-causal variants.

effect well enough. As illustrated in Figure 8-2, this might lead to focusing on non-causal variants.

The extent of allelic or locus heterogeneity in laboratory populations other than the HS, and the difficulty of interpreting the association of each sequence variant, will depend on the amount and repartition of genetic variation in the founder strains. For example, the three founder strains of the mouse Collaborative Cross (CC) that are wild derived, contribute many more variants than the other five founders, and drive the majority of the QTLs detected in the CC[135]. Therefore, heterogeneity is likely to be less pervasive in the CC QTLs than in the HS QTLs. The downside of this situation however is that most of the variants at a QTL driven by a wild strain will be compatible with the QTL effect, so that sequence informative will not really help fine-map the QTL.



Svenson K L et al. Genetics 2012;190:437-447

Figure 8-1 **Figure and legend taken from Svenson et al.[217]** Change in plasma cholesterol has a significant QTL on chromosome 3. (B) **The eight coefficients of the QTL model**

show the effects on the phenotype contributed by each founder haplotype on chr 3. (C) QTL plot for the chr 3 locus. Shading identifies a two-LOD support interval. Dashed line is maximum LOD −2. (D) Expansion of the two-LOD support interval containing 11 genes. A heatmap of the QTL P-value is shown above the gene locations with SNP locations indicated by orange vertical bars. The scale of significance (red most significant) is shown on the left. **The seven Sanger SNPs that match the founder effect pattern are marked beneath the heat map with carats (^);** five of these cluster upstream of *Foxo1*.

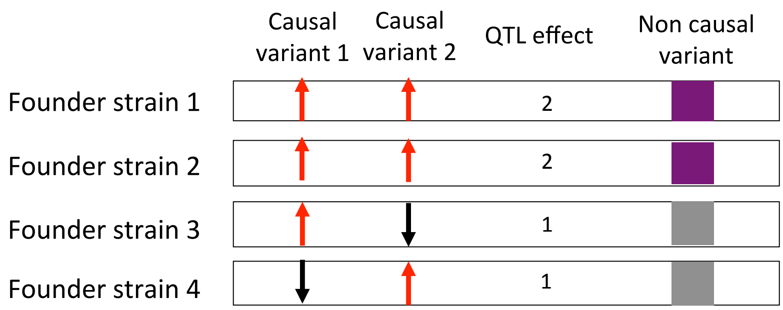


Figure 8-2 Difficulty with identifying the causal variants at QTLs arising from multiple causal variants. The genomic segment corresponding to a QTL is shown for four founder strains (for illustration purposes). Two causal variants are present at the QTL. Each has two alleles, one of which (red) increases the phenotype by 1. The effect of the QTL is the overall effect of the two variants. A third variant, whose alleles are represented in purple and grey, is shown that is not causal.

8.3 Homologous genes do not contribute to variation in the same phenotype across species

I found little overlap between the QTLs detected in the HS mice and those detected in the HS rats, and little overlap at the pathway level. While a possible explanation for these results is that different (non homologous) genes and pathways contribute to phenotypic variation in the two HS, limitations in my analysis are likely to be partly responsible. I first discuss these limitations, then examine the possibility and implications of a true lack of overlap.

Limited sampling of the genetic variation that segregates in these species is likely to contribute to the lack of overlap. If other mouse and rat heterogeneous stocks (if they existed - only one rat HS exists) were analysed, they might very well contain different sets of QTLs overlapping which would increase the total number of QTLs overlapping between mice and rats.

Perhaps more importantly, the limited power to detect QTLs in the HS experiments meant that only a few QTLs per measure could be detected; however we know that many more loci must contribute to phenotypic variation. Homologous genes could be detected as associated with a phenotype only if both of their effect sizes were large enough to allow detection, which is unlikely since only a few QTLs are detected for each trait in each population. Much larger sample sizes (e.g. 10,000) would help resolve this question.

Lack of overlap at the pathway level was at least in part due to the inability to uncover enriched pathways in each species in the first place (as seen in APPENDIX F and APPENDIX G). This is most likely because of the relatively small number of loci found for each phenotype, as well as the presence of multiple genes at each locus, which both reduce our power to detect enrichment in pathways. Using QTLs called at an arbitrary threshold is common in pathway analyses[219-221] but using all the loci in the genome weighted by their p-value, as suggested by Wang *et al.*[222], might be more powerful. However, that method was designed for pathway analysis based on results from human GWAS, where linkage disequilibrium decays very quickly so that SNPs can reasonably be attributed to the closest gene. In the mouse and rat HS, the extent of LD is much larger.

Since these limitations apply to all rodent laboratory populations, and because the two HS experiments otherwise provide unique advantages for this analysis (large number of measures measured in a similar manner in both species, high mapping resolution), I suspect it is currently not possible to accurately assess overlap between the two species. A comparison with results from human GWAS might be more powerful because it would not suffer from sampling issues on the human side, and human QTLs would be narrower. Studies have investigated the overlap between mouse or rat QTLs and human QTLs, but the rodent QTLs were so wide that they covered one third of the genome. Significance is very difficult to establish in that case.

Two possible interpretations of a true lack of overlap are that species-specific pathways are much more common than usually appreciated, and that the same pathways contribute to homologous phenotypes across species but some harbour natural genetic variation in some species and others do in other species. Our knowledge of pathways is too scant and too few genes have been identified to evaluate the second hypothesis. As for truly species-specific pathways, their existence is supported by interventional experiments as mentioned in Chapter 1[123,124], as well as gene expression analyses, which show differences in co-expression modules between species[223] in agreement with the lack of conservation in transcription-factor binding sites or in vivo binding across species[224], and protein interaction analyses[225,226]. Species-specific biology would constitute a limitation of the use of model organisms for biomedical research.

8.4 Concluding remarks

My thesis work has provided some important insights into the nature of genetic architecture in rats, and how this compares to mice and humans, the two other mammalian species where the genetic architecture of complex traits has been intensively studied. I have also been able to show how sequence information can be used to improve mapping resolution, and in some cases to identify causal variants. However in most cases my analysis did not unequivocally identify a gene involved in each phenotype. Clearly a major challenge for the future is to continue to develop approaches that will bring that goal within reach.

Appendices

APPENDIX A Mis-assembled regions of the rat genome excluded from the analysis.

Chromosome	Start (bp)	Stop (bp)
1	35,951,205	37,213,124
17	2,202,903	2,887,911
4	46,491,865	46,491,865
9	204,359	2,599,275
4	12,917,785	13,429,152
14	46,598,211	67,803,639
12	41,030,616	46,318,674
19	14,393,582	14,453,551

APPENDIX B Summary of the measures collected in the HS, and, for each, significant covariates, transformation, mapping method, logP or posterior probability threshold, number of animals for which genotypes, phenotype and covariates values were available, joint effect size of the QTLs, proportion of phenotypic variance of genetic of genetic origin (i.e. heritability), and proportion of genetic variance explained jointly by the QTLs. * highlights that the proportion of genetic variance explained by the QTLs is greater than 100%, which is not formally possible. This may happen when there is a QTL of very large effect: because the variance components were estimated under the null model, the covariance structure used to account for relatedness might be slightly incorrect. Proportions greater than 100% can also result from false positive associations.

Phenotype	**Measure**	**Covariates**	**Transformation**	**Mapping method**	**Threshold**	**No animals**	**Joint effect QTLs (%)**	**Genetic variance (%)**	**Proportion of genetic variance explained by QTLs (%)**
Anxiety (novel cage)	Distance 25' to 30'	sex,batch,is. albino	boxcox	Mixed models	4.2	1369	3.62	5.1	71
Anxiety (novel cage)	Distance first 5'	sex,batch,is. albino	boxcox	Mixed models	4.1	1367	4.54	20.3	22.4
Anxiety (shuttlebox)	Mean response latency	sex,batch,is. albino	boxcox	Mixed models	4.1	1366	5.13	28.3	18.1
Anxiety (shuttlebox)	Number of avoidances	sex,batch,is. albino		Resampling	0.4	1369	0	NA	NA
Anxiety (shuttlebox)	Number of crosses between trials	sex,batch,is. albino		Resampling	0.3	1369	1.6	NA	NA
Anxiety (shuttlebox)	Time spent freezing	sex,batch,is. albino	boxcox	Mixed models	4.3	937	0	17.5	0
Anxiety (zeromaze)	Latency to enter open section	sex,batch,is. albino	boxcox	Mixed models	4.1	1166	5.56	3.9	142.6*

Anxiety (zeromaze)	Number of entries in open section	sex,batch,is. albino		Resampling	0.7	1375	0	NA	NA
Anxiety (zeromaze)	Stretched attend postures	sex,batch,is. albino	boxcox	Mixed models	4.8	1370	0	14.5	0
Anxiety (zeromaze)	Time spent in open section	sex,batch,is. albino	boxcox	Mixed models	4.2	1161	3.86	21.3	18.1
Arterial elastic lamina rupture	Number of lesions in AA	sex		Resampling	0.7	559	10.9	NA	NA
Arterial elastic lamina rupture	Number of lesions in AA and IL	sex		Resampling	0.7	559	8.6	NA	NA
Arterial elastic lamina rupture	Number of lesions in IL	sex		Resampling	0.7	559	0	NA	NA
Arterial elastic lamina rupture	Score of lesions in AA	sex		Resampling	0.7	559	8.7	NA	NA
Arterial elastic lamina rupture	Score of lesions in AA and IL	sex		Resampling	0.7	559	6.7	NA	NA
Arterial elastic lamina rupture	Score of lesions in IL	sex		Resampling	0.7	559	0	NA	NA
Body weight	Body weight at day immunization	sex	boxcox	Mixed models	3.6	1226	34.43	73.6	46.8
Bone morphology	Axis length	sex,batch,BW_at_day9_pi,age		Mixed models	4.6	1172	0	17.2	0
Bone morphology	Bone head width	sex,batch,BW_at_day9_pi,age		Mixed models	4.1	1172	14.39	28.7	50.1

Bone morphology	Bone neck width	sex,batch,BW_at_day9_pi,age	Mixed models	4.2	1170	19.79	26.7	74.1
Bone morphology	Distal femur cortical area	sex,batch,BW_at_day9_pi,age	Mixed models	4.6	1170	5.02	34.8	14.4
Bone morphology	Distal femur cortical density	sex,batch,BW_at_day9_pi,age	Mixed models	5.8	1172	0	15.6	0
Bone morphology	Distal femur total area	sex,batch,BW_at_day9_pi,age	Mixed models	4.9	1172	0	16.6	0
Bone morphology	Distal femur total density	sex,batch,BW_at_day9_pi,age	Mixed models	3.9	1171	13.93	45.7	30.5
Bone morphology	Distal femur trabecular area	sex,batch,BW_at_day9_pi,age	Mixed models	4.9	1172	0	16.6	0
Bone morphology	Distal femur trabecular density	sex,batch,BW_at_day9_pi,age	Mixed models	4.7	1172	0	35.6	0
Bone morphology	Femoral neck cortical area	sex,batch,BW_at_day9_pi,age	Mixed models	4.7	1158	5.51	28.7	19.2
Bone morphology	Femoral neck cortical density	sex,batch,BW_at_day9_pi,age	Mixed models	5.4	1158	6.11	36.8	16.6
Bone morphology	Femoral neck elongation	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1075	3.33	18.3	18.2
Bone morphology	Femoral neck polar moment or inertia	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1144	12.06	19.7	61.2
Bone morphology	Femoral neck stiffness	sex,batch,BW_at_day9_pi,age	Mixed models	4.5	1075	2.88	17.8	16.2

Bone morphology	Femoral neck total area	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1158	10.69	27.4	39
Bone morphology	Femoral neck total density	sex,batch,BW_at_day9_pi,age	Mixed models	3.8	1158	27.45	61.3	44.8
Bone morphology	Femoral neck trabecular area	sex,batch,BW_at_day9_pi,age	Mixed models	3.9	1158	11.14	40.6	27.4
Bone morphology	Femoral neck trabecular density	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1158	4.42	47.2	9.4
Bone morphology	Femoral neck ultimate force	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1075	17.39	35.6	48.8
Bone morphology	Femoral neck work to failure	sex,batch,BW_at_day9_pi,age	Mixed models	4.1	1075	7.06	20.4	34.6
Bone morphology	Femur area	sex,batch,BW_at_day9_pi,age	Mixed models	3.9	1174	20.57	39.2	52.5
Bone morphology	Femur elongation	sex,batch,BW_at_day9_pi,age	Mixed models	4	1161	12.69	12.4	102.3*
Bone morphology	Femur midshaft cortical area	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1171	27.43	50.2	54.6
Bone morphology	Femur midshaft cortical density	sex,batch,BW_at_day9_pi,age	Mixed models	4.5	1171	25.49	49.3	51.7
Bone morphology	Femur midshaft polar moment of inertia	sex,batch,BW_at_day9_pi,age	Mixed models	4.1	1171	35.84	60.2	59.5

Bone morphology	Femur midshaft total area	sex,batch,BW_at_day9_pi,age	Mixed models	4.1	1171	40.38	62.7	64.4
Bone morphology	Femur midshaft total density	sex,batch,BW_at_day9_pi,age	Mixed models	4.4	1171	18.21	59.1	30.8
Bone morphology	Femur mineral content	sex,batch,BW_at_day9_pi,age	Mixed models	4.6	1174	21.24	56.7	37.5
Bone morphology	Femur mineral density	sex,batch,BW_at_day9_pi,age	Mixed models	5.1	1174	0	47	0
Bone morphology	Femur stiffness	sex,batch,BW_at_day9_pi,age	Mixed models	4.7	1168	0	19.5	0
Bone morphology	Femur ultimate force	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1168	3.52	22.3	15.8
Bone morphology	Femur work to failure	sex,batch,BW_at_day9_pi,age	Mixed models	4.1	1168	16.71	17.7	94.4
Bone morphology	Length to femoral head	sex,batch,BW_at_day9_pi,age	Mixed models	4.2	1172	14.57	52	28
Bone morphology	Length to trochanter	sex,batch,BW_at_day9_pi,age	Mixed models	4	1172	19.98	49.8	40.1
Bone morphology	Lumbar area	sex,batch,BW_at_day9_pi,age	Mixed models	4.3	1172	13.83	27.7	49.9
Bone morphology	Lumbar cortical area	sex,batch,BW_at_day9_pi,age	Mixed models	4.6	1173	0	33.7	0
Bone morphology	Lumbar mineral content	sex,batch,BW_at_day9_pi,age	Mixed models	4.8	1173	0	50.5	0

Bone morphology	Lumbar mineral density	sex,batch,BW_at_day9_pi,age		Mixed models	4.8	1173	7.08	51.2	13.8
Bone morphology	Lumbar total area	sex,batch,BW_at_day9_pi,age		Mixed models	4.4	1173	7.44	20.8	35.8
Bone morphology	Lumbar total density	sex,batch,BW_at_day9_pi,age		Mixed models	5	1173	0	35	0
Bone morphology	Lumbar trabecular area	sex,batch,BW_at_day9_pi,age		Mixed models	3.9	1173	11.63	17.2	67.6
Bone morphology	Lumbar trabecular density	sex,batch,BW_at_day9_pi,age		Mixed models	5.1	1173	4.46	33	13.5
Bone morphology	Lumber cortical density	sex,batch,BW_at_day9_pi,age		Mixed models	6.6	1173	4.7	25.1	18.7
Cardiovascular function	Blood pressure	sex,batch	boxcox	Mixed models	4.5	1386	0	13.2	0
Cardiovascular function	Heart weight	sex,batch,BW_at_day9_pi	boxcox	Mixed models	3.8	1174	16.7	35.1	47.6
Coat colour	Is albino			Resampling	NA	1407	97.5	NA	NA
Coat colour	Is dark brown			Resampling	NA	1407	84.8	NA	NA
Coat colour	Is light brown			Resampling	NA	1407	82.2	NA	NA
Coat colour	Is spotted			Resampling	NA	1407	68.2	NA	NA
Glucose tolerance	Area under glycemia curve	sex,batch,test_worked	boxcox	Mixed models	4.4	931	2.61	38.3	6.8
Glucose tolerance	Area under glycemia curve over baseline	sex,batch,test_worked	boxcox	Mixed models	4.2	938	15.39	27.8	55.4

