

Genetic analysis of variation in complex traits

Nick Groves-Kirkby

Linacre College, University of Oxford

Submitted for the degree of *Doctor of Philosophy*, Hilary Term 2016

Abstract | Variation is a universal property of life, and much of contemporary genetics research is directed towards understanding the causes of variation in traits. Here I present the results of my investigations into the genetic and other causes of trait variation in humans and mice. I address these questions in the context of two distinct research projects, which use pre-existing data to investigate the causes of trait variation, through a range of analytic techniques.

I first extract trait data from historic breeding records from the incipient Collaborative Cross (a genetic reference population of recombinant inbred mice) and use them to map genetic factors affecting litter size and other reproductive traits. Mapping reveals significant quantitative trait loci associated with litter size and time between litters, as well as a number of suggestive loci. I characterise the genetic effects at these loci and investigate candidate genes. The most robust finding, a litter size locus on chromosome 5, explains around 3% of observed variation and 24% of the variation attributable to genetics.

Using data obtained from the Netherlands Twin Registry—a longitudinal database of Dutch twins—I investigate the prevalence of parent of origin effects on gene expression traits in peripheral blood in humans. I first phase individuals' genotypes by parental origin and use these genotypes to calculate the heritability of over 44,000 gene expression traits partitioned into that attributable to matching and nonmatching parent of origin. I replicate prior genomewide heritability estimates for many traits, but I find little evidence of widespread parent-of-origin effects on human gene expression in blood. On further examination, the small sample size severely limits the power to detect such effects. Nonetheless, I identify approximately 200 genes enriched for immune system processes that show evidence of parent-of-origin-specific effects on heritability.

Genetic analysis of variation in complex traits

Nick Groves-Kirkby



Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2016

To Anna

Contents

Chapter 1 Introduction	2
Variation and genetics	2
Genetics and data	5
Genetic analysis of reproduction in Collaborative Cross mice	6
Parent-of-origin effects on gene expression in humans	7
Common themes	8
 Chapter 2 Reproduction in the Collaborative Cross	 10
Reproduction in mammals	10
Definitions	11
Genetics of reproduction	12
The mouse as a model organism	14
Reproduction in mice	16
Hybrid sterility	17
The Collaborative Cross	19
Breeding scheme	21
Mapping and other studies	21
Breeding performance	28
 Chapter 3 Collaborative Cross breeding records	 31
The CC at Tel Aviv University	31
Data structure	35

Cousin lines	36
Data cleaning	37
Exploratory data analysis	44
Initial observations	45
Litter size	45
Calculated dates	48
Suitability for mapping	51
Covariates	55
Inbreeding	57
Breeding histories	58
First litters	60
Cousins	60
Summary	64
Genotypes	67
Genotyped CC lines	68
Unrecorded cousins	69
Chapter 4 Genetic analysis of reproductive traits	73
Genetic analysis of complex traits	73
Heritability	74
Broad-sense heritability	74
Genetic coefficient of variation	76
Narrow-sense heritability	78

Genetic mapping of reproductive traits	79
Linear mixed effect models	80
QTL mapping	82
QTL mapping for litter size	86
Modeling generation effects	89
Effects of subpopulation	91
QTL mapping for time between litters	94
QTL mapping for other traits	94
Summary	97
Genetic effects	101
Chromosome 5 locus	101
Chromosome 1 locus	104
Chromosome 17	104
Time between litters	104
Candidate genes	108
Discussion	110
Results	110
Phenotype	118
Inbreeding depression	121
CC population	122
Data quality	124
Summary	126

Chapter 5 Parent-of-origin effects	127
The epigenetic basis of variation	127
Parent-of-origin effects	129
Examples	131
Genomic imprinting	132
Other parent-of-origin mechanisms	136
Detecting parent-of-origin effects	137
Human gene expression	138
 Chapter 6 Netherlands Twin Register data	 140
NTR cohort	140
NTR research	142
NTR genotype data	144
Expression data	150
Covariates	150
 Chapter 7 Parent of origin effects on gene expression	 163
Phasing NTR genotype data	163
Kinship	171
Method	173
Family structure	178
Heritability of gene expression	178
Total heritability	180

Parent-of-origin heritability	182
Imprinted genes	188
Discussion	192
Phasing	193
Kinship	194
Heritability	195
Strong <i>cis</i> effects	196
Context	197
Chapter 8 Conclusions and perspectives	199
Outcomes	199
Further work	200
Critical perspectives	201
Data in biological research	202
Reproducibility	203
Postscript	205
References	206

Rule 3: "The Truth Can Wait"

Chapter 1 | Introduction

One can state, without exaggeration, that the observation of and the search for similarities and differences are the basis of all human knowledge.

—Alfred Nobel, from his ‘black notebook’ (unpublished)

Variation and genetics

Variation—loosely defined as a measure of the similarities and differences between individuals, cells or groups—is a universal property of life, and a central pillar of evolutionary biology. We see great variation in all manner of traits between species, and within species. Variation is so ubiquitous that it is frequently taken for granted. We very rarely question why some humans are taller than others—we just understand that humans are different sizes. We seldom ask how we comprehend the differences between humans and dogs—we just understand that dogs are dogs, and that they are not humans.

In certain situations, however, we have great incentive to understand biological variation. One important example is in the practice of medicine. We frequently see differences in disease severity or response to treatment between otherwise similar individuals. Understanding the causes of such differences would allow the identification of high-risk patients and the selection of appropriate therapies. Similarly, understanding the causes of variation in animals and crops would contribute to agricultural production. Understanding the causes of human trait variation should also make an important contribution to how we frame discussions of history, identity, and equality,¹⁻⁴ and to our understanding of cultural activities, such as sporting performance.⁵

While individuals show variation across a range of traits, related individuals tend to be more similar than are nonrelated individuals. We intuitively understand that offspring resemble their parents, and that siblings are somewhat alike and twins more so. From the

work of early geneticists, most notably Gregor Mendel, we know that such resemblances are attributable to the action of paired heritable factors—genes—which are transmitted one from each parent. For some traits, a few such genetic factors are sufficient to explain observed variation. Classical examples of such Mendelian inheritance include Mendel’s canonical experiment with peas,⁶ and the autosomal recessive pattern of inheritance of sickle cell disease.⁷

The Human Genome Project, completed in 2003, provides a haploid reference genome and means of cataloging variation among humans at the genetic level.⁸ Around 20,000 genes are identified, fewer than generally expected and probably too few to support a one-gene–one-trait paradigm. Sequencing the human genome has enabled much progress in the understanding of genetic contributions to hereditary disease and other traits.⁹ Nevertheless, population-based mapping approaches such as genome-wide association studies (GWAS), which aim to identify common single nucleotide polymorphisms associated with specific traits, have not yet identified individual genes with effects sufficient to explain all observed variation in the majority of traits.^{10–13} Encouragingly, efforts are now starting to shed light on the whereabouts of this missing heritability.^{14–18}

Many traits do not exhibit simple Mendelian patterns of inheritance (an observation made by Mendel himself in studies of other plants and traits⁶). In fact, variation in most traits seems to arise from the actions of many genes (i.e. hundreds to thousands).^{19,20} Contemporary genetic techniques allow the mapping of only the most powerful of these effects. In addition, environmental effects (broadly defined to include any effect on trait not caused by genetic factors, and in this case including epigenetic effects) contribute to the observed trait variation.²¹ Inter-individual variation in these complex traits is nonetheless attributable to biologic processes, and thus represents a domain of study for geneticists.