Glucose tolerance	Glycemia 120' after injection	sex,batch,test_worked	boxcox	Mixed models	4.5	934	11.82	32.5	36.4
Glucose tolerance	Glycemia 30' after injection	sex,batch,test_worked	boxcox	Mixed models	4.2	930	0	31.9	0
Glucose tolerance	Glycemia 60' after injection	sex,batch,test_worked	boxcox	Mixed models	4.4	933	7.03	31.5	22.3
Glucose tolerance	Glycemia before injection	sex,batch,test_worked	boxcox	Mixed models	4.6	936	5.17	26.2	19.7
Hematology	Calculated mean cell haemoglobin concentration	sex,batch	boxcox	Mixed models	5.6	1302	0	40.5	0
Hematology	Hematocrit	sex,batch	boxcox	Mixed models	4.2	1295	18.21	33.5	54.4
Hematology	Hemoglobin	sex,batch	boxcox	Mixed models	4.5	1298	17.73	38.9	45.6
Hematology	Hemoglobin distribution width	sex,batch	boxcox	Mixed models	4.4	1307	36.49	55.4	65.9
Hematology	Lob-index	sex,batch	boxcox	Mixed models	4.9	1313	4.26	28.6	14.9
Hematology	Mean corpuscular hemoglobin	sex,batch	boxcox	Mixed models	4	1304	22.16	59.2	37.4
Hematology	Mean corpuscular volume	sex,batch	boxcox	Mixed models	4.6	1311	14.23	68.1	20.9
Hematology	Mean peroxidase index	sex,batch	boxcox	Mixed models	6.1	1297	0	39.8	0
Hematology	Mean platelet component	sex,batch	boxcox	Mixed models	4.8	1293	19.12	38.8	49.3
Hematology	Mean platelet mass	sex,batch	boxcox	Mixed models	4.8	1279	47.21	57.7	81.8
Hematology	Mean platelet volume	sex,batch	boxcox	Mixed models	4.8	1294	11.83	26.8	44.1

Hematology	Measured mean cell hemoglobin concentration	sex,batch	boxcox	Mixed models	5.5	1301	10.25	36.2	28.3
Hematology	Platelet clumps	sex,batch	boxcox	Mixed models	4.3	1229	0	NA	NA
Hematology	Platelet component distribution width	sex,batch	boxcox	Mixed models	5.3	1299	0	38.8	0
Hematology	Platelet count	sex,batch	boxcox	Mixed models	4.3	1312	22.82	41.1	55.5
Hematology	Platelet distribution width	sex,batch	boxcox	Mixed models	5.1	1304	6.32	37.7	16.8
Hematology	Plateletcrit	sex,batch,BW_at_IPGTT	boxcox	Mixed models	4.8	1309	21.58	49.3	43.8
Hematology	Proportion of basophils in WBC	sex,batch	boxcox	Mixed models	5.5	1306	0	18.3	0
Hematology	Proportion of eosinophils in WBC	sex,batch	boxcox	Mixed models	4.3	1313	7.44	41.9	17.8
Hematology	Proportion of large unstained cells in WBC	sex,batch	boxcox	Mixed models	5.6	1301	0	18.7	0
Hematology	Proportion of lymphocytes in WBC	sex,batch	boxcox	Mixed models	4.4	1302	7.67	41.3	18.6
Hematology	Proportion of monocytes in WBC	sex,batch	boxcox	Mixed models	4.7	1313	17.95	45.8	39.2
Hematology	Proportion of neutrophils in WBC	sex,batch	boxcox	Mixed models	4.2	1310	9.91	43.4	22.8

Hematology	Red blood cell count	sex,batch	boxcox	Mixed models	3.8	1293	14.59	38.5	37.9
Hematology	Red blood cell distribution width	sex,batch	boxcox	Mixed models	4.4	1300	26.72	55.2	48.4
Hematology	WBC in the Baso channel	sex,batch	boxcox	Mixed models	4.2	1313	13.4	52.9	25.3
Immunology	Absolute B cells	sex	quantile normalization by batch	Mixed models	4.2	744	21.44	50.9	42.1
Immunology	Absolute CD25+CD4+ cells	sex	quantile normalization by batch	Mixed models	4.2	727	14.47	40.3	35.9
Immunology	Absolute CD25+CD8+ cells	sex	quantile normalization by batch	Mixed models	4.2	726	10.9	25.7	42.4
Immunology	Absolute CD4+ T cells	sex	quantile normalization by batch	Mixed models	4.4	1168	3.69	57.5	6.4
Immunology	Absolute CD8+ T cells	sex	quantile normalization by batch	Mixed models	4.1	1166	19.87	59.5	33.4
Immunology	Absolute T cells	sex	quantile normalization by batch	Mixed models	4.3	1187	14.86	59	25.2
Immunology	Expression of CD25 on C84+ cells	sex	quantile normalization by batch	Mixed models	4.1	1024	4.05	17.5	23.1
Immunology	Expression of CD25 on CD4+ cells	sex	quantile normalization by batch	Mixed models	4.3	1207	5.59	24	23.3
Immunology	Expression of CD28 on T cells	sex	quantile normalization by batch	Mixed models	4.2	900	0	11.4	0

Immunology	Expression of CD45 in CD4+ cells	sex	quantile normalization by batch	Mixed models	4.2	541	17.92	36.1	49.6
Immunology	Expression of CD45 in CD8+ cells	sex	quantile normalization by batch	Mixed models	4.2	540	28.7	32	89.7
Immunology	Expression of RT1A on granulocytes	sex	quantile normalization by batch	Mixed models	4.2	1082	29.2	41.8	69.9
Immunology	Expression on RT1B on B cells	sex	quantile normalization by batch	Mixed models	4.2	813	29.62	50.1	59.1
Immunology	Proportion of B cells in in WBC	sex	quantile normalization by batch	Mixed models	4.2	1231	33.21	51.2	64.9
Immunology	Proportion of CD4- CD8- T cells	sex	quantile normalization by batch	Mixed models	4.2	769	17.4	41.9	41.5
Immunology	Proportion of CD4+ cells expressing CD25	sex	quantile normalization by batch	Mixed models	4.3	1198	11.13	46.3	24
Immunology	Proportion of CD4+ cells in T cells	sex	quantile normalization by batch	Mixed models	4.2	1255	19	55.1	34.5
Immunology	Proportion of CD4+ cells with expressing CD45RC	sex	quantile normalization by batch	Mixed models	4.1	542	20.04	27.2	73.7
Immunology	Proportion of CD4+ cells with high expression of CD25	sex	quantile normalization by batch	Mixed models	4.2	569	20.52	63.3	32.4

Immunology	Proportion of CD4+ cells with high expression of CD45RC	sex	quantile normalization by batch	Mixed models	4.2	535	12.56	49	25.6
Immunology	Proportion of CD4+ cells with low expression of CD45RC	sex	quantile normalization by batch	Mixed models	4.2	539	21.26	43.2	49.2
Immunology	Proportion of CD4+ cells with not expressing CD45RC	sex	quantile normalization by batch	Mixed models	4.1	186	39.59	63.2	62.6
Immunology	Proportion of CD4+CD8+ T cells	sex	quantile normalization by batch	Mixed models	4.2	185	19.05	40.4	47.2
Immunology	Proportion of CD8+ cells expressing CD25	sex	quantile normalization by batch	Mixed models	4.2	1147	16.05	25.5	62.9
Immunology	Proportion of CD8+ cells in T cells	sex	quantile normalization by batch	Mixed models	4.2	1256	32.28	56	57.6
Immunology	Proportion of CD8+ cells with expressing of CD45RC	sex	quantile normalization by batch	Mixed models	4.2	540	16.57	39.1	42.4
Immunology	Proportion of CD8+ cells with high expression of CD25	sex	quantile normalization by batch	Mixed models	4	183	36.93	23.5	157.1*

Immunology	Proportion of CD8+ cells with high expression of CD45RC	sex	quantile normalization by batch	Mixed models	4.1	538	0	42.8	0
Immunology	Proportion of CD8+ cells with low expression of CD45RC	sex	quantile normalization by batch	Mixed models	4.2	535	0	46.8	0
Immunology	Proportion of CD8+ cells with not expressing of CD45RC	sex	quantile normalization by batch	Mixed models	4.1	187	25.84	51.1	50.6
Immunology	Proportion of T cells expressing RT1B	sex	quantile normalization by batch	Mixed models	4.1	393	9.67	35.5	27.2
Immunology	Proportion of T cells in WBC	sex	quantile normalization by batch	Mixed models	4.1	1029	19.31	57	33.9
Immunology	Ratio of CD4+cells to CD8+ cells	sex	quantile normalization by batch	Mixed models	4.2	1248	23.3	58.6	39.8
Immunology	Ratio of T cells to B cells	sex	quantile normalization by batch	Mixed models	4.2	1021	28.67	56.8	50.5
Induced neuroinflammation	Balance disturbance	sex,batch		Resampling	NA	1401	0	NA	NA
Induced neuroinflammation	Cumulative scores	sex,batch	boxcox	Mixed models	4.3	457	0	21.5	0
Induced neuroinflammation	Died	sex,batch		Resampling	NA	1405	0	NA	NA

Induced neuroinflammation	Duration	sex,batch	boxcox	Mixed models	4.2	457	0	16.9	0
Induced neuroinflammation	First day with score>=1	sex,batch		Resampling	NA	457	0	NA	NA
Induced neuroinflammation	Lowest weight	sex,batch	boxcox	Mixed models	3.6	1383	31.35	44.9	69.8
Induced neuroinflammation	Maximum score	sex,batch	boxcox	Mixed models	4.2	457	0	21.8	0
Induced neuroinflammation	Symptoms for at least one day	sex,batch		Resampling	NA	1405	0	NA	NA
Induced neuroinflammation	Symptoms for at least two days	sex,batch		Resampling	NA	1405	0	NA	NA
Induced neuroinflammation	Weight loss compared to day 0	sex,batch	boxcox	Mixed models	4.2	1197	20.25	24.6	82.3
Induced neuroinflammation	Weight loss compared to day 9	sex,batch	boxcox	Mixed models	4.3	1357	19.14	18.7	102.4*
Renal agenesis	Renal agenesis	sex		Resampling	NA	1407	0	NA	NA
Serum biochemistry	Alanine aminotransferase	sex,batch,Haemalysis	boxcox	Mixed models	4.1	1157	0	29.6	0
Serum biochemistry	Alkaline phosphatase	sex,batch	boxcox	Mixed models	4.2	1365	24.93	48.1	51.8
Serum biochemistry	Aspartate aminotransferase	sex,batch,BW_at_day28_pi,Haemalysis	boxcox	Mixed models	4.4	1126	0	28.7	0
Serum biochemistry	Basal glycemia	sex,batch	boxcox	Mixed models	4.6	1359	4.7	15.6	30.1

Serum biochemistry	Calcium	sex,batch	boxcox	Mixed models	5.3	1358	0	21.2	0
Serum biochemistry	Chloride	sex,batch	boxcox	Mixed models	4.7	1351	7.28	28.2	25.8
Serum biochemistry	Creatinine	sex,batch	boxcox	Mixed models	4.1	1346	0	36.7	0
Serum biochemistry	HDL	sex,batch	boxcox	Mixed models	4.4	1355	12.17	38.3	31.8
Serum biochemistry	Iron	sex,batch	boxcox	Mixed models	3.7	1359	14.82	19	78
Serum biochemistry	LDL	sex,batch	boxcox	Mixed models	4.4	1347	2.28	36	6.3
Serum biochemistry	Potassium	sex,batch,Haemalysis	boxcox	Mixed models	4.7	1034	2.86	41.7	6.9
Serum biochemistry	Sodium	sex,batch	boxcox	Mixed models	4.4	1348	2.39	27.6	8.7
Serum biochemistry	Total cholesterol	sex,batch,Haemalysis	boxcox	Mixed models	4.4	1194	12.4	40.9	30.3
Serum biochemistry	Triglycerides	sex,batch,BW_at_day28_pi	boxcox	Mixed models	4.3	1295	6.76	26.1	25.9
Serum biochemistry	Urea	sex,batch,BW_at_day28_pi,Haemalysis	boxcox	Mixed models	4.2	1124	4.29	24.2	17.7
Wound healing	Hole area	sex,batch,reliable	boxcox	Mixed models	4.6	1054	5.79	14.6	39.7

APPENDIX C Summary of the QTLs detected in the HS, the mapping method used, effect size, and whether or not there are candidate variants (as defined by merge analysis) at the QTL.

Phenotype	**Measure**	**Mapping method**	**Chr**	**Start QTL (bp)**	**Stop QTL (bp)**	**logP/RMIP**	**Effect size**	**Has candidate variant**
Anxiety (novel cage)	Distance 25' to 30'	Mixed models	2	212,339,307	216,428,065	5.6	3.6	no
Anxiety (novel cage)	Distance first 5'	Mixed models	7	40,970,110	45,621,360	4.1	4.5	yes
Anxiety (shuttlebox)	Mean response latency	Mixed models	2	80,228,740	84,832,232	4.3	5.1	yes
Anxiety (zeromaze)	Latency to enter open section	Mixed models	5	148,032,372	152,573,310	4.4	2.6	no
Anxiety (zeromaze)	Latency to enter open section	Mixed models	20	30,779,805	35,105,807	5	3	no
Anxiety (zeromaze)	Time spent in open section	Mixed models	8	90,360,975	94,877,907	4.5	3.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Body weight	Body weight at day immunization	Mixed models	2	129,257,158	133,732,902	4.6	6.1	no
Body weight	Body weight at day immunization	Mixed models	2	138,471,008	142,614,192	5.4	9.7	no
Body weight	Body weight at day immunization	Mixed models	2	216,921,362	220,119,428	7.9	7.4	yes
Body weight	Body weight at day immunization	Mixed models	3	23,637,280	36,416,292	5.2	1.9	no
Body weight	Body weight at day immunization	Mixed models	4	27,024,851	31,510,205	4.6	0.9	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Body weight	Body weight at day immunization	Mixed models	4	103,453,924	108,048,784	4.3	3.5	no
Body weight	Body weight at day immunization	Mixed models	8	81,857,162	86,244,836	4.8	4.6	no
Body weight	Body weight at day immunization	Mixed models	9	69,979,174	74,467,926	4.6	8.1	no
Body weight	Body weight at day immunization	Mixed models	12	7,483,933	12,147,931	4.1	9.1	yes
Body weight	Body weight at day immunization	Mixed models	16	85,599,698	90,217,239	4	1.8	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Bone head width	Mixed models	2	115,053,708	119,671,384	4.2	4.1	yes
Bone morphology	Bone head width	Mixed models	7	110,032,878	113,889,214	6.2	8.3	yes
Bone morphology	Bone head width	Mixed models	8	83,497,817	87,206,739	6.6	3.1	no
Bone morphology	Bone head width	Mixed models	X	22,737,140	27,326,052	4.3	4	no
Bone morphology	Bone neck width	Mixed models	1	156,268,618	160,898,972	4.2	3.4	yes
Bone morphology	Bone neck width	Mixed models	1	248,786,924	253,414,546	4.2	3.6	no
Bone morphology	Bone neck width	Mixed models	3	12,017,175	20,061,775	9.8	6.6	no
Bone morphology	Bone neck width	Mixed models	4	53,447,006	57,992,998	4.4	3.8	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Bone neck width	Mixed models	15	107,346,994	109,746,510	4.8	3.8	no
Bone morphology	Bone neck width	Mixed models	17	87,918,971	90,857,457	8.6	6	no
Bone morphology	Bone neck width	Mixed models	20	13,497,072	17,306,136	6.3	3.7	no
Bone morphology	Distal femur cortical area	Mixed models	13	86,052,287	90,183,387	5.5	5	no
Bone morphology	Distal femur total density	Mixed models	2	152,744,366	157,219,858	4.6	7.6	yes
Bone morphology	Distal femur total density	Mixed models	5	48,773,007	53,141,595	4.9	2.2	no
Bone morphology	Distal femur total density	Mixed models	10	102,313,675	107,052,207	3.9	5.9	yes
Bone morphology	Distal femur total density	Mixed models	X	44,129,745	48,579,177	4.7	4.3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femoral neck cortical area	Mixed models	10	88,815,649	93,165,763	4.9	5.5	no
Bone morphology	Femoral neck cortical density	Mixed models	2	180,061,094	184,080,322	5.8	6.1	no
Bone morphology	Femoral neck elongation	Mixed models	3	119,597,501	124,127,543	4.4	3.3	no
Bone morphology	Femoral neck polar moment or inertia	Mixed models	5	4,407,737	8,957,033	4.4	4.1	yes
Bone morphology	Femoral neck polar moment or inertia	Mixed models	7	43,710,704	47,559,914	6.2	4.3	no
Bone morphology	Femoral neck polar moment or inertia	Mixed models	19	30,390,841	34,962,031	4.3	4.5	no
Bone morphology	Femoral neck stiffness	Mixed models	14	14,061,617	18,524,251	4.6	2.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femoral neck total area	Mixed models	3	42,197,070	46,571,334	4.9	3.3	no
Bone morphology	Femoral neck total area	Mixed models	5	4,470,805	8,893,965	4.7	5	yes
Bone morphology	Femoral neck total area	Mixed models	5	155,941,615	160,405,793	4.6	2.6	no
Bone morphology	Femoral neck total area	Mixed models	6	66,405,253	70,766,805	4.9	0.8	no
Bone morphology	Femoral neck total density	Mixed models	1	170,051,092	174,654,246	4.3	7.3	yes
Bone morphology	Femoral neck total density	Mixed models	5	4,025,968	8,215,792	5.3	6.4	yes
Bone morphology	Femoral neck total density	Mixed models	6	130,159,342	134,693,746	4.4	5.1	no
Bone morphology	Femoral neck total density	Mixed models	10	102,482,740	107,052,600	4.3	11.8	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femoral neck total density	Mixed models	17	30,797,494	35,470,790	4.1	5.1	no
Bone morphology	Femoral neck trabecular area	Mixed models	3	42,048,824	46,719,580	4.1	3.1	no
Bone morphology	Femoral neck trabecular area	Mixed models	5	4,191,288	8,050,472	6.2	6.2	no
Bone morphology	Femoral neck trabecular area	Mixed models	8	98,845,406	103,488,914	4.2	3	yes
Bone morphology	Femoral neck trabecular density	Mixed models	1	102,219,485	106,760,825	4.4	4.4	no
Bone morphology	Femoral neck ultimate force	Mixed models	4	95,683,000	100,197,714	4.5	4.5	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femoral neck ultimate force	Mixed models	8	90,568,443	95,016,767	4.7	5.4	yes
Bone morphology	Femoral neck ultimate force	Mixed models	10	44,645,600	48,415,584	6.4	7.7	no
Bone morphology	Femoral neck ultimate force	Mixed models	18	75,068,266	79,591,914	4.5	2.6	no
Bone morphology	Femoral neck work to failure	Mixed models	10	82,396,545	86,984,557	4.3	3.2	no
Bone morphology	Femoral neck work to failure	Mixed models	13	42,220,830	46,729,966	4.5	4.1	no
Bone morphology	Femur area	Mixed models	1	146,197,892	150,706,600	4.5	2.3	no
Bone morphology	Femur area	Mixed models	5	111,007,492	114,896,042	6.1	7.4	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur area	Mixed models	5	155,847,821	165,889,789	5.4	4.7	no
Bone morphology	Femur area	Mixed models	5	170,584,825	173,068,538	5.6	6	no
Bone morphology	Femur area	Mixed models	14	9,185,966	13,550,120	4.9	5.7	no
Bone morphology	Femur elongation	Mixed models	8	36,788,399	41,370,391	4.3	2.7	no
Bone morphology	Femur elongation	Mixed models	8	121,095,076	125,698,968	4.3	3.6	no
Bone morphology	Femur elongation	Mixed models	11	40,014,789	43,688,643	6.7	3.6	no
Bone morphology	Femur elongation	Mixed models	13	57,972,944	62,645,986	4.1	4.5	yes
Bone morphology	Femur midshaft cortical area	Mixed models	3	19,821,238	24,117,082	5.1	7.6	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur midshaft cortical area	Mixed models	4	90,902,206	95,434,364	4.4	7.7	no
Bone morphology	Femur midshaft cortical area	Mixed models	5	170,253,190	173,068,538	4.4	7.4	no
Bone morphology	Femur midshaft cortical area	Mixed models	10	91,203,990	95,157,124	5.9	13.7	no
Bone morphology	Femur midshaft cortical area	Mixed models	10	102,860,493	107,383,881	4.5	4.5	yes
Bone morphology	Femur midshaft cortical area	Mixed models	16	61,718,203	66,271,523	4.4	4.4	no
Bone morphology	Femur midshaft cortical density	Mixed models	2	177,250,896	180,611,046	7.5	2.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur midshaft cortical density	Mixed models	6	38,241,510	41,517,214	7.7	3.8	yes
Bone morphology	Femur midshaft cortical density	Mixed models	6	93,019,628	97,145,216	5.5	7	yes
Bone morphology	Femur midshaft cortical density	Mixed models	9	42,346,982	45,947,194	6.8	12	no
Bone morphology	Femur midshaft cortical density	Mixed models	20	1	2,191,953	4.9	6.4	yes
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	1	201,265,878	205,189,952	6	4.8	no
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	2	134,927,504	138,889,644	5.9	8.8	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	4	96,051,838	106,296,290	6.4	6.8	no
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	5	170,033,053	173,068,538	6	12.7	no
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	10	90,877,745	95,483,369	4.3	16.1	no
Bone morphology	Femur midshaft polar moment of inertia	Mixed models	11	30,599,269	35,105,331	4.5	9.7	no
Bone morphology	Femur midshaft total area	Mixed models	1	167,340,781	171,991,511	4.1	2.2	no
Bone morphology	Femur midshaft total area	Mixed models	2	43,959,565	48,572,613	4.2	11.2	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur midshaft total area	Mixed models	4	95,732,433	100,148,281	4.7	7.1	no
Bone morphology	Femur midshaft total area	Mixed models	5	170,474,673	173,068,538	5.5	12.5	no
Bone morphology	Femur midshaft total area	Mixed models	7	77,077,466	81,146,584	5.6	11	no
Bone morphology	Femur midshaft total area	Mixed models	10	44,399,078	48,662,106	5.1	5.8	no
Bone morphology	Femur midshaft total area	Mixed models	10	91,151,727	95,209,387	5.7	17.4	no
Bone morphology	Femur midshaft total area	Mixed models	11	30,704,307	35,000,293	5.1	10.6	no
Bone morphology	Femur midshaft total density	Mixed models	10	57,899,832	62,370,770	4.6	8.4	yes
Bone morphology	Femur midshaft total density	Mixed models	10	99,708,728	110,711,601	5.8	3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur midshaft total density	Mixed models	13	79,211,931	83,557,081	4.9	3.5	no
Bone morphology	Femur midshaft total density	Mixed models	14	8,690,526	12,962,922	5.1	4.5	no
Bone morphology	Femur mineral content	Mixed models	5	110,272,347	114,727,787	4.6	7.1	no
Bone morphology	Femur mineral content	Mixed models	10	44,497,288	48,563,896	5.6	5.7	no
Bone morphology	Femur mineral content	Mixed models	10	90,981,458	95,379,656	4.8	11.7	no
Bone morphology	Femur mineral content	Mixed models	18	78,288,032	82,452,056	5.4	6	no
Bone morphology	Femur ultimate force	Mixed models	18	13,097,598	17,667,068	4.3	3.5	no
Bone morphology	Femur work to failure	Mixed models	2	25,790,756	30,395,378	4.3	3.6	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Femur work to failure	Mixed models	5	130,081,145	133,799,979	6.5	6.1	yes
Bone morphology	Femur work to failure	Mixed models	8	21,568,703	26,173,297	4.3	4.3	yes
Bone morphology	Femur work to failure	Mixed models	10	95,120,025	99,351,693	5.2	5.6	no
Bone morphology	Length to femoral head	Mixed models	1	18,435,975	22,845,341	4.8	6.1	yes
Bone morphology	Length to femoral head	Mixed models	10	44,432,287	48,628,897	5.3	3.9	no
Bone morphology	Length to femoral head	Mixed models	10	90,938,668	95,422,446	4.6	1.8	no
Bone morphology	Length to femoral head	Mixed models	18	78,266,911	82,473,177	5.3	4	no
Bone morphology	Length to trochanter	Mixed models	2	136,110,601	140,702,799	4.3	4.4	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Length to trochanter	Mixed models	5	110,199,867	114,800,267	4.3	5.2	no
Bone morphology	Length to trochanter	Mixed models	6	53,206,065	57,863,995	4.1	2	yes
Bone morphology	Length to trochanter	Mixed models	10	44,646,835	48,414,349	6.4	4.7	no
Bone morphology	Length to trochanter	Mixed models	18	78,379,374	82,360,714	5.9	3.9	no
Bone morphology	Length to trochanter	Mixed models	20	32,016,012	35,665,316	6.7	4.3	no
Bone morphology	Lumbar area	Mixed models	1	231,016,806	235,372,388	4.9	3.2	no
Bone morphology	Lumbar area	Mixed models	3	1,845,276	5,723,902	6.1	3.7	no
Bone morphology	Lumbar area	Mixed models	9	96,397,323	100,737,125	4.9	3.3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Lumbar area	Mixed models	12	21,964,772	22,405,748	16.4	6.9	no
Bone morphology	Lumbar mineral density	Mixed models	1	195,565,176	199,903,482	4.9	7.1	yes
Bone morphology	Lumbar total area	Mixed models	1	86,292,375	90,853,025	4.4	4	yes
Bone morphology	Lumbar total area	Mixed models	2	18,349,590	22,878,736	4.5	2.5	no
Bone morphology	Lumbar total area	Mixed models	3	25,832,944	30,156,274	5	3.3	no
Bone morphology	Lumbar trabecular area	Mixed models	1	86,300,178	90,845,222	4.4	3.6	yes
Bone morphology	Lumbar trabecular area	Mixed models	2	146,448,743	151,112,383	4.1	2.3	no
Bone morphology	Lumbar trabecular area	Mixed models	8	1,963,299	6,686,095	4	3.6	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Bone morphology	Lumbar trabecular area	Mixed models	20	21,101,004	25,751,416	4.1	4.5	yes
Bone morphology	Lumbar trabecular density	Mixed models	13	1	1,898,715	7	4.5	no
Bone morphology	Lumber cortical density	Mixed models	15	58,985,183	62,528,673	7	4.7	no
Cardiovascular function	Heart weight	Mixed models	1	202,149,330	206,633,814	4.6	5.5	yes
Cardiovascular function	Heart weight	Mixed models	5	142,009,210	146,713,074	4	2.5	yes
Cardiovascular function	Heart weight	Mixed models	10	37,243,161	47,929,176	3.9	6.4	no
Cardiovascular function	Heart weight	Mixed models	13	56,470,137	67,956,785	4.3	3.6	no
Glucose tolerance	Area under glycemia curve	Mixed models	2	202,127,341	206,658,781	4.4	2.6	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Glucose tolerance	Area under glycemia curve over baseline	Mixed models	2	80,496,696	85,105,628	4.2	5	yes
Glucose tolerance	Area under glycemia curve over baseline	Mixed models	2	192,963,657	197,506,909	4.4	3.9	no
Glucose tolerance	Area under glycemia curve over baseline	Mixed models	3	147,102,806	151,528,632	4.7	4.2	no
Glucose tolerance	Area under glycemia curve over baseline	Mixed models	10	9,614,095	14,031,271	4.7	4.4	no
Glucose tolerance	Glycemia 120' after injection	Mixed models	2	157,476,106	161,666,442	5.3	8.6	yes
Glucose tolerance	Glycemia 120' after injection	Mixed models	15	39,128,803	43,543,589	4.7	4.1	no
Glucose tolerance	Glycemia 30' after injection	Mixed models	X	146,832,363	151,273,313	4.7	2.3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Glucose tolerance	Glycemia 60' after injection	Mixed models	2	153,823,195	158,123,335	5	7	yes
Glucose tolerance	Glycemia before injection	Mixed models	8	37,037,965	41,485,455	4.7	5.2	no
Hematology	Hematocrit	Mixed models	4	90,275,481	94,601,809	5	3.4	no
Hematology	Hematocrit	Mixed models	12	1,784,579	5,783,623	5.8	8.6	yes
Hematology	Hematocrit	Mixed models	15	51,629,414	56,199,828	4.3	3.2	no
Hematology	Hematocrit	Mixed models	19	30,223,924	34,580,406	4.9	5.9	no
Hematology	Hemoglobin	Mixed models	2	129,590,265	133,399,795	6.3	8.6	no
Hematology	Hemoglobin	Mixed models	4	90,214,860	94,662,430	4.7	3.8	no
Hematology	Hemoglobin	Mixed models	12	1,620,681	5,770,271	5.4	8.4	yes
Hematology	Hemoglobin distribution width	Mixed models	4	8,291,182	23,204,902	16.7	29.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Hemoglobin distribution width	Mixed models	9	53,957,135	58,368,897	4.8	12.1	no
Hematology	Hemoglobin distribution width	Mixed models	20	2,167,452	6,165,694	5.8	14.1	yes
Hematology	Hemoglobin distribution width	Mixed models	X	34,918,379	39,247,311	5	4.5	no
Hematology	Hemoglobin distribution width	Mixed models	X	142,501,628	146,962,986	4.6	2.3	no
Hematology	Lob-index	Mixed models	19	19,529,372	23,497,972	5.9	4.3	no
Hematology	Mean corpuscular hemoglobin	Mixed models	2	125,834,928	130,354,342	4.5	2.2	no
Hematology	Mean corpuscular hemoglobin	Mixed models	8	99,031,005	103,662,303	4.2	3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Mean corpuscular hemoglobin	Mixed models	10	82,301,560	85,523,486	7.8	6.1	yes
Hematology	Mean corpuscular hemoglobin	Mixed models	10	89,601,242	94,189,260	4.3	1.5	yes
Hematology	Mean corpuscular hemoglobin	Mixed models	11	67,954,200	72,087,118	5.5	1	yes
Hematology	Mean corpuscular hemoglobin	Mixed models	12	1	2,804,092	4.5	3.6	yes
Hematology	Mean corpuscular hemoglobin	Mixed models	19	48,847,073	54,614,907	13.1	8.5	yes
Hematology	Mean corpuscular volume	Mixed models	10	82,091,228	86,077,586	5.9	5.8	yes