Nowadays, we understand that individuals show considerable variation at the genetic level, in the form of mutations, epigenetic marks and structural variants. Such variation

propagates into downstream processes (e.g. transcription and translation), and ultimately presents as observed trait variation. The processes involved are complex, involving multiple factors, mechanisms and interactions. Environmental effects also seem to have an important role. A number of projects aim to catalog and make sense of human genetic variation (e.g. Encyclopedia of DNA Elements [ENCODE], 1000 Genomes Project, International HapMap Project, and Human Variome Project).^{22–25} Nevertheless, much about the processes involved remains unknown, both in terms of their general nature and in terms of specific mechanisms and effects.

While we often think of variation as a property of individuals (e.g. height or disease susceptibility), in fact multicellular organisms can show considerable variation in form and function at the cellular level (~ 200 distinct cell types in humans; estimated 3.72×10^{13} total cells²⁶). Such variation is in some ways remarkable, because within a given individual each of these cell types is the product of an identical genetic code. Within each cell type, a subset of genes are typically expressed, specific to cellular function; such expression is often dynamic and responsive. The factors governing how the genetic code is processed and read—epigenetic marks and modifications—are also crucial determinants of observed trait variation. Epigenetic processes can be modified by environmental factors, with the caveat that environmental factors might be external to the individual, or a manifestation of complex internal interactions. Epigenetics is a burgeoning field of research, and the full extent of the effects of epigenetic processes on trait variation is yet to be defined.²⁷

Thus, much remains unknown about the causes of inter-individual variation. Genetic factors are an important contributor; however, very few phenotypes conform to the textbook example of a ‘gene for’ a trait. Unravelling the cause of variation in a trait is certainly aided by identifying the genetic factors involved, but is also dependent on understanding how they function and interact with each other and with a broadly defined environment. Much of contemporary genetics research is in one way or another involved in addressing such questions. In this thesis I report research that builds on these approaches,

in the context of a number of high-level questions:

- To what extent and under what circumstances do genetic factors explain observed trait variation?
- What specific genetic (and other) factors explain observed trait variation?
- By what mechanisms do these genetic factors affect trait variation?
- Are these mechanisms predictable and generalizable across traits, individuals, and populations?
- To what extent is knowing about these factors useful for medical and other applications?

I present here the results of my investigations into the genetic and other causes of trait variation in humans and mice. I address these questions in the context of two distinct research projects, which use pre-existing data to investigate the causes of trait variation, through application of a range of analytic techniques. Below I provide brief outlines of these projects and their major outcomes. While this is in many ways a thesis in two parts, I endeavour to identify links and highlight common themes throughout.

Genetics and data

The other notable focus of this thesis is the use of data in genetic research. The past years have seen an immense increase in the quantity of data available to a researcher in genetics. An assumption exists—probably correctly—that more data leads to better research, both in terms of resolution and robustness. Furthermore, genetic and other data are widely believed to be the key that unlocks a new wave of innovation in health care and related technologies. Such ideas are held not just among researchers and clinicians, but are also gaining traction among the press and public.^{28–33}

The obvious source of the growth of data in genetics research is the increasing availability of sequencing data, and the rapid decreases in the cost of obtaining such sequences. A related development is an increase in the availability of high-powered computing clusters, which enables analysis of large data sets, and in cloud and server methods for data storage. Sequence data is complemented by other large data sets, including those reporting gene and protein expression. Analyses run on such data generate additional datasets in the form of results and intermediaries. Finally, published articles and results comprise a growing body of data in themselves, as do curated resources such as the ENSEMBL and UCSC genome browsers.^{34,35}

The growth in genetic and related data offers the opportunity for novel research and discovery. The challenge in this setting is to identify good, well-documented data resources that are suitable for addressing important research questions. In this regard, comparatively little published guidance on issues of data management and good practice exists in the genetics literature. As the research reported in this thesis relies primarily on computational methods and data analysis, I take the opportunity to comment on some of these pertinent issues.

Genetic analysis of reproduction in Collaborative Cross mice

The first part of this thesis investigates the factors affecting reproductive traits—chiefly in the form of litter size (i.e. fecundity)—in inbred mice from the Collaborative Cross genetic reference population. Reproduction is a fundamental property of life; yet we see dramatic differences in reproductive strategies between species, including among mammals. We also observe differences in reproductive performance between individuals within a species. An understanding of the genetic and other factors affecting mammalian reproduction would potentially contribute to a number of domains, including improvement of agriculture and animal husbandry, treatment of infertility, development of novel contraceptives, and

government policy in multiple areas.

Inbred mouse lines are a cornerstone of genetic research. Individuals within an inbred line are (theoretically) genetically identical, allowing any differences within and between lines to be investigated. The Collaborative Cross population is a collection of recombinant inbred lines with known founders designed to enable trait mapping and systems genetics approaches. I extract trait data from historic breeding records from the incipient Collaborative Cross lines (Chapter 2) and use them to map genetic factors affecting litter size and other reproductive traits (Chapter 3). Mapping reveals significant quantitative trait loci associated with litter size and time between litters, as well as a number of suggestive loci for litter size and other traits. I then apply a number of techniques to dissect these loci, to characterise the effect sizes and to identify candidate genes (Chapter 4). The most robust finding, a litter size locus on chromosome 5, explains around 3% of observed variation and 24% of the variation attributable to genetics.

Parent-of-origin effects on gene expression in humans

The second part of this thesis discusses parent-of-origin effects on gene expression in humans. While it is possible to identify genetic factors that show association with specific traits, many complex traits show substantial ‘missing heritability’. A number of possible explanations for this missing heritability are suggested, including epigenetic effects. One particular class of epigenetic effects is parent-of-origin effects, in which the parental origin of an allele (i.e. maternal or paternal) is associated with its contribution to a given trait. Such effects have been observed in a number of species and traits but it remains an open question how widespread they are, and what their contribution is to observed trait variation.

I first provide a review of the current state of knowledge on parent-of-origin effects (Chapter 5). Using data obtained from the Netherlands Twin Registry—a longitudinal database

comprising approximately 50% of Dutch twins (Chapter 6)—I investigate the prevalence of parent of origin effects on gene expression traits in peripheral blood in humans. I first identify individuals for whom both parental genotypes are available, and use these data to phase the individuals' genotypes by parental origin. I then calculate the heritability of over 44,000 gene expression traits partitioned into that attributable to matching and nonmatching parent of origin (Chapter 7). I replicate prior genomewide heritability estimates for many traits, but I find little evidence of widespread parent-of-origin effects on human gene expression in blood. On further examination, the small sample size severely limits the power to detect such effects as a result of comparatively low variance in kinship between unrelated individuals. Nonetheless, I identify approximately 200 genes enriched for immune system processes that show evidence of parent-of-origin-specific effects on heritability.