Hematology	Mean corpuscular volume	Mixed models	14	76,227,760	80,568,850	4.9	3.1	yes
Hematology	Mean corpuscular volume	Mixed models	19	53,114,463	55,795,689	9.2	5.4	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Mean platelet component	Mixed models	6	118,279,200	122,224,918	6	11.8	no
Hematology	Mean platelet component	Mixed models	9	53,314,769	85,775,988	11.6	7.1	yes
Hematology	Mean platelet component	Mixed models	11	18,650,912	22,570,132	6	4.7	no
Hematology	Mean platelet mass	Mixed models	1	193,981,495	197,882,733	6.1	13	yes
Hematology	Mean platelet mass	Mixed models	2	209,009,239	212,910,797	6.1	11.6	no
Hematology	Mean platelet mass	Mixed models	2	225,193,752	228,439,362	7.8	14.1	yes
Hematology	Mean platelet mass	Mixed models	3	167,833,969	170,027,449	10.5	9.9	no
Hematology	Mean platelet mass	Mixed models	4	79,734,074	84,066,312	5	15.5	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Mean platelet mass	Mixed models	5	141,258,770	145,549,974	5.1	3.5	no
Hematology	Mean platelet mass	Mixed models	9	52,528,387	88,106,078	37.3	18.1	yes
Hematology	Mean platelet mass	Mixed models	13	67,222,863	71,400,623	5.4	9.7	no
Hematology	Mean platelet mass	Mixed models	17	17,143,009	21,475,265	5	2.3	no
Hematology	Mean platelet mass	Mixed models	18	60,426,319	63,987,697	6.9	6.6	no
Hematology	Mean platelet volume	Mixed models	3	26,155,975	30,373,359	5.3	6.7	no
Hematology	Mean platelet volume	Mixed models	9	70,162,857	74,284,243	5.5	2.3	no
Hematology	Mean platelet volume	Mixed models	11	18,945,695	22,650,879	6.6	5.3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Measured mean cell hemoglobin concentration	Mixed models	4	15,564,048	19,396,454	6.2	6.5	yes
Hematology	Measured mean cell hemoglobin concentration	Mixed models	6	60,005,731	63,114,357	8.1	5.3	no
Hematology	Platelet clumps	Mixed models	4	156,706,146	169,444,576	9.3	4.6	yes
Hematology	Platelet clumps	Mixed models	8	100,571,280	104,808,826	5.2	2.9	yes
Hematology	Platelet clumps	Mixed models	11	16,078,020	20,444,072	4.9	2.8	no
Hematology	Platelet count	Mixed models	9	53,073,639	76,583,611	16.3	18.1	yes
Hematology	Platelet count	Mixed models	11	14,468,503	18,540,417	5.6	6.8	yes
Hematology	Platelet distribution width	Mixed models	9	65,912,428	78,126,163	9	6.3	yes
Hematology	Plateletcrit	Mixed models	3	164,288,458	168,546,234	5.2	2.8	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Plateletcrit	Mixed models	9	49,807,586	76,486,574	14.4	15.3	yes
Hematology	Plateletcrit	Mixed models	12	1	7,471,179	5.5	6.8	yes
Hematology	Proportion of eosinophils in WBC	Mixed models	3	59,785,304	64,243,322	4.6	7.4	no
Hematology	Proportion of lymphocytes in WBC	Mixed models	1	125,570,566	129,321,970	6.5	5.1	no
Hematology	Proportion of lymphocytes in WBC	Mixed models	10	13,512,372	17,998,460	4.6	3.4	no
Hematology	Proportion of monocytes in WBC	Mixed models	1	250,374,319	254,359,131	5.9	8.5	yes
Hematology	Proportion of monocytes in WBC	Mixed models	8	119,864,424	128,916,092	11.2	10.4	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Proportion of neutrophils in WBC	Mixed models	1	125,309,032	129,583,504	5.1	5	no
Hematology	Proportion of neutrophils in WBC	Mixed models	10	13,612,555	17,898,277	5.1	4.6	no
Hematology	Proportion of neutrophils in WBC	Mixed models	20	21,873,671	26,431,969	4.4	1.6	no
Hematology	Red blood cell count	Mixed models	1	150,626,476	159,921,500	4.4	2.7	yes
Hematology	Red blood cell count	Mixed models	4	90,149,695	94,727,595	4.3	2.4	no
Hematology	Red blood cell count	Mixed models	6	42,210,942	46,901,266	4	4.8	no
Hematology	Red blood cell count	Mixed models	10	46,386,104	51,124,698	3.9	4	no
Hematology	Red blood cell count	Mixed models	15	51,419,143	56,060,333	4.2	3	no
Hematology	Red blood cell distribution width	Mixed models	2	128,833,347	133,361,091	4.5	5.1	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	Red blood cell distribution width	Mixed models	3	79,566,239	83,877,249	5	9.3	no
Hematology	Red blood cell distribution width	Mixed models	4	112,857,035	116,649,807	6.4	4.3	no
Hematology	Red blood cell distribution width	Mixed models	9	53,929,869	58,396,163	4.6	9.3	yes
Hematology	Red blood cell distribution width	Mixed models	15	69,784,862	74,234,324	4.7	2.4	no
Hematology	WBC in the Baso channel	Mixed models	3	43,848,050	48,132,714	5.1	6.6	no
Hematology	WBC in the Baso channel	Mixed models	3	93,038,327	96,683,919	6.7	3.5	no
Hematology	WBC in the Baso channel	Mixed models	9	89,577,151	94,038,951	4.6	4.6	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Hematology	WBC in the Baso channel	Mixed models	X	33,909,560	38,520,228	4.2	4	no
Immunology	Absolute B cells	Mixed models	3	74,061,018	78,656,594	4.3	7.7	yes
Immunology	Absolute B cells	Mixed models	4	82,080,115	86,242,491	5.4	4.1	no
Immunology	Absolute B cells	Mixed models	4	107,695,034	112,288,522	4.3	3.8	no
Immunology	Absolute B cells	Mixed models	9	89,591,881	93,842,847	5.2	4.7	yes
Immunology	Absolute B cells	Mixed models	10	70,039,087	74,323,019	5.1	6.3	no
Immunology	Absolute CD25+CD4+ cells	Mixed models	19	50,708,817	54,963,031	5.2	14.5	yes
Immunology	Absolute CD25+CD8+ cells	Mixed models	20	1,813,977	5,984,767	5.4	10.9	yes
Immunology	Absolute CD4+ T cells	Mixed models	3	8,619,200	12,992,376	4.9	3.7	no
Immunology	Absolute CD8+ T cells	Mixed models	3	92,893,655	97,314,231	4.7	6	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Absolute CD8+ T cells	Mixed models	16	77,530,269	82,131,129	4.3	2.9	no
Immunology	Absolute CD8+ T cells	Mixed models	20	1	8,895,081	17.7	12.1	yes
Immunology	Absolute T cells	Mixed models	3	8,642,567	12,969,009	5	2.9	no
Immunology	Absolute T cells	Mixed models	16	77,602,177	82,059,221	4.6	2.8	no
Immunology	Absolute T cells	Mixed models	20	2,370,534	6,814,752	4.7	2.8	yes
Immunology	Expression of CD25 on C84+ cells	Mixed models	18	58,961,713	63,575,843	4.2	4.1	no
Immunology	Expression of CD25 on CD4+ cells	Mixed models	17	10,595,813	15,185,177	4.3	5.6	yes
Immunology	Expression of CD45 in CD4+ cells	Mixed models	13	50,900,689	55,972,579	16.3	17.9	yes
Immunology	Expression of CD45 in CD8+ cells	Mixed models	13	45,309,980	55,890,818	17.5	19.3	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Expression of CD45 in CD8+ cells	Mixed models	13	63,619,011	67,781,011	5.4	10.1	no
Immunology	Expression of CD45 in CD8+ cells	Mixed models	14	94,161,020	98,632,124	4.6	6.6	no
Immunology	Expression of RT1A on granulocytes	Mixed models	2	197,819,922	202,418,698	4.3	5.7	no
Immunology	Expression of RT1A on granulocytes	Mixed models	10	42,879,308	45,735,228	8.8	7.4	yes
Immunology	Expression of RT1A on granulocytes	Mixed models	20	149,845	9,846,504	29.5	21.2	no
Immunology	Expression on RT1B on B cells	Mixed models	2	128,928,781	133,341,617	4.8	10.2	no
Immunology	Expression on RT1B on B cells	Mixed models	17	26,630,201	27,550,391	13.8	18.4	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Expression on RT1B on B cells	Mixed models	20	759,262	4,886,940	5.5	6.3	yes
Immunology	Proportion of B cells in in WBC	Mixed models	1	182,362,275	186,672,851	5	10.2	yes
Immunology	Proportion of B cells in in WBC	Mixed models	1	253,689,747	258,230,679	4.4	2.6	no
Immunology	Proportion of B cells in in WBC	Mixed models	2	200,696,284	211,881,753	6.1	7.2	yes
Immunology	Proportion of B cells in in WBC	Mixed models	3	102,720,285	107,178,909	4.6	6.4	yes
Immunology	Proportion of B cells in in WBC	Mixed models	3	118,220,224	122,833,228	4.2	3.2	no
Immunology	Proportion of B cells in in WBC	Mixed models	10	27,097,648	31,586,054	4.6	8.3	yes
Immunology	Proportion of B cells in in WBC	Mixed models	17	41,422,406	45,693,202	5.1	1.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of B cells in in WBC	Mixed models	18	27,194,905	31,683,133	4.6	5.9	no
Immunology	Proportion of B cells in in WBC	Mixed models	20	1	2,658,096	4.3	7.9	yes
Immunology	Proportion of CD4-CD8- T cells	Mixed models	2	87,284,977	91,476,877	5.3	13	yes
Immunology	Proportion of CD4-CD8- T cells	Mixed models	10	83,735,367	87,292,691	7	8.7	yes
Immunology	Proportion of CD4+ cells expressing CD25	Mixed models	10	10,797,869	14,970,773	5.4	6.3	yes
Immunology	Proportion of CD4+ cells expressing CD25	Mixed models	10	83,819,856	87,771,448	5.9	6.1	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD4+ cells in T cells	Mixed models	5	146,031,404	150,645,910	4.2	2.6	no
Immunology	Proportion of CD4+ cells in T cells	Mixed models	20	1,280,570	8,190,061	35.4	14.2	no
Immunology	Proportion of CD4+ cells in T cells	Mixed models	20	14,832,736	19,427,332	4.3	6.8	yes
Immunology	Proportion of CD4+ cells with expressing CD45RC	Mixed models	2	54,059,408	58,652,242	4.3	5.9	no
Immunology	Proportion of CD4+ cells with expressing CD45RC	Mixed models	13	36,860,138	62,539,705	15.3	16.4	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD4+ cells with high expression of CD25	Mixed models	4	63,458,263	67,955,569	4.5	7.5	no
Immunology	Proportion of CD4+ cells with high expression of CD25	Mixed models	10	84,273,925	87,317,379	8.3	16.3	yes
Immunology	Proportion of CD4+ cells with high expression of CD45RC	Mixed models	13	45,986,106	64,684,059	10.6	12.6	no
Immunology	Proportion of CD4+ cells with low expression of CD45RC	Mixed models	1	124,516,402	128,965,244	4.7	6.4	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD4+ cells with low expression of CD45RC	Mixed models	13	46,982,398	64,822,605	13	17.3	yes
Immunology	Proportion of CD4+ cells with not expressing CD45RC	Mixed models	13	45,163,567	55,204,599	16	39.6	yes
Immunology	Proportion of CD4+ cells with not expressing CD45RC	Mixed models	X	33,168,099	37,756,375	4.3	16.8	no
Immunology	Proportion of CD4+CD8+ T cells	Mixed models	10	70,976,247	75,190,767	5.3	19.1	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD8+ cells expressing CD25	Mixed models	3	30,354,682	34,820,774	4.6	5	no
Immunology	Proportion of CD8+ cells expressing CD25	Mixed models	3	42,852,608	47,457,536	4.3	5.8	no
Immunology	Proportion of CD8+ cells expressing CD25	Mixed models	6	112,199,874	116,772,890	4.3	3.5	yes
Immunology	Proportion of CD8+ cells expressing CD25	Mixed models	10	83,848,501	87,528,205	6.6	5.2	yes
Immunology	Proportion of CD8+ cells in T cells	Mixed models	7	96,909,649	98,600,337	11.8	6.6	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD8+ cells in T cells	Mixed models	7	128,253,601	132,751,249	4.5	3.1	no
Immunology	Proportion of CD8+ cells in T cells	Mixed models	8	123,549,355	127,919,741	4.9	4.7	yes
Immunology	Proportion of CD8+ cells in T cells	Mixed models	10	83,008,758	87,274,648	5.1	6	no
Immunology	Proportion of CD8+ cells in T cells	Mixed models	20	1,418,127	8,232,721	37.4	15.6	no
Immunology	Proportion of CD8+ cells with expressing of CD45RC	Mixed models	13	50,488,781	55,973,003	12.9	16.6	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of CD8+ cells with high expression of CD25	Mixed models	1	165,293,645	169,742,437	4.7	12.3	no
Immunology	Proportion of CD8+ cells with high expression of CD25	Mixed models	6	103,394,748	107,567,028	5.4	24.8	yes
Immunology	Proportion of CD8+ cells with high expression of CD25	Mixed models	19	52,292,698	56,804,282	4.5	18	yes
Immunology	Proportion of CD8+ cells with not expressing of CD45RC	Mixed models	13	49,493,437	52,748,917	7.7	25.8	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Proportion of T cells expressing RT1B	Mixed models	1	129,422,508	133,995,344	4.3	9.7	no
Immunology	Proportion of T cells in WBC	Mixed models	1	184,596,766	188,939,986	4.9	9.9	yes
Immunology	Proportion of T cells in WBC	Mixed models	1	202,721,780	214,993,316	7.1	11.3	yes
Immunology	Ratio of CD4+cells to CD8+ cells	Mixed models	5	59,891,981	64,518,863	4.2	8.3	yes
Immunology	Ratio of CD4+cells to CD8+ cells	Mixed models	8	124,191,088	128,741,884	4.4	4.2	yes
Immunology	Ratio of CD4+cells to CD8+ cells	Mixed models	20	1,418,127	7,789,334	36.5	14.6	no
Immunology	Ratio of T cells to B cells	Mixed models	1	183,583,696	187,405,534	6.3	10	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Immunology	Ratio of T cells to B cells	Mixed models	3	8,524,410	13,087,166	4.4	2.4	no
Immunology	Ratio of T cells to B cells	Mixed models	9	58,300,199	62,770,203	4.6	3.4	no
Immunology	Ratio of T cells to B cells	Mixed models	10	33,182,791	37,748,549	4.4	11	yes
Immunology	Ratio of T cells to B cells	Mixed models	11	29,734,815	34,368,313	4.2	6.2	no
Immunology	Ratio of T cells to B cells	Mixed models	15	11,143,183	15,735,207	4.3	1.6	no
Immunology	Ratio of T cells to B cells	Mixed models	17	41,424,945	45,690,663	5.1	2	no
Induced neuroinflammation	Lowest weight	Mixed models	1	7,048,196	11,572,398	4.5	7.3	yes
Induced neuroinflammation	Lowest weight	Mixed models	2	223,173,939	227,576,861	4.8	5.4	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Induced neuroinflammation	Lowest weight	Mixed models	3	121,448,652	126,253,256	3.7	4.1	yes
Induced neuroinflammation	Lowest weight	Mixed models	4	26,596,752	31,380,866	3.8	2	yes
Induced neuroinflammation	Lowest weight	Mixed models	7	39,990,895	44,613,279	4.2	4.2	no
Induced neuroinflammation	Lowest weight	Mixed models	9	63,250,814	67,798,968	4.4	8.7	yes
Induced neuroinflammation	Lowest weight	Mixed models	9	101,404,244	106,151,352	3.9	4.5	no
Induced neuroinflammation	Lowest weight	Mixed models	11	60,711,368	65,481,892	3.8	3.8	no
Induced neuroinflammation	Lowest weight	Mixed models	12	7,519,165	12,112,699	4.3	6.3	yes
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	2	169,793,140	174,397,944	4.3	3.6	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	2	178,616,315	182,951,401	5	3.1	no
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	7	40,048,651	44,555,523	4.5	4	no
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	7	84,034,035	88,478,831	4.7	3.2	no
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	11	60,232,962	64,688,818	4.6	3.4	no
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	18	28,656,695	32,347,745	6.6	5.5	no
Induced neuroinflammation	Weight loss compared to day 0	Mixed models	20	2,476,091	7,074,515	4.3	4.9	yes
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	1	104,229,041	108,713,635	4.6	3.9	yes
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	4	62,020,592	66,515,872	4.5	4.6	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	7	40,182,930	44,421,244	5.2	4.2	no
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	7	98,016,583	102,576,619	4.4	4.6	no
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	7	121,823,133	126,278,217	4.6	2.9	no
Induced neuroinflammation	Weight loss compared to day 9	Mixed models	18	28,362,808	32,641,632	5.1	4.5	no
Serum biochemistry	Alkaline phosphatase	Mixed models	1	156,991,582	161,516,762	4.5	4.4	yes
Serum biochemistry	Alkaline phosphatase	Mixed models	3	1	12,854,462	26.5	13	yes
Serum biochemistry	Alkaline phosphatase	Mixed models	3	18,490,907	23,107,423	4.2	1.8	yes
Serum biochemistry	Alkaline phosphatase	Mixed models	10	82,409,166	86,627,032	5.3	5.1	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Serum biochemistry	Alkaline phosphatase	Mixed models	12	437,668	4,199,684	6.4	2.7	yes
Serum biochemistry	Alkaline phosphatase	Mixed models	X	45,364,524	49,959,668	4.3	2.8	no
Serum biochemistry	Basal glycemia	Mixed models	8	85,348,595	89,793,429	4.7	4.7	no
Serum biochemistry	Chloride	Mixed models	9	32,722,634	36,503,206	6.4	4.2	yes
Serum biochemistry	Chloride	Mixed models	10	16,920,839	21,060,443	5.5	3.7	no
Serum biochemistry	HDL	Mixed models	1	58,280,289	79,757,998	7.1	7.9	yes
Serum biochemistry	HDL	Mixed models	1	264,962,331	267,865,338	6.1	5	yes
Serum biochemistry	Iron	Mixed models	2	57,035,149	61,780,925	3.9	3.4	yes

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Serum biochemistry	Iron	Mixed models	3	151,217,639	155,683,075	4.6	3.9	no
Serum biochemistry	Iron	Mixed models	9	105,337,287	109,954,567	4.2	2.8	no
Serum biochemistry	Iron	Mixed models	10	75,044,243	79,772,411	3.9	3.9	no
Serum biochemistry	Iron	Mixed models	18	70,893,392	75,109,876	5.3	4.2	yes
Serum biochemistry	Iron	Mixed models	X	66,838,929	71,436,545	4.3	3.3	no
Serum biochemistry	LDL	Mixed models	7	95,123,031	99,491,715	4.9	2.3	no
Serum biochemistry	Potassium	Mixed models	7	62,580,786	66,685,230	5.5	2.9	no
Serum biochemistry	Potassium	Mixed models	X	54,168,230	58,526,630	4.9	3.9	no

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Serum biochemistry	Sodium	Mixed models	7	20,178,419	24,699,787	4.5	2.4	no
Serum biochemistry	Total cholesterol	Mixed models	1	65,844,068	78,635,862	7.3	8.1	yes
Serum biochemistry	Total cholesterol	Mixed models	1	264,698,503	267,865,338	4.8	5.1	yes
Serum biochemistry	Triglycerides	Mixed models	4	74,801,045	79,281,275	4.6	4.5	yes
Serum biochemistry	Triglycerides	Mixed models	10	78,171,290	82,643,278	4.6	3.1	no
Serum biochemistry	Urea	Mixed models	3	44,003,820	48,428,898	4.7	4.3	yes
Wound healing	Hole area	Mixed models	3	21,800,537	25,904,233	5.5	5.8	yes
Anxiety (shuttlebox)	Number of crosses between trials	Resampling	6	1	3,371,571	0.4	1.4	-
Coat colour	Is albino	Resampling	1	141,746,261	145,746,261	0.5	96	-

Phenotype	Measure	Mapping method	Chr	Start QTL (bp)	Stop QTL (bp)	logP/RMIP	Effect size	Has candidate variant
Coat colour	Is dark brown	Resampling	1	141,657,551	145,657,551	0.9	25.3	-
Coat colour	Is dark brown	Resampling	3	143,400,126	147,400,126	0.6	49.7	-
Coat colour	Is light brown	Resampling	1	141,657,551	145,657,551	0.7	19.1	-
Coat colour	Is light brown	Resampling	3	143,400,126	147,400,126	1	46.8	-
Coat colour	Is spotted	Resampling	1	141,657,551	145,657,551	1	27.7	-
Coat colour	Is spotted	Resampling	14	32,428,489	36,428,489	1	38.1	-
Arterial elastic lamina rupture	Number of lesions in AA	Resampling	5	164,908	4,164,908	1	10.6	-
Arterial elastic lamina rupture	Number of lesions in AA and IL	Resampling	5	164,908	4,164,908	0.9	8.6	-
Arterial elastic lamina rupture	Score of lesions in AA	Resampling	5	164,908	4,164,908	1	8.6	-
Arterial elastic lamina rupture	Score of lesions in AA and IL	Resampling	5	164,908	4,164,908	1	6.7	-

APPENDIX D Overlap between the QTLs detected in the HS and the QTLs reported in the Rat Genome Database for the same phenotypes.

Measure	Chr.	Start HS QTL (bp)	Stop HS QTL (bp)	QTL ID in RGD	Start RGD QTL (bp)	Stop RGD QTL (bp)	Type of cross	p-value of overlap
Score of lesions in abdominal aorta	5	164,908	4,164,908	631551	1,368,223	5,112,800	F2	<1/1000
Total cholesterol	1	65,844,068	78,635,862	631512	71,659,707	90,282,193		0.007
Total cholesterol	1	264,698,503	267,865,338	631690	243,502,887	267,910,886		
Total cholesterol	1	264,698,503	267,865,338	631835	226,083,572	267,910,886	backcross	
Total cholesterol	1	264,698,503	267,865,338	631836	244,611,037	267,910,886	backcross	
Heart weight	1	202,149,330	206,633,814	1358292	201,920,676	246,920,676		0.034
Heart weight	10	37,243,161	47,929,176	631532	18,217,625	53,793,117	intercross	
Heart weight	13	56,470,137	67,956,785	1558644	25,198,204	70,198,204		
Femur midshaft polar moment of inertia	1	201,265,878	205,189,952	2293654	176,612,333	221,612,333	intercross	0.058
Femur midshaft polar moment of inertia	10	90,877,745	95,483,369	2293663	74,167,134	110,718,848	intercross	
Femur midshaft polar moment of inertia	4	96,051,838	106,296,290	1578658	61,483,655	106,483,655		
Lumbar mineral density	1	195,565,176	199,903,482	2300174	176,612,333	221,612,333	intercross	0.157

Body weight at day immunization	12	7,483,933	12,147,931	2303568	1	25,457,135	intercross	
Body weight at day immunization	2	129,257,158	133,732,902	1358887	24,474,676	163,154,227		
Body weight at day immunization	2	138,471,008	142,614,192	1358887	24,474,676	163,154,227		
Body weight at day immunization	2	216,921,362	220,119,428	1358900	163,154,358	227,150,051		
Body weight at day immunization	2	129,257,158	133,732,902	1358908	24,474,676	163,154,227		
Body weight at day immunization	2	138,471,008	142,614,192	1358908	24,474,676	163,154,227		
								0.3
Body weight at day immunization	3	23,637,280	36,416,292	1354589	30,253,942	76,620,970	intercross	
Body weight at day immunization	3	23,637,280	36,416,292	1354604	30,253,942	103,304,908	intercross	
Body weight at day immunization	3	23,637,280	36,416,292	1558654	6,373,335	26,674,263		
Body weight at day immunization	4	103,453,924	108,048,784	1357342	75,732,943	119,369,308		
Body weight at day immunization	4	103,453,924	108,048,784	1549843	60,262,965	104,415,981		
Body weight at day immunization	4	27,024,851	31,510,205	2303585	11,706,134	56,706,134	intercross	

Body weight at day immunization	4	103,453,924	108,048,784	70167	75,732,943	119,369,308		
Body weight at day immunization	8	81,857,162	86,244,836	1331837	85,328,126	103,665,018		
Body weight at day immunization	8	81,857,162	86,244,836	1358912	54,364,071	112,242,906		
Body weight at day immunization	8	81,857,162	86,244,836	1582243	57,308,572	89,558,994	intercross	
Body weight at day immunization	4	103,453,924	108,048,784	1549839	60,262,965	116,780,394		
Weight loss compared to day 9	18	28,362,808	32,641,632	70178	13,239,641	58,239,641	F2	0.501
Weight loss compared to day 9	4	62,020,592	66,515,872	2317577	65,821,936	71,729,738		
Femur midshaft total area	10	91,151,727	95,209,387	2293646	74,167,134	110,718,848	intercross	0.7

APPENDIX E Measures collected in both rat and mouse HS.

Phenotype	Measure
Anxiety (novel cage)	Distance 0' to 30'
Anxiety (zeromaze)	Number of entries in open section
Anxiety (zeromaze)	Time spent in open section
Glucose tolerance	Area under glycemia curve
Glucose tolerance	Area under glycemia curve over baseline
Glucose tolerance	Glycemia before injection
Hematology	Aboslute number of basophils
Hematology	Absolute number of lymphocytes
Hematology	Absolute number of monocytes
Hematology	Absolute number of neutrophils
Hematology	Hematocrit
Hematology	Hemoglobin
Hematology	Mean corpuscular hemoglobin
Hematology	Mean corpuscular volume
Hematology	Mean platelet volume
Hematology	Measured mean cell hemoglobin concentration
Hematology	Platelet count
Hematology	Plateletcrit
Hematology	Red blood cell count
Hematology	Red blood cell distribution width
Hematology	WBC in the Baso channel
Immunology	Proportion of B cells in WBC
Immunology	Proportion of CD4+ cells in T cells
Immunology	Proportion of CD8+ cells in T cells

Immunology	Proportion of T cells in WBC
Immunology	Ratio of CD4+cells to CD8+ cells
Serum biochemistry	Alanine aminotransferase
Serum biochemistry	Alkaline phosphatase
Serum biochemistry	Aspartate aminotransferase
Serum biochemistry	Calcium
Serum biochemistry	Chloride
Serum biochemistry	Creatinine
Serum biochemistry	HDL
Serum biochemistry	LDL
Serum biochemistry	Total cholesterol
Serum biochemistry	Triglycerides
Serum biochemistry	Urea
Wound healing	Hole area

APPENDIX F KEGG pathway enrichment analysis for rat QTLs. Pathways with an empirical p-value smaller than 0.05 are shown. Those enriched at a corrected p-value smaller than 0.01 are marked with *. KEGG pathways enriched in both the rat and mouse QTLs for homologous measures are highlighted in blue.

Measure	**Number of genes in KEGG pathway**	**Number of intervals with genes in KEGG pathway**	**Empirical p-value**	**Corrected p-value**	**KEGG pathway name**
Aboslute number of basophils	95	2	0.0227954	0.355729	Amoebiasis
Aboslute number of basophils	146	2	0.045191	0.489302	Phagosome
Absolute number of lymphocytes	95	3	0.0493901	0.905019	Amoebiasis
Absolute number of lymphocytes	73	2	0.0257948	0.755049	Antigen processing and presentation
Absolute number of lymphocytes	103	3	0.0409918	0.84943	T cell receptor signaling pathway
Absolute number of monocytes	64	4	0.0395921	0.765647	Adipocytokine signaling pathway

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Absolute number of monocytes	29	2	0.0423915	0.789242	African trypanosomiasis
Absolute number of monocytes	66	4	0.029994	0.695261	ECM-receptor interaction
Absolute number of monocytes	39	2	0.0169966	0.509498	Fat digestion and absorption
Absolute number of monocytes	76	5	0.00019996	0.0069986*	Glycerophospholipid metabolism
Absolute number of monocytes	12	2	0.00039992	0.0115977	Glycosphingolipid biosynthesis - ganglio series
Absolute number of monocytes	90	4	0.0229954	0.641472	GnRH signaling pathway
Absolute number of monocytes	41	3	0.0223955	0.632873	Intestinal immune network for IgA production
Absolute number of monocytes	61	3	0.04979	0.840032	Long-term potentiation
Absolute number of monocytes	1018	6	0.00019996	0.0069986*	Metabolic pathways

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Absolute number of monocytes	20	2	0.00759848	0.259548	Pentose and glucuronate interconversions
Absolute number of monocytes	146	4	0.0171966	0.511898	Phagosome
Absolute number of monocytes	68	3	0.0359928	0.728854	Salivary secretion
Absolute number of monocytes	117	4	0.0477904	0.826835	Spliceosome
Absolute number of monocytes	9	2	0.00039992	0.0115977	Valine, leucine and isoleucine biosynthesis
Absolute number of neutrophils	73	3	0.0209958	0.65127	Antigen processing and presentation
Absolute number of neutrophils	49	4	0.0133973	0.482304	mTOR signaling pathway
Absolute number of neutrophils	45	3	0.0421916	0.871826	Type II diabetes mellitus
Alkaline phosphatase	39	3	0.0483903	0.910818	ABC transporters
Alkaline phosphatase	66	5	0.00379924	0.204159	ECM-receptor interaction

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Alkaline phosphatase	73	4	0.0467906	0.910818	Fc epsilon RI signaling pathway
Alkaline phosphatase	74	4	0.029794	0.772046	Hematopoietic cell lineage
Alkaline phosphatase	117	6	0.0253949	0.729654	Neurotrophin signaling pathway
Alkaline phosphatase	44	3	0.024795	0.712857	Notch signaling pathway
Alkaline phosphatase	67	5	0.0327934	0.80224	Pancreatic cancer
Alkaline phosphatase	94	6	0.0361928	0.832633	Ribosome
Alkaline phosphatase	77	4	0.0421916	0.885023	Systemic lupus erythematosus
Alkaline phosphatase	45	4	0.0317936	0.797441	Type II diabetes mellitus
Alkaline phosphatase	80	4	0.0361928	0.832633	Viral myocarditis
Area under glycemia curve	13	2	0.0165967	0.480104	Terpenoid backbone biosynthesis
Area under glycemia curve over baseline	80	4	0.0433913	0.896221	Gap junction

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Chloride	20	3	0.00079984	0.0209958	Circadian rhythm - mammal
Chloride	84	3	0.0345931	0.707459	Pyrimidine metabolism
Creatinine	136	3	0.0195961	0.418116	Cell adhesion molecules (CAMs)
Glycemia before injection1	254	4	0.0307938	0.788642	Neuroactive ligand-receptor interaction
Glycemia before injection1	117	3	0.0421916	0.853429	Spliceosome
Glycemia before injection2	254	4	0.0307938	0.788642	Neuroactive ligand-receptor interaction
Glycemia before injection2	117	3	0.0421916	0.853429	Spliceosome
HDL	31	2	0.0421916	0.829834	Basal transcription factors
HDL	33	2	0.0261948	0.719456	Ether lipid metabolism
HDL	94	3	0.0185963	0.636873	Pancreatic secretion
HDL	68	3	0.039792	0.80084	Renal cell carcinoma

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
HDL	51	2	0.0429914	0.834233	Retinol metabolism
HDL	142	5	0.00959808	0.438312	RNA transport
Hematocrit	120	5	0.0203959	0.75145	Axon guidance
Hematocrit	118	5	0.0103979	0.503499	Cell cycle
Hematocrit	104	4	0.0405919	0.906219	Oocyte meiosis
Hematocrit	43	3	0.0171966	0.674865	Proteasome
Hemoglobin	22	2	0.0433913	0.940412	Mismatch repair
Hemoglobin	254	6	0.0321936	0.888222	Neuroactive ligand-receptor interaction
Hemoglobin	94	4	0.0165967	0.589482	Ribosome
Hemoglobin	41	3	0.0311938	0.874625	Vasopressin-regulated water reabsorption
Hole area	118	2	0.0113977	0.142372	Cell cycle
LDL	83	2	0.0405919	0.631274	ErbB signaling pathway
LDL	21	2	0.009998	0.290742	Nitrogen metabolism
LDL	118	2	0.0471906	0.673865	Oxidative phosphorylation

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
LDL	117	2	0.0431914	0.647271	Parkinson's disease
LDL	73	2	0.0275945	0.532893	Small cell lung cancer
Mean corpuscular hemoglobin	66	4	0.0423915	0.944411	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Mean corpuscular hemoglobin	74	5	0.0481904	0.965807	Hematopoietic cell lineage
Mean corpuscular hemoglobin	41	4	0.025195	0.817836	Intestinal immune network for IgA production
Mean corpuscular hemoglobin	74	6	0.0157968	0.661068	mRNA surveillance pathway
Mean corpuscular hemoglobin	117	7	0.024795	0.806239	Neurotrophin signaling pathway
Mean corpuscular hemoglobin	70	5	0.0265947	0.832234	Ribosome biogenesis in eukaryotes
Mean corpuscular hemoglobin	45	4	0.0469906	0.961608	Type II diabetes mellitus
Mean corpuscular volume	19	3	0.00219956	0.127974	Dorso-ventral axis formation

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Mean corpuscular volume	23	2	0.0217956	0.719456	Glycosaminoglycan biosynthesis - heparan sulfate
Mean corpuscular volume	74	3	0.0417916	0.915817	mRNA surveillance pathway
Mean corpuscular volume	254	6	0.0161968	0.591882	Neuroactive ligand-receptor interaction
Mean corpuscular volume	21	2	0.0229954	0.75105	Nitrogen metabolism
Mean corpuscular volume	146	5	0.0217956	0.719456	Phagosome
Mean corpuscular volume	32	3	0.0271946	0.830834	SNARE interactions in vesicular transport
Mean corpuscular volume	13	2	0.0123975	0.465107	Terpenoid backbone biosynthesis
Mean platelet volume	44	2	0.0365927	0.84863	Amino sugar and nucleotide sugar metabolism