Common themes

These two projects share several common biological themes, including relationships between generations, the effects of genotype on observed phenotype and phenotypic variation, and conflicts between parental genomes. Of these possibilities I have chosen to focus my narrative on the topic of the genetic and other factors affecting trait variation, as such questions have implicitly and explicitly guided the design, execution and interpretation of the research. Nonetheless, throughout this thesis I will endeavour to highlight areas of overlap between the two projects as they become apparent.

Both projects also rely heavily on data and computational analysis. The application of data to investigate complex problems has been a common theme throughout my doctoral work, and thus a notable component of this thesis will be a discussion of the role of data and other resources in the contemporary life sciences and beyond. Much of this discussion can be found in the concluding chapter (Chapter 8).

Finally, in the spirit of good research practice, I have made every effort to ensure that the work reported here reflects open, reproducible research. All data analysis is done using the open-source language R (<https://www.r-project.org>). The thesis itself is written in R Markdown (<http://rmarkdown.rstudio.com/>) and the code used to generate the vast majority of figures and analyses is embedded within the raw text file. On occasion, code blocks are presented in the final text when they enhance understanding. A final electronic version will be deposited in the Oxford University Research Archive (<http://ora.ouls.ox.ac.uk/>) and all relevant files will be available via an open repository on GitHub (<https://github.com/>). I evaluate my experiences with such open reproducible research in the concluding chapter (Chapter 8).

Chapter 2 | Reproduction in the Collaborative Cross

It is certainly true that the natural scheme for preservation of the species discourages incest in an extremely effective manner. . . For stragglers able to survive these hazards, a barrier of sterility looms as the tenth generation of inbreeding is approached. Few scientists seriously believed that a viable strain of inbreds could be continued, even if briefly achieved.

—Leonell C Strong, in *Origins of Inbred Mice* (1978) Ed. Herbert C Morse III³⁶

Reproduction in mammals

Reproduction is a fundamental property of life, and is central to determining evolutionary fitness. Among mammals sexual reproduction is the norm, with each zygote inheriting one set of chromosomes from each parent. Sexual reproduction maintains trait variation in a population, necessary for the action of natural selection. Within *Mammalia*, monotremes, marsupials and placentals have notable differences in reproductive strategy, and substantial variation occurs between species within each clade. As the investigations in this thesis primarily involve mice and humans, I focus the discussion on placental mammals.

Successful mammalian reproduction depends on the correct function of multiple biological processes, including sex determination of the parents (beginning in embryonic development), sexual maturation, gametogenesis, endocrine function, sexual behaviour and physical function, fertilization, early embryonic development of the offspring, gestation, parturition, lactation, and neonatal care. These processes occur over what can be a substantial period of time, and they interact with environmental stimuli. Reproduction in mammals is, therefore, undoubtedly a complex trait, comprising multiple interacting processes.

Placental mammals show marked inter-species differences in reproductive traits including litter size, gestation time, birth weight, and investment in postnatal care. Variation in

such traits also exists within species, including humans and mice. Although heritability of reproductive traits is generally reported to be lower than other quantitative traits, much variation remains heritable and attributable to genetic factors.^{37–40}

Mammalian reproduction is studied in diverse contexts, including health care (i.e. infertility and contraception; often shared between gynecological and andrological specialties), animal husbandry, conservation, and ethics. Research approaches include clinical sciences, developmental biology, epigenetics, molecular biology and genomics. While much published work exists in each of these disciplines, articles tend to be domain-specific with comparatively little overlap between disciplines. Thus, while much information on the factors affecting mammalian reproduction exists, it is not well synthesised and curated.

Definitions

Some confusion exists as to how precisely to describe reproductive capacity, particularly in the uses of the terms ‘fertility’ and ‘fecundity’. In fact, failure to adequately define these terms at the outset of this project resulted in ambiguous scope and poorly defined project goals, which I discuss in more detail in Chapter 8. Throughout this thesis I endeavor to adhere to a strict biological definition, whereby fecundity is a measure of the reproductive output of an individual (e.g. the number of offspring produced), and fertility is a measure of the potential for reproduction (e.g. fertile versus infertile, and grades in between). In human demography, however, the two terms seem to have broadly acquired the opposite meaning, and day-to-day they are often used interchangeably.

In the context of mammalian reproduction traits such as litter size or total offspring production can be regarded as a measurement of the fecundity of an individual or mating, which can be representative of the fertility of an individual, mating or inbred line. Other measurable traits such as gestation time or time to first litter are reproductive traits, also representative of fertility.

How best to measure and define reproductive phenotypes to enable their investigation remains an open question. Simple measures such as litter size are appealing, as they are easy to measure. However, as highlighted above, litter size represents the cumulative action of multiple biological and environmental processes. It is likely to be something of a crude tool for identifying specific factors associated with these processes. By contrast, biologically relevant phenotypes such as ovulation rate require more resources to evaluate, and often destruction of the organism under investigation.

Genetics of reproduction

Mammalian reproduction is a complex trait, representing the outcome of many different biological processes. Several thousand gene products are ultimately suggested to be involved in reproductive success. To date hundreds of genes are implicated in these processes and shown to be associated with reproductive phenotypes. Many of these genes are associated with developmental abnormalities, or with male factors—spermatogenesis is a remarkably complex process with ~4% of mammalian genes specifically expressed in the male germline.⁴¹ Thus, many such genes are identified in the context of sometimes extreme reproductive defects, rather than in the context of continuous natural variation in reproductive traits. These findings have been extensively reviewed elsewhere.^{42–45}

In the clinical setting, research similarly focuses on disorders of reproduction, typically in couples who present with infertility. The first genes identified to affect reproduction were associated with sexual differentiation (e.g. *SRY* in Turner and Klinefelter Syndromes), or hormone production and sensitivity (e.g. *GnRH*, *KAL1*, *LHB*, and *FSHB* in disorders of the hypothalamic–pituitary–gonadal axis). Subsequent investigations using mouse knockout models have revealed over 500 null mutations associated with reproductive processes; however, while the action of hormonal and sexual differentiation genes is comparable across species, few of the mouse genes necessary for gametogenesis have known mutations in their human orthologs.⁴⁵

More-recently, candidate gene and genome-wide association studies in women have identified loci associated with endometriosis, uterine fibroids, polycystic ovary syndrome and premature ovarian failure, as well as age at menarche and menopause.^{46,47} These loci follow the expected pattern of many genes with small effect sizes. Similarly, small genomewide studies have reported loci associated with male fertility.^{48–51} Nonetheless, despite the many genes linked to fertility traits, clinically relevant mechanisms and interventions are not well defined; most clinical presentations still result in the diagnosis of idiopathic infertility.