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Mean platelet volume	32	2	0.0233953	0.715257	Fructose and mannose metabolism
Mean platelet volume	15	2	0.0115977	0.455109	One carbon pool by folate
Mean platelet volume	66	3	0.0289942	0.784243	RNA degradation
Measured mean cell hemoglobin concentration	136	3	0.0403919	0.858428	Cell adhesion molecules (CAMs)
Measured mean cell hemoglobin concentration	96	3	0.00839832	0.312138	Natural killer cell mediated cytotoxicity
Measured mean cell hemoglobin concentration	16	2	0.00739852	0.282943	Phenylalanine metabolism
Measured mean cell hemoglobin concentration	34	2	0.0283943	0.722655	Pyruvate metabolism
Number of entries in open section	71	3	0.0445911	0.912418	Chronic myeloid leukemia
Number of entries in open section	175	4	0.0215957	0.713057	Focal adhesion

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Number of entries in open section	41	3	0.00059988	0.0303939	Intestinal immune network for IgA production
Number of entries in open section	96	3	0.044791	0.916617	Natural killer cell mediated cytotoxicity
Number of entries in open section	15	2	0.0285943	0.79864	One carbon pool by folate
Number of entries in open section	70	3	0.0459908	0.922416	Ribosome biogenesis in eukaryotes
Number of entries in open section	110	3	0.0331934	0.857628	Vascular smooth muscle contraction
Platelet count	66	2	0.00019996	0.00019996*	RNA degradation
Platelet count	80	2	0.00019996	0.00019996*	Viral myocarditis
Plateletcrit	41	2	0.0327934	0.693261	Intestinal immune network for IgA production
Plateletcrit	43	3	0.00439912	0.115977	Proteasome
Proportion of B cells in WBC	29	7	0.00079984	0.0683863	African trypanosomiasis

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of B cells in WBC	95	8	0.0217956	0.812837	Amoebiasis
Proportion of B cells in WBC	31	5	0.0143971	0.658068	Basal transcription factors
Proportion of B cells in WBC	118	10	0.0069986	0.441912	Cell cycle
Proportion of B cells in WBC	74	6	0.0291942	0.884623	Hematopoietic cell lineage
Proportion of B cells in WBC	166	11	0.0143971	0.658068	Huntington's disease
Proportion of B cells in WBC	129	8	0.0289942	0.882823	Jak-STAT signaling pathway
Proportion of B cells in WBC	64	6	0.0411918	0.954209	Long-term depression
Proportion of B cells in WBC	61	6	0.0215957	0.806839	Long-term potentiation
Proportion of CD4+ cells in T cells	8	2	0.00979804	0.373525	Fatty acid elongation in mitochondria
Proportion of CD4+ cells in T cells	22	2	0.0263947	0.704659	Galactose metabolism

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of CD4+ cells in T cells	33	2	0.0459908	0.865427	Primary immunodeficiency
Proportion of CD8+ cells in T cells	54	4	0.0209958	0.724855	Amyotrophic lateral sclerosis (ALS)
Proportion of CD8+ cells in T cells	93	6	0.00339932	0.210958	Chagas disease (American trypanosomiasis)
Proportion of CD8+ cells in T cells	60	5	0.0039992	0.235353	Colorectal cancer
Proportion of CD8+ cells in T cells	202	7	0.0295941	0.821836	Endocytosis
Proportion of CD8+ cells in T cells	129	6	0.0315937	0.832633	Jak-STAT signaling pathway
Proportion of CD8+ cells in T cells	103	6	0.00719856	0.358928	Osteoclast differentiation
Proportion of CD8+ cells in T cells	84	5	0.0235953	0.745851	Pyrimidine metabolism
Proportion of CD8+ cells in T cells	61	4	0.0289942	0.817636	RIG-I-like receptor signaling pathway

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of CD8+ cells in T cells	80	5	0.0379924	0.873425	TGF-beta signaling pathway
Proportion of CD8+ cells in T cells	9	2	0.0425915	0.890222	Valine, leucine and isoleucine biosynthesis
Proportion of T cells in WBC	40	4	0.029794	0.823835	Aldosterone-regulated sodium reabsorption
Proportion of T cells in WBC	95	5	0.0261948	0.788042	Amoebiasis
Proportion of T cells in WBC	17	2	0.0489902	0.94821	Ascorbate and aldarate metabolism
Proportion of T cells in WBC	65	5	0.0421916	0.915017	Bacterial invasion of epithelial cells
Proportion of T cells in WBC	65	5	0.0261948	0.788042	Cardiac muscle contraction
Proportion of T cells in WBC	118	6	0.04979	0.955409	Cell cycle
Proportion of T cells in WBC	80	6	0.0169966	0.65227	Fc gamma R-mediated phagocytosis

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of T cells in WBC	23	3	0.040192	0.908218	Glycosaminoglycan biosynthesis - heparan sulfate
Proportion of T cells in WBC	24	3	0.0333933	0.864027	Histidine metabolism
Proportion of T cells in WBC	42	4	0.0173965	0.658268	Lysine degradation
Proportion of T cells in WBC	96	5	0.0493901	0.954809	Natural killer cell mediated cytotoxicity
Proportion of T cells in WBC	22	3	0.0489902	0.94821	Nicotinate and nicotinamide metabolism
Proportion of T cells in WBC	50	4	0.0409918	0.908218	Non-small cell lung cancer
Proportion of T cells in WBC	16	3	0.0485903	0.942012	Phenylalanine metabolism
Proportion of T cells in WBC	34	3	0.0369926	0.894621	Prion diseases
Proportion of T cells in WBC	32	4	0.0169966	0.65227	SNARE interactions in vesicular transport

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of T cells in WBC	42	3	0.0481904	0.940812	Tryptophan metabolism
Ratio of CD4+cells to CD8+ cells	39	3	0.0259948	0.716857	ABC transporters
Ratio of CD4+cells to CD8+ cells	23	2	0.0193961	0.612478	Glycosaminoglycan biosynthesis - heparan sulfate
Ratio of CD4+cells to CD8+ cells	76	4	0.039992	0.835033	Hypertrophic cardiomyopathy (HCM)
Ratio of CD4+cells to CD8+ cells	20	2	0.0469906	0.858428	Pentose and glucuronate interconversions
Ratio of CD4+cells to CD8+ cells	67	4	0.009998	0.388322	PPAR signaling pathway
Ratio of CD4+cells to CD8+ cells	145	5	0.0119976	0.452709	Purine metabolism
Ratio of CD4+cells to CD8+ cells	9	2	0.0271946	0.730654	Valine, leucine and isoleucine biosynthesis
Red blood cell count	120	5	0.034793	0.866427	Axon guidance

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Red blood cell count	31	3	0.020196	0.724055	Basal transcription factors
Red blood cell count	51	3	0.0215957	0.739652	Hedgehog signaling pathway
Red blood cell count	77	3	0.0241952	0.778644	Systemic lupus erythematosus
Red blood cell count	31	3	0.0471906	0.917816	Tyrosine metabolism
Red blood cell count	141	6	0.0119976	0.4987	Wnt signaling pathway
Red blood cell distribution width	54	3	0.0463907	0.945611	Amyotrophic lateral sclerosis (ALS)
Red blood cell distribution width	120	5	0.045191	0.939412	Axon guidance
Red blood cell distribution width	161	6	0.0167966	0.681064	Calcium signaling pathway
Red blood cell distribution width	19	2	0.0367926	0.902619	Dorso-ventral axis formation

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Red blood cell distribution width	121	5	0.0385923	0.904819	Insulin signaling pathway
Red blood cell distribution width	31	3	0.0461908	0.941212	Tyrosine metabolism
Time spent in open section	250	4	0.0419916	0.886023	MAPK signaling pathway
Time spent in open section	29	2	0.0089982	0.370526	Thyroid cancer
Total cholesterol	60	5	0.0209958	0.728454	Colorectal cancer
Total cholesterol	33	3	0.0353929	0.90122	Ether lipid metabolism
Triglycerides	80	2	0.019996	0.679064	Viral myocarditis
WBC in the Baso channel	95	5	0.0423915	0.931814	Amoebiasis

APPENDIX G KEGG pathway enrichment analysis for mouse QTLs. Pathways with an empirical p-value smaller than 0.05 are shown. Those enriched at a corrected p-value smaller than 0.01 are marked with *. KEGG pathways enriched in both the rat and mouse QTLs for homologous measures are highlighted in blue.

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Aboslute number of basophils	56	2	0.0121976	0.254949	NOD-like receptor signaling pathway
Absolute number of monocytes	44	3	0.0315937	0.739052	ABC transporters
Absolute number of monocytes	34	4	0.0213957	0.612478	Basal transcription factors
Absolute number of monocytes	89	4	0.0261948	0.659868	Prostate cancer
Alanine aminotransferase	130	6	0.0059988	0.30234	Axon guidance
Alanine aminotransferase	27	3	0.0361928	0.881224	Collecting duct acid secretio
Alanine aminotransferase	27	2	0.0229954	0.769846	Galactose metabolism
Alanine aminotransferase	101	3	0.0475905	0.934813	Pancreatic secretion

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Alanine aminotransferase	84	4	0.0411918	0.907419	TGF-beta signaling pathway
Alkaline phosphatase	73	2	0.0491902	0.537692	Complement and coagulatio cascades
Alkaline phosphatase	60	2	0.00019996	0.00019996*	Taste transduction
Area under glycemia curve	78	2	0.0365927	0.60168	Drug metabolism - cytochro P450
Area under glycemia curve	86	2	0.0155969	0.232553	Gap junction
Area under glycemia curve	68	2	0.0365927	0.60168	Metabolism of xenobiotics b cytochrome P450
Aspartate aminotransferase	54	3	0.0411918	0.857628	Arginine and proline metabolism
Aspartate aminotransferase	17	2	0.0159968	0.489302	Ascorbate and aldarate metabolism
Aspartate aminotransferase	6	2	0.0183963	0.513097	Cyanoamino acid metabolis
Aspartate aminotransferase	34	2	0.0469906	0.894421	Ether lipid metabolism
Aspartate aminotransferase	46	3	0.0139972	0.44791	Fatty acid metabolism

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Aspartate aminotransferase	44	2	0.0453909	0.878624	Lysine degradation
Aspartate aminotransferase	31	3	0.0029994	0.120176	Propanoate metabolism
Calcium	68	4	0.0203959	0.70186	Adipocytokine signaling pathway
Calcium	32	4	0.0089982	0.437912	Alanine, aspartate and gluta metabolism
Calcium	63	4	0.0205959	0.706259	Colorectal cancer
Calcium	86	4	0.0469906	0.89842	Progesterone-mediated oocy maturation
Calcium	44	3	0.0415917	0.883023	Tryptophan metabolism
Calcium	49	3	0.0265947	0.787243	Type II diabetes mellitus
Calcium	7	2	0.0137972	0.572885	Ubiquinone and other terper quinone biosynthesis
Calcium	24	3	0.00659868	0.330734	Vitamin digestion and absorption

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Chloride	53	4	0.0439912	0.933613	Amyotrophic lateral sclerosi (ALS)
Chloride	17	3	0.019796	0.723255	Ascorbate and aldarate metabolism
Chloride	23	2	0.0337932	0.896621	Asthma
Chloride	23	3	0.0235953	0.775645	beta-Alanine metabolism
Chloride	30	3	0.0459908	0.940612	Butanoate metabolism
Chloride	46	4	0.0213957	0.746651	Fatty acid metabolism
Chloride	57	5	0.0029994	0.191962	Glycolysis / Gluconeogenes
Chloride	78	4	0.0375925	0.907618	PPAR signaling pathway
Chloride	48	3	0.0069986	0.362927	Staphylococcus aureus infec
Chloride	46	4	0.00739852	0.381324	Steroid hormone biosynthes
Chloride	92	5	0.00319936	0.214957	Systemic lupus erythematos
Chloride	44	5	0.0125975	0.545891	Tryptophan metabolism

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Distance 0' to 30'	41	2	0.0473905	0.865227	Aminoacyl-tRNA biosynthe
Distance 0' to 30'	113	3	0.0195961	0.593481	Amoebiasis
Distance 0' to 30'	197	4	0.00319936	0.0987802	Focal adhesion
Distance 0' to 30'	34	2	0.00779844	0.242951	Glycine, serine and threonin metabolism
Distance 0' to 30'	85	3	0.035193	0.778244	mRNA surveillance pathway
Distance 0' to 30'	108	3	0.0141972	0.439912	Oocyte meiosis
Distance 0' to 30'	86	3	0.0171966	0.4979	Progesterone-mediated oocy maturation
Distance 0' to 30'	162	4	0.0107978	0.369926	Protein processing in endoplasmic reticulum
Distance 0' to 30'	70	3	0.0291942	0.69966	RNA degradation
Distance 0' to 30'	108	3	0.0239952	0.672466	T cell receptor signaling pathway
Glycemia before injection 1	18	2	0.0267946	0.827634	Steroid biosynthesis

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Glycemia before injection 10	55	5	0.0219956	0.772446	Basal cell carcinoma
Glycemia before injection 2	54	4	0.0263947	0.817037	Arginine and proline metabolism
Glycemia before injection 2	18	2	0.0239952	0.792442	Steroid biosynthesis
Glycemia before injection 3	54	4	0.0281944	0.840432	Arginine and proline metabolism
Glycemia before injection 3	37	3	0.0455909	0.946611	Starch and sucrose metaboli
Glycemia before injection 4	18	3	0.0079984	0.433713	Glyoxylate and dicarboxylat metabolism
Glycemia before injection 4	37	3	0.0433913	0.935013	Starch and sucrose metaboli
Glycemia before injection 5	18	3	0.00679864	0.429714	Glyoxylate and dicarboxylat metabolism
Glycemia before injection 5	108	7	0.0355929	0.891422	Oocyte meiosis
Glycemia before injection 6	108	7	0.035193	0.884623	Oocyte meiosis
Glycemia before injection 6	163	8	0.004999	0.30134	Phagosome
Glycemia before injection 7	163	8	0.00439912	0.292541	Phagosome
Glycemia before injection 7	134	8	0.0409918	0.922216	Tight junction

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Glycemia before injection 8	86	7	0.0105979	0.541092	Progesterone-mediated oocy maturation
Glycemia before injection 8	134	8	0.0389922	0.909418	Tight junction
Glycemia before injection 9	55	5	0.0173965	0.712657	Basal cell carcinoma
Glycemia before injection 9	86	7	0.00979804	0.519496	Progesterone-mediated oocy maturation
HDL	123	7	0.0119976	0.541092	Cell cycle
HDL	238	7	0.0489902	0.945011	Cytokine-cytokine receptor interaction
HDL	70	4	0.0125975	0.560288	Gastric acid secretion
HDL	56	3	0.0373925	0.888222	NOD-like receptor signaling pathway
HDL	20	3	0.0179964	0.683063	Proximal tubule bicarbonate reclamation
HDL	83	7	0.0229954	0.766647	Small cell lung cancer
HDL	125	7	0.0111978	0.515297	Spliceosome

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Hematocrit	46	2	0.0193961	0.569486	Steroid hormone biosynthes
Hemoglobin	44	2	0.004999	0.124175	ABC transporters
Hemoglobin	43	2	0.00239952	0.0435913	Fat digestion and absorption
Hemoglobin	86	2	0.0105979	0.260548	Gap junction
Hole area	53	4	0.0177964	0.591882	Amyotrophic lateral sclerosi (ALS)
Hole area	85	4	0.0217956	0.678864	Apoptosis
Hole area	83	5	0.00059988	0.0339932	Arachidonic acid metabolisr
Hole area	6	3	0.00019996	0.0177964*	Cyanoamino acid metabolis
Hole area	26	4	0.00039992	0.0245951	Histidine metabolism
Hole area	169	6	0.0355929	0.845231	Huntington's disease
Hole area	66	3	0.0163967	0.543691	Long-term potentiation
Hole area	56	4	0.00319936	0.141372	NOD-like receptor signaling pathway
Hole area	19	2	0.0385923	0.873025	One carbon pool by folate

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Hole area	35	3	0.0387922	0.876225	Other types of O-glycan biosynthesis
Hole area	17	3	0.0139972	0.485303	Phenylalanine metabolism
Hole area	34	2	0.0377924	0.867626	Primary immunodeficiency
Hole area	92	5	0.0175965	0.584283	Pyrimidine metabolism
Hole area	10	3	0.0255949	0.720256	Taurine and hypotaurine metabolism
Hole area	29	3	0.0195961	0.641272	Thyroid cancer
Hole area	36	4	0.00379924	0.171366	Tyrosine metabolism
Hole area	24	3	0.00179964	0.0873825	Vitamin digestion and absorption
LDL	8	2	0.0135973	0.394921	Fatty acid elongation in mitochondria
LDL	27	3	0.00319936	0.0767846	Homologous recombination
LDL	150	4	0.0163967	0.464107	Jak-STAT signaling pathwa
LDL	45	3	0.0273945	0.666067	Malaria
LDL	22	2	0.0363927	0.74985	Mismatch repair