While male factors account for the majority of genes implicated in reproductive processes, the contributions of male and female factors to clinical infertility are about equal,⁵² and female factors are reported to have much greater impact on overall determination of reproductive traits. This apparent contradiction makes some sense in the context of the biology of reproduction—male factors can be expected to affect the success or failure of conception, but female biology governs ovulation, gestation, parturition and other reproductive processes. The genetic factors so far associated with female reproduction are perhaps less-extensively reviewed^{53–55} than are the genetics of male factor infertility.^{52,56–65}

Knowledge of normal reproductive function is clinically relevant, for preventative approaches to treating infertility, for family planning and for the application of contraceptive therapies. In this regard, large-scale genomewide studies of nonpathogenic human reproductive traits are somewhat impractical, owing to long generation times and the invasive nature of clinical investigations. While many of the underlying pathways are well known, better understanding of the specific genetic and other factors that affect these traits might better inform clinical decision making and therapeutic discovery.

One domain in which investigation into the factors affecting quantitative trait variation has been done is animal husbandry and agriculture. Productivity is important to the practice of agriculture, and the genetics of reproductive traits have been investigated in several species, chiefly cows, pigs, goats and sheep. In particular, studies in sheep

have identified genes involved in the transforming growth factor beta superfamily (*GDF9*, *BMP15* and *BMPRIB*), which are associated with offspring production through their effects on ovulation rates.⁶⁶ These genes are since shown to have similar effects in human and mouse reproduction. While genomewide association studies and other approaches have identified regions associated with reproductive traits, better predictive markers of reproductive output are required as fecundity is often below the maximum attainable level, or has declined owing to selection for milk or wool production.

The mouse as a model organism

The house mouse (*Mus musculus*) is the most popular mammalian model organism for the study of health and disease genetics. Mice breed quickly, are relatively cheap, tolerate inbreeding well (at least compared to other mammals), and exhibit a wide range of genetic and phenotypic diversity, similar to humans. Mice are commensal with humans, with a worldwide distribution comprising several subspecies (Figure 1). They have been kept as pets since as long as 3,000 years ago, and have been used in biomedical research since the early 16th Century and in genetic research since the rediscovery of Mendel’s laws in 1900.⁶⁷

A key breakthrough in mouse genetic research was the development of the first fully inbred lines in the 1920s. Such lines facilitated reproducible experiments and enabled investigation of the observed genetic linkage of traits.⁶⁹ Inbred strains are now defined as having undergone strict brother–sister mating for at least 20 consecutive generations, resulting in ~99% homozygosity.⁷⁰ Most classical laboratory inbred mouse strains are derived from pet-shop and fancy mice, which were often bred without specific protocol; selected for good health, ease of handling and docility; and subject to a range of bottleneck effects. In fact, classical laboratory strains are all derived from no more than a few *M. m. domesticus* females, with inter-strain polymorphism contributed by six *M. m. musculus*

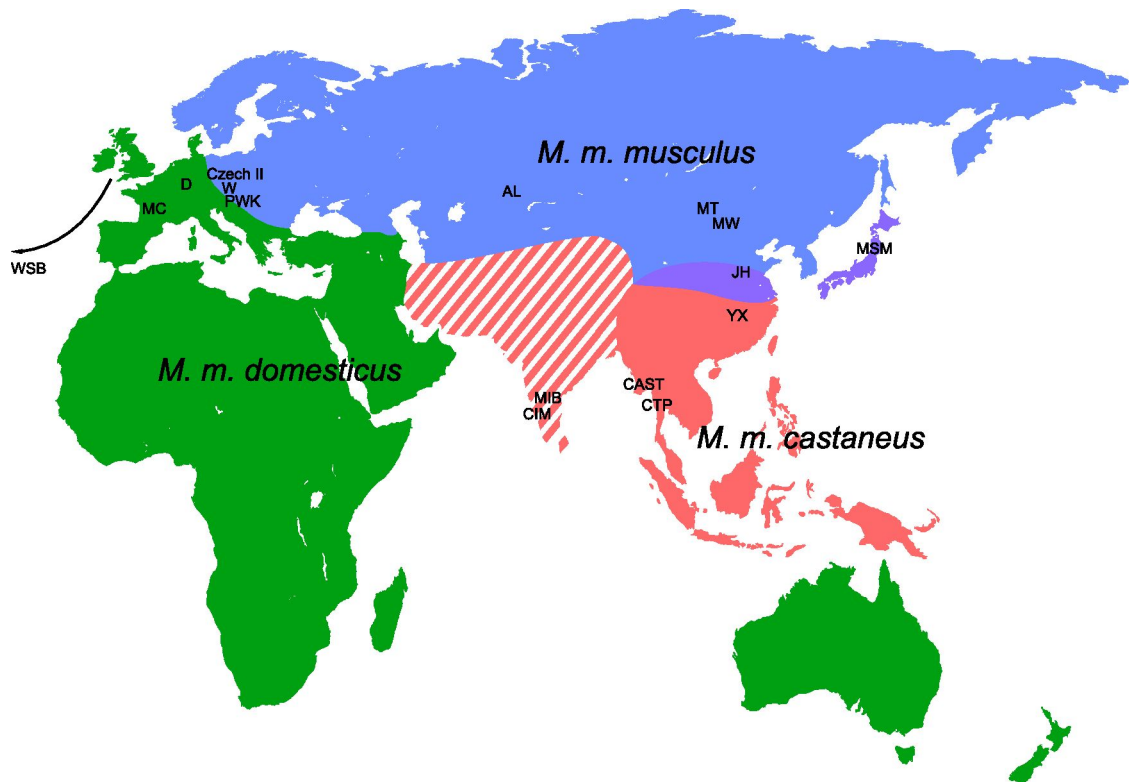


Figure 1: Worldwide distribution of mouse subspecies. Colour code relates to the individual subspecies (green—*M. m. domesticus*, blue—*M. m. musculus*, red—*M. m. castaneus*). Purple colour indicates the hybrid zone of *M. m. musculus* and *M. m. castaneus*. A much narrower hybrid zone of *M. m. domesticus* and *M. m. musculus* exists in central Europe. The definition of *M. m. castaneus* in the Indian subcontinent is complex. © 2013 Lilue *et al.*. Modified with permission under CC-BY.⁶⁸

males. A number of resources document the unusual origins of modern laboratory mice.^{36,67,71–74}

The Jackson Laboratory currently offers around 200 inbred strains to researchers, and many more have been generated, from varying origins and with diffuse characteristics.^{75–77} The mouse is now by a distance the mammal with the greatest number of genetically distinct laboratory strains. Notably, while the origins—and many contemporary uses—of inbred mice lie in investigating the genetic causes of disease, the creation and maintenance of inbred stocks provides insight into fundamental biological processes such as those surrounding reproduction, inbreeding and speciation.

Over time, a number of mouse genetic resources have been developed for trait mapping, including recombinant inbred panels, advanced intercross lines and heterogeneous stocks. These resources have typically been derived from classical inbred strains, which show limited allelic diversity. They have facilitated trait mapping in mice, but variously suffer from lack of phenotypic diversity, widespread linkage and limited allelic diversity in regions of the genome. A key consequence of these limitations is low mapping precision. Novel resources such as the Collaborative Cross (see below) and Diversity Outbred populations have been developed to address these issues.⁷⁸

Reproduction in mice

The process of reproduction in mice is generally similar to that of other mammals. As a result of the long history of breeding mice for research, details of reproductive processes in the mouse are well characterized and documented.^{79,80} Mice have a high reproductive potential, with a gestation period of 18–21 days, multiple pups born per litter, immediate postpartum estrus, and a generation time of 8–12 weeks. The mouse thus represents a good candidate for the investigation of reproductive processes. Females produce multiple litters in their lifetime, typically up to 5 in inbred strains and up to 10 when not inbred.

Inbred females show greatly reduced fecundity from the age of 8–10 months.⁸⁰

Reproductive traits investigated in mice include litter size, ovarian function, gestation time, ovulation rate and response to gonadotropins, number born alive, number of implantations, testis weight, sperm count, nursing properties, and stress resistance during pregnancy. Some of these traits have been associated with genetic factors, but much variation remains unexplained. Substantial variability is seen in reproductive traits among mouse populations, and differences among inbred strains are common. Most features of mouse reproduction are alterable by environmental factors, including physical environment and presence of other animals; these environmental effects also vary among inbred strains.⁷⁹

Much of the current knowledge on reproduction in mice has developed alongside the breeding and maintenance of research populations including inbred lines. Inbreeding is associated with a strong decline in fertility and other fitness traits, and as a result the creation of fully inbred mouse lines is difficult. This ‘inbreeding depression’ probably results from accumulation of homozygosity for mutations with recessive effects.⁸¹ In mice, inbreeding is repeatedly demonstrated to reduce the number of offspring born alive by approximately 0.5 per 10% inbreeding.⁸² Nonetheless, the loss of fitness effect of inbreeding depression can be overcome, and inbred strains exist with a range of reproductive phenotypes including good breeders.

Hybrid sterility

An important contributor to fertility phenotypes in mice is hybrid sterility, in which matings of individuals from different mouse subspecies produce sterile hybrid offspring. Hybrid sterility is described in a range of taxa, and contributes to speciation by restricting gene flow between populations. It generally results from incompatible epistatic interactions between independently evolving genes in two or more subpopulations (the

Dobzhansky–Muller model⁸³). Hybrid sterility is often asymmetric; in mice, sterile F1 males result from one direction of cross, consistent with Haldane’s rule that sterility affects the heterozygous sex,⁸⁴ and indicating a role for sex chromosomes. The relative closeness of evolutionary distance between mouse subspecies (compared with other model organisms) makes the mouse a popular model for investigation of causal hybrid sterility and speciation genes.

In mice, a well-documented narrow hybrid zone exists between *M. m. musculus* and *M. m. domesticus* populations in eastern and western Europe, respectively (see Figure 1), within which males show significantly reduced (but variable) fertility.⁸⁵ Laboratory crosses between representatives of these two subspecies show that F1 hybrid sterility is asymmetric, with male sterility caused by the interaction of one locus, *Hst1* on chromosome 17, subsequently identified as *Prdm9*,⁸⁶ with *Hstx1* on the middle region of *M. m. musculus* chromosome X.⁸⁷ These loci are necessary but not sufficient for sterility, and thus other weak loci also seem to contribute to this process. Recessive loci also contribute to hybrid male sterility traits, indicating a need for investigation beyond the F1 generation.⁸⁸

An underlying mechanism of hybrid sterility in these laboratory crosses is failure of pairing of paternal and maternal chromosomes during meiosis (termed asynapsis), the effects of which are more severe in males than in females.⁸⁹ The asymmetric effect on asynapsis has been localized to a 4.7Mb region of chromosome X (named *Hstx2*, which overlaps with *Hstx1*).⁹⁰ This region contains several plausible candidate genes for the *Hstx1* and *Hstx2* loci.

Genomewide association mapping of sterility traits in mice from the European hybrid zone identifies multiple associated loci; however, these loci are expected to have epistatic effects, and few cases of incompatibilities exist in which all partner loci are known.⁹¹ Several such interactions are identified in an expression study of F2 hybrids, including several autosome–X-chromosome interactions and novel autosome–autosome interactions.⁹²

Studies of other mouse subspecies hybrids also identify a number of autosome and sex chromosome loci associated with hybrid sterility.^{93–95} While some overlap of identified chromosome regions is seen in these crosses, different genes seem to be involved. Similarly, differences seem to exist in the genes involved in determining sterility in crosses of different *M. m. musculus* and *M. m. domesticus* strains.⁸⁸ Female hybrid sterility,⁹⁶ assortative mating,⁹⁷ and reduced hybrid immune function might also contribute to reproductive isolation between mouse subspecies.

The Collaborative Cross

The Collaborative Cross (CC; csbio.unc.edu/CCstatus/) is a systems genetics resource of recombinant inbred lines descended from eight common mouse inbred strains. The CC was initiated in 2002, with the aim of providing a powerful tool for the dissection of the genetic architecture of complex traits.^{98,99} Simulation¹⁰⁰ and intermediate analyses^{101,102} report successful creation of the CC, with balanced allele frequencies, an even spread of recombination sites, good phenotypic distributions, and an anticipated genetic mapping resolution of ~1Mb. The CC is anticipated to offer a number of improvements over existing mouse genetic resources.

The eight CC founder strains were chosen on the basis of their popularity in mouse genetics, availability of sequence data, medical relevance, and their genetic diversity (Table 1). To maximise diversity, wild-derived strains from the three main mouse subspecies were included. Most common laboratory mouse strains and resources are overwhelmingly *M. m. domesticus* in origin and consequently they represent a relatively small part (~26%) of the possible genetic diversity.^{103,104} This limited diversity is not evenly distributed across the genome, resulting in blind spots for genetic mapping. By contrast, the eight CC founder strains capture 89% of the known genetic diversity in mice, randomly spread across the genome.¹⁰⁵

Table 1: Collaborative Cross founder strains

Name	Letter	Colour	Subspecies	SNP diff from ref ¹⁰⁶	Origin	Breeding performance	Comments
A/J	A	Yellow	<i>M.m.domesticus</i>	4,198,324	Classical	Challenging	Cancer & Immunology
C57BL/6J	B	Grey	<i>M.m.domesticus</i>	9,844	Classical	Good	Reference; overweight
129S1/SvImJ	C	Pink	<i>M.m.domesticus</i>	4,458,004	Classical	Good	Targeted mutations
NOD/ShiLtJ	D	Dark blue	<i>M.m.domesticus</i>	4,323,530	Classical	Good	T1D model
NZO/H1LtJ	E	Light blue	<i>M.m.domesticus</i>	4,492,372	Classical	N/A	Obesity & T2D model
CAST/EiJ	F	Green	<i>M.m.castaneum</i>	17,673,726	Wild-derived	Challenging	Wild-derived
PWK/PhJ	G	Red	<i>M.m.musculus</i>	17,202,436	Wild-derived	N/A	Wild-derived; hybrid sterile
WSB/EiJ	H	Purple	<i>M.m.domesticus</i>	6,045,573	Wild-derived	N/A	Wild-derived