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Mean corpuscular hemoglobin	160	5	0.0209958	0.74845	Alzheimer's disease
Mean corpuscular hemoglobin	78	4	0.0159968	0.616077	Glycerophospholipid metabolism
Mean corpuscular hemoglobin	95	4	0.0281944	0.842032	GnRH signaling pathway
Mean corpuscular hemoglobin	118	5	0.00959808	0.460708	Leukocyte transendothelial migration
Mean corpuscular hemoglobin	54	5	0.00119976	0.0705859	Non-small cell lung cancer
Mean corpuscular hemoglobin	77	3	0.0495901	0.934013	Protein digestion and absorp
Mean corpuscular hemoglobin	83	4	0.0145971	0.583083	Small cell lung cancer
Mean corpuscular hemoglobin	108	5	0.014797	0.592082	T cell receptor signaling pathway
Mean corpuscular hemoglobin	116	5	0.0157968	0.613277	Vascular smooth muscle contraction
Mean corpuscular hemoglobin	75	5	0.00439912	0.20036	VEGF signaling pathway
Mean corpuscular volume	160	5	0.00779844	0.338332	Alzheimer's disease

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Mean corpuscular volume	53	3	0.0141972	0.566487	Amyotrophic lateral sclerosi (ALS)
Mean corpuscular volume	130	5	0.0111978	0.469106	Axon guidance
Mean corpuscular volume	42	3	0.030194	0.806839	Bladder cancer
Mean corpuscular volume	100	4	0.0245951	0.741652	Chagas disease (American trypanosomiasis)
Mean corpuscular volume	31	3	0.0137972	0.545291	Citrate cycle (TCA cycle)
Mean corpuscular volume	22	2	0.0389922	0.863227	Dorso-ventral axis formatio
Mean corpuscular volume	95	4	0.0145971	0.570486	GnRH signaling pathway
Mean corpuscular volume	118	4	0.0219956	0.695261	Leukocyte transendothelial migration
Mean corpuscular volume	54	4	0.0257948	0.745051	Non-small cell lung cancer
Mean corpuscular volume	74	4	0.00779844	0.338332	Salivary secretion

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Mean corpuscular volume	116	4	0.0353929	0.840032	Vascular smooth muscle contraction
Mean corpuscular volume	75	5	0.00519896	0.226355	VEGF signaling pathway
Mean platelet volume	197	4	0.0229954	0.572286	Focal adhesion
Mean platelet volume	25	2	0.0211958	0.527495	Glycosylphosphatidylinosito PI)-anchor biosynthesis
Mean platelet volume	125	4	0.0113977	0.320936	Spliceosome
Measured mean cell hemoglobin concentration	113	4	0.015197	0.562288	Amoebiasis
Measured mean cell hemoglobin concentration	53	2	0.0479904	0.931614	Amyotrophic lateral sclerosi (ALS)
Measured mean cell hemoglobin concentration	27	2	0.00739852	0.294341	Collecting duct acid secretic
Measured mean cell hemoglobin concentration	84	5	0.0133973	0.513097	ECM-receptor interaction
Measured mean cell hemoglobin concentration	197	7	0.0211958	0.674265	Focal adhesion
Measured mean cell hemoglobin concentration	117	5	0.0487902	0.936213	Oxidative phosphorylation

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Measured mean cell hemoglobin concentration	77	5	0.00279944	0.105779	Protein digestion and absorp
Measured mean cell hemoglobin concentration	35	2	0.0341932	0.861228	Regulation of autophagy
Measured mean cell hemoglobin concentration	83	5	0.00019996	0.0169966*	Small cell lung cancer
Measured mean cell hemoglobin concentration	60	2	0.0227954	0.70026	Taste transduction
Number of entries in open section	81	2	0.0089982	0.185963	Hematopoietic cell lineage
Number of entries in open section	126	2	0.0289942	0.557688	Neurotrophin signaling path
Number of entries in open section	67	2	0.0209958	0.462507	p53 signaling pathway
Platelet count	45	3	0.0131974	0.434313	Malaria
Platelet count	22	2	0.0259948	0.659468	Mismatch repair
Platelet count	84	3	0.0429914	0.838632	TGF-beta signaling pathway
Plateletcrit	62	3	0.0467906	0.844631	Glioma
Plateletcrit	71	2	0.024795	0.642072	Melanoma
Proportion of B cells in WBC	130	6	0.0365927	0.89782	Axon guidance
Proportion of B cells in WBC	143	6	0.00179964	0.110578	Cell adhesion molecules (CAMs)
Proportion of B cells in WBC	35	3	0.0387922	0.918816	DNA replication

Measure	**Number of genes in KEGG pathway**	**Number of intervals with genes in KEGG pathway**	**Empirical p-value**	**Corrected p-value**	**KEGG pathway name**
Proportion of B cells in WBC	213	9	0.0019996	0.117776	Endocytosis
Proportion of B cells in WBC	81	7	0.00019996	0.0203959*	Hematopoietic cell lineage
Proportion of B cells in WBC	150	5	0.0385923	0.917217	Jak-STAT signaling pathwa
Proportion of B cells in WBC	118	5	0.0327934	0.877425	Leukocyte transendothelial migration
Proportion of B cells in WBC	92	6	0.00619876	0.34933	Pyrimidine metabolism
Proportion of B cells in WBC	48	3	0.0395921	0.930814	Staphylococcus aureus infec
Proportion of CD8+ cells in T cell	113	4	0.0371926	0.835833	Amoebiasis
Proportion of CD8+ cells in T cell	35	2	0.0429914	0.863227	DNA replication
Proportion of T cells in WBC	48	4	0.0289942	0.79964	Amino sugar and nucleotide sugar metabolism
Proportion of T cells in WBC	23	3	0.00019996	0.0193961*	Asthma
Proportion of T cells in WBC	39	3	0.0403919	0.910818	Carbohydrate digestion and absorption
Proportion of T cells in WBC	174	6	0.0315937	0.84863	Chemokine signaling pathw

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
Proportion of T cells in WBC	73	3	0.0357928	0.881424	Complement and coagulatio cascades
Proportion of T cells in WBC	55	3	0.039992	0.910218	Cytosolic DNA-sensing pathway
Proportion of T cells in WBC	27	3	0.0191962	0.674865	Galactose metabolism
Proportion of T cells in WBC	64	4	0.0241952	0.724055	Leishmaniasis
Proportion of T cells in WBC	92	4	0.00179964	0.103179	Systemic lupus erythematos
Ratio of CD4+cells to CD8+ cells	113	4	0.0371926	0.835833	Amoebiasis
Ratio of CD4+cells to CD8+ cells	35	2	0.0429914	0.863227	DNA replication
Red blood cell distribution width	27	2	0.0361928	0.74905	Collecting duct acid secretio
Red blood cell distribution width	197	5	0.00419916	0.14937	Focal adhesion
Red blood cell distribution width	22	2	0.0195961	0.519096	Mismatch repair
Red blood cell distribution width	17	2	0.00379924	0.132773	Other glycan degradation
Red blood cell distribution width	163	3	0.0295941	0.65067	Phagosome
Red blood cell distribution width	43	2	0.00979804	0.318136	Vasopressin-regulated wate reabsorption

Measure	**Number of genes in KEGG pathway**	**Number of intervals with genes in KEGG pathway**	**Empirical p-value**	**Corrected p-value**	**KEGG pathway name**
Total cholesterol	44	2	0.0485903	0.914417	Aldosterone-regulated sodiu reabsorption
Total cholesterol	25	2	0.0387922	0.873225	Phototransduction
Triglycerides	39	2	0.0113977	0.335533	Carbohydrate digestion and absorption
Triglycerides	26	2	0.00039992	0.0191962	Histidine metabolism
Triglycerides	67	2	0.009998	0.293141	Retinol metabolism
Triglycerides	37	2	0.0345931	0.774645	Starch and sucrose metaboli
Urea	44	2	0.00719856	0.295941	ABC transporters
Urea	34	2	0.0219956	0.672865	Basal transcription factors
Urea	67	2	0.0361928	0.862827	Retinol metabolism
WBC in the Baso channel	69	5	0.0029994	0.147171	Bacterial invasion of epithel cells
WBC in the Baso channel	55	5	0.0223955	0.694861	Basal cell carcinoma
WBC in the Baso channel	14	3	0.0163967	0.617277	Terpenoid backbone biosynthesis

Measure	Number of genes in KEGG pathway	Number of intervals with genes in KEGG pathway	Empirical p-value	Corrected p-value	KEGG pathway name
WBC in the Baso channel	29	3	0.0385923	0.85163	Thyroid cancer

References

1. EURATRANS Consortium. http://www.euratrans.eu/.
2. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* **38**, 879-87 (2006).
3. Ensembl. Rat assembly and gene annotation. http://www.ensembl.org/Rattus_norvegicus/Info/Annotation - genebuild.
4. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
5. Kendler, K.S. & Greenspan, R.J. The nature of genetic influences on behavior: lessons from "simpler" organisms. *The American journal of psychiatry* **163**, 1683-94 (2006).
6. Falconer DS, M.T. *Introduction to quantitative genetics*, 180 p (London, 1996).
7. Michael Lynch, B.W. *Genetics and Analysis of Quantitative Traits*, 980 p. (Sunderland (Massachusetts), 1998).
8. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics* **9**, 255-66 (2008).
9. Sullivan, P.F., Neale, M.C. & Kendler, K.S. Genetic epidemiology of major depression: review and meta-analysis. *The American journal of psychiatry* **157**, 1552-62 (2000).
10. Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187-97 (2011).
11. Mackay, T.F. The genetic architecture of quantitative traits. *Annual review of genetics* **35**, 303-39 (2001).
12. Gottesman, II & Shields, J. A polygenic theory of schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* **58**, 199-205 (1967).
13. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1193-8 (2012).
14. Larsson, H. *et al.* Genetic and environmental influences on adult attention deficit hyperactivity disorder symptoms: a large Swedish population-based study of twins. *Psychological medicine* **43**, 197-207 (2013).
15. Maes, H.H., Neale, M.C. & Eaves, L.J. Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* **27**, 325-51 (1997).
16. Abney, M., McPeek, M.S. & Ober, C. Broad and narrow heritabilities of quantitative traits in a founder population. *American journal of human genetics* **68**, 1302-7 (2001).
17. Barker, J.S. Inter-locus interactions: a review of experimental evidence. *Theoretical population biology* **16**, 323-46 (1979).
18. Miller, R.H., Legates, J.E. & Cockerham, C.C. Estimation of Nonadditive Hereditary Variance in Traits of Mice. *Genetics* **48**, 177-88 (1963).

19. Wagner, G.P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature reviews. Genetics* **12**, 204-13 (2011).
20. Otto, S.P. Two steps forward, one step back: the pleiotropic effects of favoured alleles. *Proceedings. Biological sciences / The Royal Society* **271**, 705-14 (2004).
21. Davignon, J. Beneficial cardiovascular pleiotropic effects of statins. *Circulation* **109**, III39-43 (2004).
22. Mitchell, B.D. *et al.* Genetic Analysis of the IRS: Pleiotropic Effects of Genes Influencing Insulin Levels on Lipoprotein and Obesity Measures. *Arteriosclerosis, Thrombosis, and Vascular Biology* **16**, 281-288 (1996).
23. Meredith, W. A model for analyzing heritability in the presence of correlated genetic and environmental effects. *Behavior Genetics* **3**, 271-7 (1973).
24. Lyons, M.J. *et al.* Do genes influence exposure to trauma? A twin study of combat. *American journal of medical genetics* **48**, 22-7 (1993).
25. Plomin, R., DeFries, J.C. & Loehlin, J.C. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological bulletin* **84**, 309-22 (1977).
26. Kendler, K.S. & Baker, J.H. Genetic influences on measures of the environment: a systematic review. *Psychological medicine* **37**, 615-26 (2007).
27. Gerlai, R. & Csanyi, V. Genotype-environment interaction and the correlation structure of behavioral elements in paradise fish (Macropodus opercularis). *Physiology & behavior* **47**, 343-56 (1990).
28. Valdar, W. *et al.* Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959-84 (2006).
29. Vink, J.M. *et al.* Sex differences in genetic architecture of complex phenotypes? *PloS one* **7**, e47371 (2012).
30. Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744-50 (1986).
31. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
32. CF Genetic Analysis Consortium. http://www.genet.sickkids.on.ca/cftr.
33. Correlation between genotype and phenotype in patients with cystic fibrosis. The Cystic Fibrosis Genotype-Phenotype Consortium. *The New England journal of medicine* **329**, 1308-13 (1993).
34. Cutting, G.R. Modifier genes in Mendelian disorders: the example of cystic fibrosis. *Annals of the New York Academy of Sciences* **1214**, 57-69 (2010).
35. Palanca Suela, S. *et al.* CASP8 D302H polymorphism delays the age of onset of breast cancer in BRCA1 and BRCA2 carriers. *Breast cancer research and treatment* **119**, 87-93 (2010).
36. Estivill, X. Complexity in a monogenic disease. *Nature Genetics* **12**, 348-50 (1996).
37. Scriver, C.R. & Waters, P.J. Monogenic traits are not simple: lessons from phenylketonuria. *Trends in Genetics* **15**, 267-272 (1999).
38. King, M.-C., Marks, J.H., Mandell, J.B. & The New York Breast Cancer Study, G. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science* **302**, 643-646 (2003).
39. Ghoussaini, M., Pharoah, P.D.P. & Easton, D.F. Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning? *The American journal of pathology*.

40. Kajiwara, K., Berson, E.L. & Dryja, T.P. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* **264**, 1604-1608 (1994).
41. Katsanis, N. *et al.* Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* **293**, 2256-9 (2001).
42. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536-9 (1996).
43. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends in genetics : TIG* **17**, 502-10 (2001).
44. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
45. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
46. Hindorff, L., Junkins, H., Mehta, J. & Manolio, T. A catalog of published genome-wide association studies 2010. *Ref Type: Generic* (2011).
47. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362-7 (2009).
48. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
49. Visscher, P.M. Sizing up human height variation. *Nature Genetics* **40**, 489-90 (2008).
50. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-9 (2010).
51. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981-90 (2012).
52. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467-71 (2013).
53. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
54. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nature Genetics* **43**, 1127-30 (2011).
55. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**, 124-37 (2001).
56. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-43 (2008).
57. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
58. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).
59. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666-70 (2010).
60. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
61. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics* **10**, 392-404 (2009).
62. Phillips, P.C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics* **9**, 855-67 (2008).

63. Hunter, D.J. Gene-environment interactions in human diseases. *Nature reviews. Genetics* **6**, 287-98 (2005).
64. Evans, D.M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics* **43**, 761-7 (2011).
65. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics* **42**, 985-90 (2010).
66. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41**, 703-7 (2009).
67. Carrasquillo, M.M. *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genetics* **32**, 237-44 (2002).
68. Hemani, G., Knott, S. & Haley, C. An evolutionary perspective on epistasis and the missing heritability. *PLoS genetics* **9**, e1003295 (2013).
69. Zubenko, G.S., Hughes, H.B., 3rd & Stiffler, J.S. D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals. *Molecular psychiatry* **6**, 413-9 (2001).
70. Gray-McGuire, C. *et al.* Genome scan of human systemic lupus erythematosus by regression modeling: evidence of linkage and epistasis at 4p16-15.2. *American journal of human genetics* **67**, 1460-9 (2000).
71. Tsai, C.T. *et al.* Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis* **195**, 172-80 (2007).
72. Atkinson, A., Barbier, M., Afridi, S., Fumoux, F. & Rihet, P. Evidence for epistasis between hemoglobin C and immune genes in human P. falciparum malaria: a family study in Burkina Faso. *Genes and immunity* **12**, 481-9 (2011).
73. Caspi, A. *et al.* Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301**, 386-9 (2003).
74. Takeshita, T., Mao, X.Q. & Morimoto, K. The contribution of polymorphism in the alcohol dehydrogenase beta subunit to alcohol sensitivity in a Japanese population. *Human genetics* **97**, 409-13 (1996).
75. Weiss, L.A., Pan, L., Abney, M. & Ober, C. The sex-specific genetic architecture of quantitative traits in humans. *Nature Genetics* **38**, 218-22 (2006).
76. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *American journal of human genetics* **89**, 607-18 (2011).
77. Tishkoff, S.A. & Williams, S.M. Genetic analysis of African populations: human evolution and complex disease. *Nature reviews. Genetics* **3**, 611-21 (2002).
78. Mackay, T.F. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173-8 (2012).
79. Johannesson, M. *et al.* A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. *Genome Res* **19**, 150-8 (2009).
80. Churchill, G.A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* **36**, 1133-7 (2004).

81. Flint, J. Mapping quantitative traits and strategies to find quantitative trait genes. *Methods* **53**, 163-74 (2011).
82. Flint, J. & Mackay, T.F. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **19**, 723-33 (2009).
83. Flint, J., Valdar, W., Shifman, S. & Mott, R. Strategies for mapping and cloning quantitative trait genes in rodents. *Nature reviews. Genetics* **6**, 271-86 (2005).
84. Acevedo-Arozena, A. *et al.* ENU mutagenesis, a way forward to understand gene function. *Annual review of genomics and human genetics* **9**, 49-69 (2008).
85. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).
86. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-80 (1989).
87. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42**, 30-5 (2010).
88. Risch, N. Genetic linkage and complex diseases, with special reference to psychiatric disorders. *Genetic epidemiology* **7**, 3-16; discussion 17-45 (1990).
89. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).
90. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).
91. Saccone, S.F. *et al.* SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic acids research* **38**, W201-W209 (2010).
92. The ENCODE Project: ENCyclopedia Of DNA Elements. http://www.genome.gov/10005107.
93. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
94. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
95. Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annual review of biochemistry* **57**, 159-97 (1988).
96. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic acids research* **30**, 38-41 (2002).
97. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
98. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-63 (2009).
99. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
100. Rijnkels, M. *et al.* Epigenetic modifications unlock the milk protein gene loci during mouse mammary gland development and differentiation. *PloS one* **8**, e53270 (2013).
101. Helms, C. *et al.* A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. *Nature Genetics* **35**, 349-56 (2003).

102. McCarthy, M.I. & Hirschhorn, J.N. Genome-wide association studies: potential next steps on a genetic journey. *Human molecular genetics* **17**, R156-65 (2008).
103. Vivek Kumar, F.P.-M.d.V., Gary Churchill, Joseph S. Takahashi. QTL analysis utilizing closely related mouse substrains identifies Cytoplasmic FMRP Interacting Protein 2 (CYFIP2) as a regulator of cocaine response. in *Complex Trait Consoritum meeting* (Madison, Wisconsin, USE, 2013).
104. Yalcin, B., Flint, J. & Mott, R. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* **171**, 673-81 (2005).
105. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289-94 (2011).
106. Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nature Genetics* **41**, 677-87 (2009).
107. Wang, W., Fu, J.-F., Gong, F.-Q., Zhu, W.-H. & Shen, Z. Rare hypertension as a result of 17Œ±-hydroxylase deficiency. in *Journal of Pediatric Endocrinology and Metabolism* Vol. 24 333 (2011).
108. Hillebrandt, S. *et al.* Complement factor 5 is a quantitative trait gene that modifies liver fibrogenesis in mice and humans. *Nature Genetics* **37**, 835-43 (2005).
109. Wang, X. *et al.* Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nature Genetics* **37**, 365-72 (2005).
110. Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506-11 (2003).
111. Wang, X., Ishimori, N., Korstanje, R., Rollins, J. & Paigen, B. Identifying Novel Genes for Atherosclerosis through Mouse-Human Comparative Genetics. *The American Journal of Human Genetics* **77**, 1-15 (2005).
112. Korstanje, R. & DiPetrillo, K. Unraveling the genetics of chronic kidney disease using animal models. *American journal of physiology. Renal physiology* **287**, F347-52 (2004).
113. Wang, X. & Paigen, B. Genome-wide search for new genes controlling plasma lipid concentrations in mice and humans. *Current opinion in lipidology* **16**, 127-37 (2005).
114. Wang, X. & Paigen, B. Quantitative trait loci and candidate genes regulating HDL cholesterol: a murine chromosome map. *Arteriosclerosis, Thrombosis, and Vascular Biology* **22**, 1390-401 (2002).
115. Stoll, M. *et al.* New target regions for human hypertension via comparative genomics. *Genome Research* **10**, 473-82 (2000).
116. Sugiyama, F. *et al.* Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics* **71**, 70-7 (2001).
117. Klein, R.F. Genetic regulation of bone mineral density in mice. *Journal of musculoskeletal & neuronal interactions* **2**, 232-6 (2002).
118. de la Pompa, J.L. *et al.* Conservation of the Notch signalling pathway in mammalian neurogenesis. *Development* **124**, 1139-48 (1997).
119. Kyriakis, J.M. & Avruch, J. Mammalian Mitogen-Activated Protein Kinase Signal Transduction Pathways Activated by Stress and Inflammation. *Physiological Reviews* **81**, 807-869 (2001).
120. Smith, E.D. *et al.* Quantitative evidence for conserved longevity pathways between divergent eukaryotic species. *Genome Research* **18**, 564-70 (2008).

121. Jaglo, K.R. *et al.* Components of the Arabidopsis C-repeat/dehydration-responsive element binding factor cold-response pathway are conserved in Brassica napus and other plant species. *Plant physiology* **127**, 910-7 (2001).
122. Wagner, K.U. Models of breast cancer: quo vadis, animal modeling? *Breast cancer research : BCR* **6**, 31-8 (2004).
123. Claeys, I. *et al.* Insulin-related peptides and their conserved signal transduction pathway. *Peptides* **23**, 807-16 (2002).
124. Rao, M. Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. *Developmental biology* **275**, 269-86 (2004).
125. Baud, A. *et al.* Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nature Genetics* **45**, 767-75 (2013).
126. Saar, K. *et al.* SNP and haplotype mapping for genetic analysis in the rat. *Nature Genetics* **40**, 560-6 (2008).
127. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics* **43**, 648-55 (2011).
128. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nature Genetics* **24**, 23-5 (2000).
129. Jacob, H.J. *et al.* A genetic linkage map of the laboratory rat, Rattus norvegicus. *Nature Genetics* **9**, 63-9 (1995).
130. History of the Norway Rat (Rattus norvegicus). (2003, 2004).
131. Pravenec, M. *et al.* A genetic linkage map of the rat derived from recombinant inbred strains. *Mammalian genome : official journal of the International Mammalian Genome Society* **7**, 117-27 (1996).
132. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* **37**, 243-53 (2005).
133. Gibbs, R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
134. Heinig, M. *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467**, 460-4 (2010).
135. Aylor, D.L. *et al.* Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome research* **21**, 1213-22 (2011).
136. Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C. & Flint, J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 12649-54 (2000).
137. Hansen, C. & Spuhler, K. Development of the National Institues of Health genetically heterogeneous rat stock. *Alcoholism: Clinical and Experimental Research* **8**, 477-479 (1984).
138. Affymetrix. BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array. (2007).
139. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-75 (2007).
140. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
141. Ke, X. *et al.* The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human molecular genetics* **13**, 577-88 (2004).
142. Valdar, W., Holmes, C.C., Mott, R. & Flint, J. Mapping in structured populations by resample model averaging. *Genetics* **182**, 1263-77 (2009).

143. Evans, A.L., Brown, W., Kenyon, C.J., Maxted, K.J. & Smith, D.C. Improved system for measuring systolic blood pressure in the conscious rat. *Medical & biological engineering & computing* **32**, 101-2 (1994).
144. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
145. Bennett, B.J. *et al.* A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome research* **20**, 281-90 (2010).
146. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics* **2**, e41 (2006).
147. Hayes, B.J., Visscher, P.M. & Goddard, M.E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research* **91**, 47-60 (2009).
148. Solberg, L.C. *et al.* A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mammalian genome : official journal of the International Mammalian Genome Society* **17**, 129-46 (2006).
149. Fiddy, S., Cattermole, D., Xie, D., Duan, X.Y. & Mott, R. An integrated system for genetic analysis. *BMC Bioinformatics* **7**, 210 (2006).
150. Venables, W.N.R., B. D. *Modern Applied Statistics with S*, (New York, 2002).
151. Mousseau, T.A. & Roff, D.A. Natural selection and the heritability of fitness components. *Heredity* **59 (Pt 2)**, 181-97 (1987).
152. Abecasis, G.R., Cardon, L.R. & Cookson, W.O. A general test of association for quantitative traits in nuclear families. *American journal of human genetics* **66**, 279-92 (2000).
153. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American journal of human genetics* **52**, 506-16 (1993).
154. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *American journal of human genetics* **67**, 170-81 (2000).
155. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
156. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature methods* **8**, 833-5 (2011).
157. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-54 (2010).
158. Hu, Q. *et al.* The Yak genome database: an integrative database for studying yak biology and high-altitude adaption. *BMC genomics* **13**, 600 (2012).
159. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* **44**, 825-30 (2012).
160. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82 (2011).
161. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-31 (2008).
162. Manolio, T.A., Brooks, L.D. & Collins, F.S. A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation* **118**, 1590-605 (2008).

163. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).
164. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241-7 (1995).
165. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-71 (1994).
166. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356-69 (2008).
167. Chanock, S.J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).
168. Ioannidis, J.P. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).
169. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670-2 (1999).
170. Melchinger, A.E., Utz, H.F. & Schon, C.C. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* **149**, 383-403 (1998).
171. Rat Genome Database. http://rgd.mcw.edu/.
172. Flint, J. & Mott, R. Finding the molecular basis of quantitative traits: successes and pitfalls. *Nature reviews. Genetics* **2**, 437-45 (2001).
173. Kover, P.X. *et al.* A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS genetics* **5**, e1000551 (2009).
174. Baud, A. *et al.* Genome sequencing and genetic mapping to dissect the genetic basis of complex traits. *M S-Medecine Sciences* **29**, 671-674 (2013).
175. Israely, I. *et al.* Deletion of the neuron-specific protein delta-catenin leads to severe cognitive and synaptic dysfunction. *Current biology : CB* **14**, 1657-63 (2004).
176. Berkel, S. *et al.* Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nature Genetics* **42**, 489-91 (2010).
177. Wallstr√∂m, E. & Olsson, T. Rat Models of Experimental Autoimmune Encephalomyelitis. in *Sourcebook of Models for Biomedical Research* (ed. Conn, P.M.) 547-556 (Humana Press, 2008).
178. Koch, M.A. *et al.* The transcription factor T-bet controls regulatory T cell homeostasis and function during type 1 inflammation. *Nature immunology* **10**, 595-602 (2009).
179. Shirihai, O.S., Gregory, T., Yu, C., Orkin, S.H. & Weiss, M.J. ABC-me: a novel mitochondrial transporter induced by GATA-1 during erythroid differentiation. *The EMBO journal* **19**, 2492-502 (2000).
180. Hyde, B.B. *et al.* The mitochondrial transporter ABC-me (ABCB10), a downstream target of GATA-1, is essential for erythropoiesis in vivo. *Cell death and differentiation* **19**, 1117-26 (2012).
181. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
182. Veyrieras, J.B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics* **4**, e1000214 (2008).
183. Hinrichs, A.S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590-8 (2006).

184. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**, 29-34 (1999).
185. Carlson, M., Falcon, S., Pages, H. & Li, N. KEGG. db: A set of annotation maps for KEGG. *R package version* **2**(2009).
186. Lee, P.H., O'Dushlaine, C., Thomas, B. & Purcell, S.M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797-9 (2012).
187. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
188. Adkins, R.M., Gelke, E.L., Rowe, D. & Honeycutt, R.L. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Molecular biology and evolution* **18**, 777-91 (2001).
189. Springer, M.S., Murphy, W.J., Eizirik, E. & O'Brien, S.J. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1056-61 (2003).
190. Halligan, D.L., Oliver, F., Eyre-Walker, A., Harr, B. & Keightley, P.D. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS genetics* **6**, e1000825 (2010).
191. Ferreira, M.A. *et al.* Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. *American journal of human genetics* **86**, 88-92 (2010).
192. Yalcin, B. *et al.* Commercially available outbred mice for genome-wide association studies. *PLoS genetics* **6**, e1001085 (2010).
193. Ziegler, A. *et al.* Mutation hotspots due to sunlight in the p53 gene of nonmelanoma skin cancers. *Proc Natl Acad Sci U S A* **90**, 4216-20 (1993).
194. Klein, J., Satta, Y., O'HUigin, C. & Takahata, N. The molecular descent of the major histocompatibility complex. *Annu Rev Immunol* **11**, 269-95 (1993).
195. Segurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci U S A* **109**, 18493-8 (2012).
196. Leffler, E.M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578-82 (2013).
197. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
198. International Multiple Sclerosis Genetics, C. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-9 (2011).
199. Mells, G.F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **43**, 329-32 (2011).
200. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
201. Pelak, K. *et al.* Host determinants of HIV-1 control in African Americans. *J Infect Dis* **201**, 1141-9 (2010).
202. Matsuda, M. *et al.* Involvement of phospholipase C-related inactive protein in the mouse reproductive system through the regulation of gonadotropin levels. *Biol Reprod* **81**, 681-9 (2009).
203. Mizokami, A. *et al.* Phospholipase C-related inactive protein is involved in trafficking of gamma2 subunit-containing GABA(A) receptors to the cell surface. *J Neurosci* **27**, 1692-701 (2007).

204. Takenaka, K. *et al.* Role of phospholipase C-L2, a novel phospholipase C-like protein that lacks lipase activity, in B-cell receptor signaling. *Mol Cell Biol* **23**, 7329-38 (2003).
205. Nie, H., Maika, S.D., Tucker, P.W. & Gottlieb, P.D. A role for SATB1, a nuclear matrix association region-binding protein, in the development of CD8SP thymocytes and peripheral T lymphocytes. *J Immunol* **174**, 4745-52 (2005).
206. Nie, H., Yao, X., Maika, S.D. & Tucker, P.W. SATB1 is required for CD8 coreceptor reversal. *Mol Immunol* **46**, 207-11 (2008).
207. Zhang, X. & Firestein, S. The olfactory receptor gene superfamily of the mouse. *Nat Neurosci* **5**, 124-33 (2002).
208. Pal, C. & Hurst, L.D. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33**, 392-5 (2003).
209. Nei, M. Genome evolution: let's stick together. *Heredity (Edinb)* **90**, 411-2 (2003).
210. Birney, E. Medaka Genetic Reference Panel.
211. Williams, R.W., Gu, J., Qi, S. & Lu, L. The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome biology* **2**, RESEARCH0046 (2001).
212. Craddock, N., O'Donovan, M.C. & Owen, M.J. Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes. *Molecular psychiatry* **13**, 649-53 (2008).
213. Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nature Genetics* **44**, 297-301 (2012).
214. Schmidts, M. *et al.* Combined NGS approaches identify mutations in the intraflagellar transport gene IFT140 in skeletal ciliopathies with early progressive kidney Disease. *Human mutation* **34**, 714-24 (2013).
215. Peden, J.F. & Farrall, M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Human molecular genetics* **20**, R198-205 (2011).
216. Barcellos, L.F. *et al.* Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. *Human molecular genetics* **15**, 2813-24 (2006).
217. Svenson, K.L. *et al.* High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* **190**, 437-47 (2012).
218. King, E.G., Macdonald, S.J. & Long, A.D. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. *Genetics* **191**, 935-49 (2012).
219. Holmans, P. *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American journal of human genetics* **85**, 13-24 (2009).
220. Beissbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-5 (2004).
221. Goodson, M. *et al.* Cofilin-1: a modulator of anxiety in mice. *PLoS genetics* **8**, e1002970 (2012).
222. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics* **81**, 1278-83 (2007).
223. Miller, J.A., Horvath, S. & Geschwind, D.H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of*

the National Academy of Sciences of the United States of America **107**, 12698-703 (2010).

224. Soccio, R.E. *et al.* Species-specific strategies underlying conserved functions of metabolic transcription factors. *Molecular endocrinology* **25**, 694-706 (2011).
225. Suthram, S., Sittler, T. & Ideker, T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* **438**, 108-12 (2005).
226. Gandhi, T.K. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics* **38**, 285-93 (2006).