Breeding scheme

The CC breeding scheme is shown in Figure 2. Each CC line results from mating of the eight founder strains in a funnel design (to generate heterozygous animals containing approximately equal contributions from each founder), followed by many generations of inbreeding. The order of the parent strains across funnels is randomised to ensure a diverse range of genotypes in the final (fully inbred) CC lines, and also balanced sex chromosome and mitochondrial contributions. At each generation of inbreeding, mating is initiated from randomly selected siblings from the previous generation. As breeding has progressed, the randomisation process has been adjusted to avoid funnel combinations that breed poorly or are infertile.¹⁰⁷

Breeding began in 2004 at three sites: Oak Ridge National Laboratory, TN (now moved to University of North Carolina [UNC]);¹⁰⁷ International Livestock Research Institute, Kenya (now moved to Tel Aviv University, Israel [TAU]);¹⁰⁸ and Western Australian Institute for Medical Research, Australia.¹⁰⁹ Updates were published in 2008 from each of these sites, outlining the status of the three breeding populations. In 2012, the first CC inbred lines were announced as ready for distribution.¹¹⁰

Mapping and other studies

To date a number of traits has been mapped in the partially inbred (pre-CC) mice (Table 2), including infection response,^{111–114} white head spotting,¹¹⁵ coat colour,¹¹⁶ body weight,^{115,117} gene expression,^{112,115,118,119} haematological parameters,¹²⁰ energy balance,¹¹⁷ albuminuria,¹²¹ immune cell phenotypes,¹²² allergic inflammation,^{118,123} and multiple other phenotypes.¹²⁴ These results illustrate the suitability and potential of the CC as a resource for trait mapping.

CC mice and founders have been used in a number of other studies (Table 3). Individual CC inbred lines have the potential to act as models for human disease, for example irritable

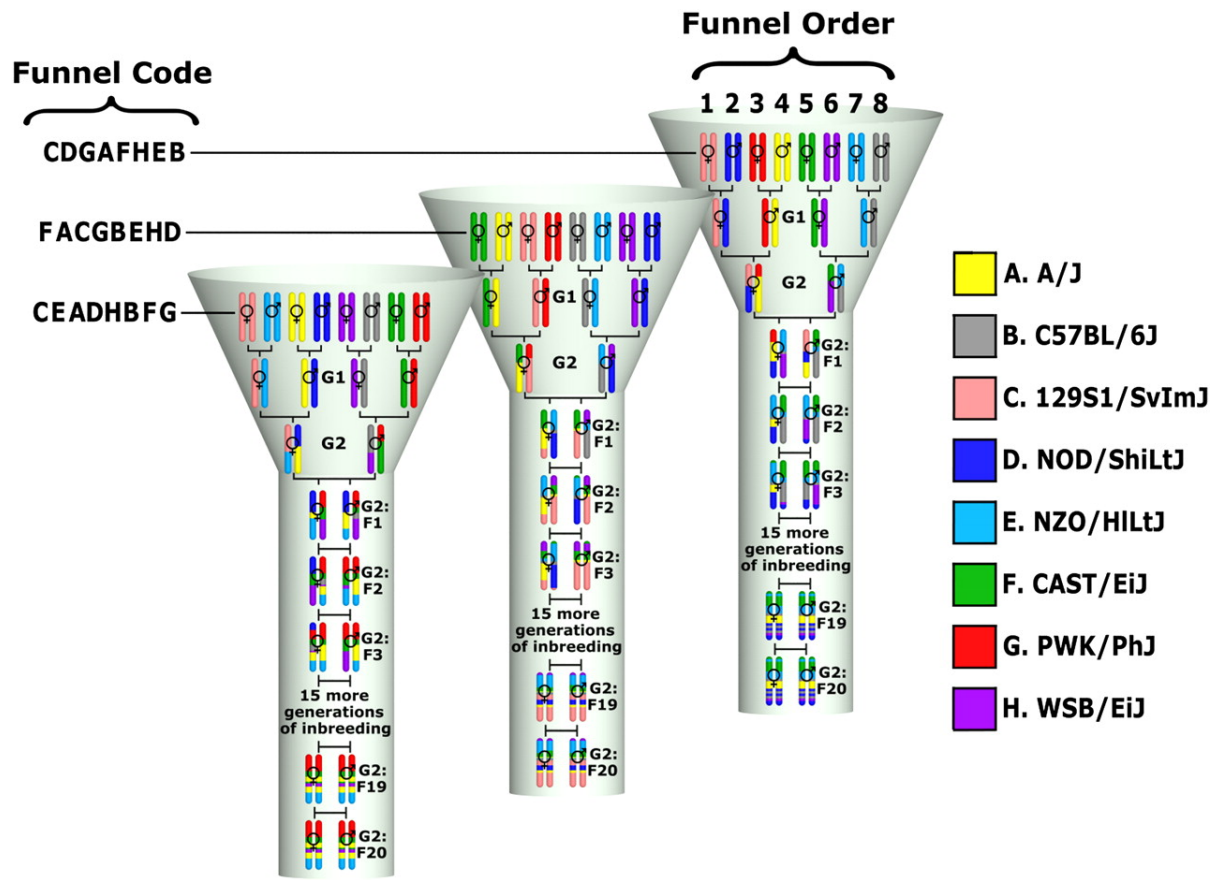


Figure 2: Breeding scheme of Collaborative Cross lines. The figure shows the breeding scheme for three independent CC lines. Each line has a funnel section followed by an inbreeding section. The eight founder strains are arranged in different positions (1–8) in each line, and this order determines the funnel code on the basis of a single letter code for each line. Founder order is randomized and not repeated across lines. The colors used for founder strains are seen throughout this article. Each mouse is represented by a pair of homologous autosomes and a symbol denoting its sex. © 2015 Genetics Society of America. Reproduced with permission

bowel disease.¹²⁵ Intercrossed CC lines were used to demonstrate genetic variation in susceptibility to Ebola hemorrhagic fever and West Nile virus, and proposed as a means to investigate pathogenesis and test therapeutic strategies.^{126,127} A number of studies have shown heritability of traits and genetic variation in phenotype, demonstrating suitability of these traits for genetic mapping. Other studies have documented phenotypes in the CC founder strains.

Despite the demonstrated suitability of the CC for mapping studies and other experimentation the CC has been slow to achieve its potential. One reason for this delay is the slower than expected breeding of the inbred lines, and thus a limited availability of the final inbred lines to date.¹⁰² At the time of writing (April 2015), 69 CC lines are listed as available to order from UNC (csbio.unc.edu/CCstatus/index.py?run=AvailableLines). Despite this availability, however, the mapping studies done to date have almost exclusively used the incipient (pre-CC) lines of the cross rather than the distributable and completed lines.

To examine the reach of the CC to date, I searched PubMed for publications related to the CC using the search term “Collaborative Cross” (done 23 April 2015). The search returns 129 articles, 73 of which relate to the CC. As expected the number of publications remained low until 2010, and has increased rapidly in the past 5 years. I categorize these articles according to their content (Figure 3a). I identify 16 mapping studies and a further 19 papers describing methods for use in mapping and analysis. Encouragingly, several studies are starting to use the CC founders in their experimental populations, which will facilitate integration of differing data sets based on the CC genotypes. Several additional mapping studies have been published in the remainder of 2015.^{128–130}

Table 2: **Trait mapping in the Collaborative Cross**

Phenotype	CC population	Outcomes	Reference
Infection response (<i>Klebsiella</i>)	Pre-CC (F8-F17)	3 QTLs	Vered <i>et al.</i> (2014) ¹¹⁴
Infection response (influenza A)	Pre-CC & founders	4 QTLs	Ferris <i>et al.</i> (2013) ¹¹¹
Infection response (influenza A)	Pre-CC	21 eQTLs	Bottomly <i>et al.</i> (2012) ¹¹²
Infection response (<i>Aspergillus</i>)	Pre-CC (F6-F12)	7 QTLs	Durrant <i>et al.</i> (2011) ¹¹³
White head spotting	Pre-CC (F5-F12)	1 QTL	Aylor <i>et al.</i> (2011) ¹¹⁵
Body weight	Pre-CC (F5-F12)	1 QTL	Aylor <i>et al.</i> (2011) ¹¹⁵
Gene expression	Pre-CC (F5-F12)	6,182 eQTLs	Aylor <i>et al.</i> (2011) ¹¹⁵
Gene expression	Pre-CC	5,744 local eQTLs	Weiser <i>et al.</i> (2014) ¹¹⁹
Haematological parameters	Pre-CC (F5-F14) & founders	5 QTLs	Kelada <i>et al.</i> (2012) ¹²⁰
Energy balance & body weight	Pre-CC (F5-F12) & founders	5 QTLs (+8 suggestive QTLs)	Mathes <i>et al.</i> (2011) ¹¹⁷
Multiple phenotypes	Pre-CC (~F7)	8 QTLs (+9 suggestive QTLs)	Philip <i>et al.</i> (2011) ¹²⁴
Albuminuria	Pre-CC (F5-F8)	1 QTL	Thaisz <i>et al.</i> (2012) ¹²¹
Immune cell phenotypes	Pre-CC (F5+) & founders	Several QTLs	Phillippi <i>et al.</i> (2013) ¹²²
Allergic inflammation	Pre-CC (F5-F8)	2 QTLs & >6,457 eQTLs	Kelada <i>et al.</i> (2014) ¹¹⁸
Neutrophilic inflammation	Pre-CC (F5-F14) & Founders	3 QTLs	Rutledge <i>et al.</i> (2014) ¹²³
Coat colour	CC	Confirm 6 QTLs	Ram <i>et al.</i> (2014) ¹¹⁶
Chr2 transmission ratio distortion	CC	Causal CNV	Didion <i>et al.</i> (2015) ¹³¹
Bone microarchitecture	Pre-CC (F11-F37)	6 QTLs	Levy <i>et al.</i> (2015) ¹²⁸
Motor performance	CC	45 QTLs	Mao <i>et al.</i> (2015) ¹²⁹
Body weight	CC	14 QTLs	Mao <i>et al.</i> (2015) ¹²⁹
Infection response (SARS-CoV)	Pre-CC	4 QTLs	Gralinski <i>et al.</i> (2015) ¹³⁰

Table 3: Nonmapping studies in Collaborative Cross mice

Phenotype	CC population	Outcomes	Reference
Spontaneous colitis	CC (CC011/Unc)	3 QTLs & model	Rogala <i>et al.</i> (2013) ¹²⁵
Ebola hemorrhagic fever	CC-RIX	Disease characterisation & model	Rasmussen <i>et al.</i> (2014) ¹²⁶
West Nile virus	CC-RIX	Disease characterisation & model	Graham <i>et al.</i> (2015) ¹²⁷
Drug response (haloperidol)	CC founders	Dissect genetic architecture	Crowley <i>et al.</i> (2014) ¹³²
Infection response (influenza A & SARS-CoV)	CC founders	lncRNA expression	Josset <i>et al.</i> (2014) ¹³³
Infection response (influenza A & SARS-CoV)	CC founders	Complex transcriptional response	Xiong <i>et al.</i> (2014) ¹³⁴
ER stress response	CC founders	Genetic variation in response	Chow <i>et al.</i> (2015) ¹³⁵
Splicing landscape	CC founders	Strain-specific splicing	Zheng <i>et al.</i> (2015) ¹³⁶
Glomerulonephritis severity	CC founders	Identify complex polygenic trait	Xiao <i>et al.</i> (2013) ¹³⁷
Prostate carcinoma	CC founders	Genetic variation in progression	Patel <i>et al.</i> (2013) ¹³⁸
Selection for physical activity	Pre-CC (F1)	Response to selection	Zombeck <i>et al.</i> (2011) ¹³⁹
Drug response (bleomycin)	Pre-CC (F7–F10)	Genetic variation in response	Gelinas <i>et al.</i> (2011) ¹⁴⁰
Gut microbiota	Pre-CC (F10–F15)	Genetic variation in microbiota	Kovacs <i>et al.</i> (2011) ¹⁴¹
Gut microbiota	CC founders	Genetic variation in microbiota	Campbell <i>et al.</i> (2012) ¹⁴²
Microbiota response to diet	CC founders	Diet–strain interactions	O’Connor <i>et al.</i> (2014) ¹⁴³
Melanoma	CC	Genetic variation in tumorigenesis	Ferguson <i>et al.</i> (2013) ¹⁴⁴
Multiple traits	Pre-CC & CC	Significant heritability	Iraqi <i>et al.</i> (2014) ¹⁴⁵
Periodontitis susceptibility	Pre-CC (F10–F18)	Significant heritability	Shusterman <i>et al.</i> (2013) ¹⁴⁶
Infection response (<i>P. aeruginosa</i>)	CC (F20–F39)	Genetic variation in phenotype	Loré <i>et al.</i> (2015) ¹⁴⁷
Induced lung inflammation	CC founders	Genetic variation in susceptibility	Scoville <i>et al.</i> (2015) ¹⁴⁸

As CC mapping studies have to date been done using the incipient CC lines, they have been conducted almost exclusively by the CC breeding centres. Over time the quantity of mapping and other studies will offer a measure of the uptake of the CC resource beyond the breeding centres. In this regard, CC papers represent 8 of 95 (8.42%) papers identified on a PubMed search for “mouse” + “qtl” from 2014. Although very roughly calculated, this proportion provides a baseline for monitoring impact of the CC over time.

I also use citation data to investigate the wider impact of the CC to date. Overall, the 73 CC papers have generated 1,320 citations in 887 articles outside the 73 (data from Thomson Reuters Web of Science; done 23 April 2015). Initially, all citations were to overview papers describing the structure and purpose of the CC (Figure 3b). Encouragingly, since the first mapping studies in 2011, both the total number of citations has increased and a sizable proportion of citations is attributable to mapping and other experimental studies. The 16 mapping studies have generated downstream citations from authors in 25 countries (not shown). It would be interesting to further map these results to see the geographical and institutional downstream reach of the CC publications.

At present, the users of the CC resource are clustered around the project breeding centres, without much overlap between these centres. Step-by-step guidelines to using the CC resource for mapping and other studies would be very welcome, as would consensus on the methods for following up on genetic associations discovered through mapping. In this regard, a centralised knowledge base, perhaps in the form of a collaborative wiki would be useful, as would a clear platform for data integration (for example, GeneNetwork;¹⁴⁹ www.genenetwork.org). Two 2015 papers highlighting CC bioinformatics pathways and shared data through the Mouse Phenome Database are very encouraging.^{150,151}

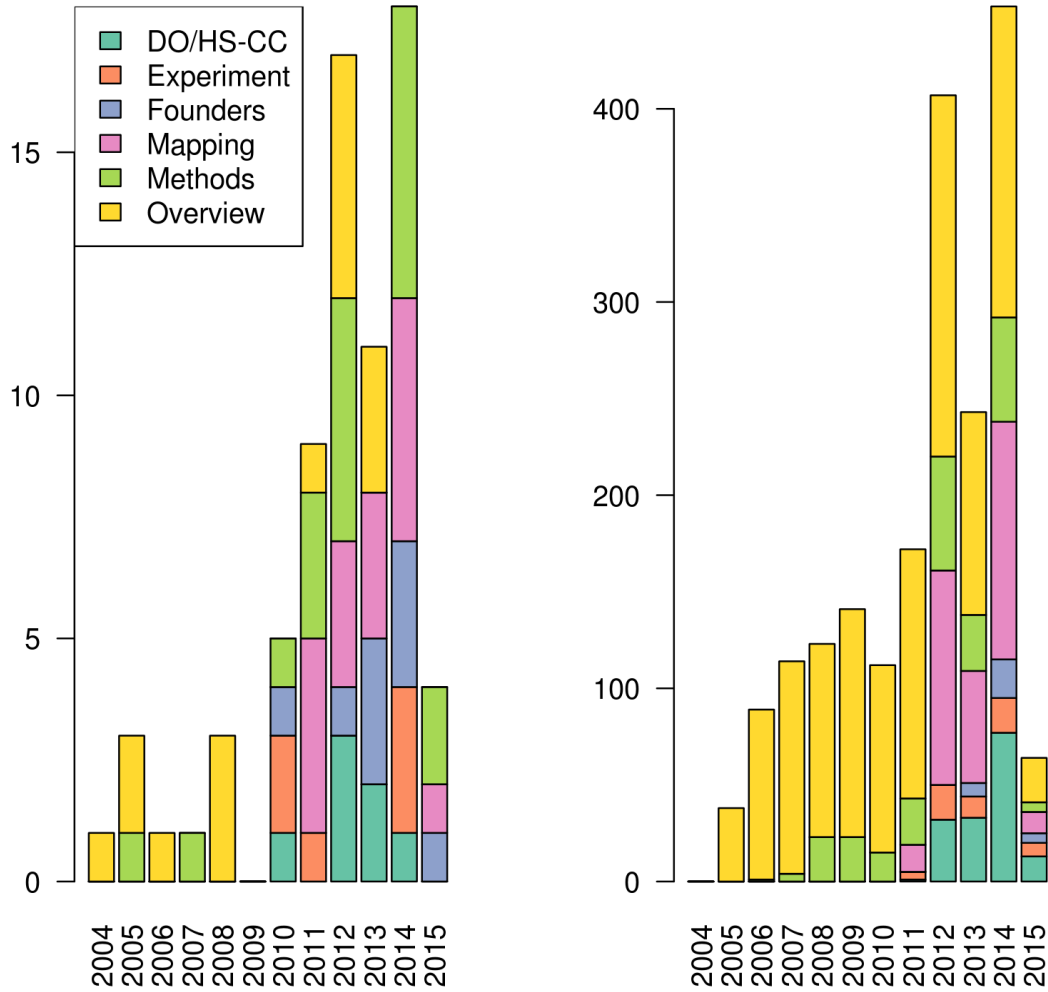


Figure 3: Published Collaborative Cross papers by category. The left panel (a) shows published articles categorized according to their purpose, into overviews, methods articles, mapping studies, investigation of founder strains, nonmapping experiments, and studies in the related DO or HS-CC mice. Mapping studies have been published steadily since 2011, and novel methods continue to be developed and discussed. Data obtained by searching PubMed for the term ‘Collaborative Cross’, and excluding nonrelevant articles. The right panel (b) shows Citations to the first 73 published articles on the Collaborative Cross by year. The majority of citations have been to overviews of the project. Since the first mapping studies were published in 2011 they have generated a number of downstream citations. Data from Thomson Reuters Web of Science, excluding citations within the CC corpus. Abbreviations: CC, Collaborative Cross; DO, Diversity Outbred; HS-CC, Heterogeneous Stock Collaborative Cross.

Breeding performance

The CC breeding program is somewhat behind schedule, owing to slower-than-expected inbreeding,¹⁰² and in large part to the difficulty of creating fully inbred lines from the eight founders. The initial CC plans anticipated a strain attrition rate of 30–40%;⁹⁸ in fact, extinction rates have been considerably greater, reported in excess of 70% in the UNC arm.¹⁰¹ As a result, the anticipated final number of inbred lines has dropped from 1,000, first to approximately 500 and now in all likelihood to fewer.

Breeding performance in the CC seems to be a line-specific heritable trait, as shown by parent–offspring regression coefficients for reproductive traits.¹²⁴ This observation supports previous reports that traits such as litter size and time to first litter are heritable (h^2 in the range 0.1–0.2) in mice^{40,152} and in agricultural and domesticated animals.^{153,154} Furthermore, selection studies on litter size have been successful in laboratory mouse populations, with significant changes seen within 30 generations of breeding.^{82,155} Genetic components associated with litter sizes have been identified, including the genes *Bmp15* and *Gdf9*, which act through modulation of ovarian function and are also reported to affect reproduction in sheep and other mammals.^{156–160}

Data from the UNC arm of the CC, obtained from the Mouse Phenome Database (phenome.jax.org)¹⁶¹ show that the eight founder strains of the CC exhibit notable variation in litter size (Figure 4a).¹⁶² Whereas NOD/ShiLtJ mice have a mean litter size greater than 8 pups, the wild-derived strains PWK/PhJ and WSB/EiJ typically have much smaller litters. With the exception of the C57BL/6J mice, sample sizes are low ($n = 3$ to $n = 7$). For pre-CC females, mean litter sizes were typically within the range of the founders (Figure 4b), indicating a trait amenable to mapping in the CC. A study of gestation length in inbred mouse strains reports much the same litter sizes for the founder strains.¹⁶³

Taken together, these observations suggest that while the eight founder strains of the

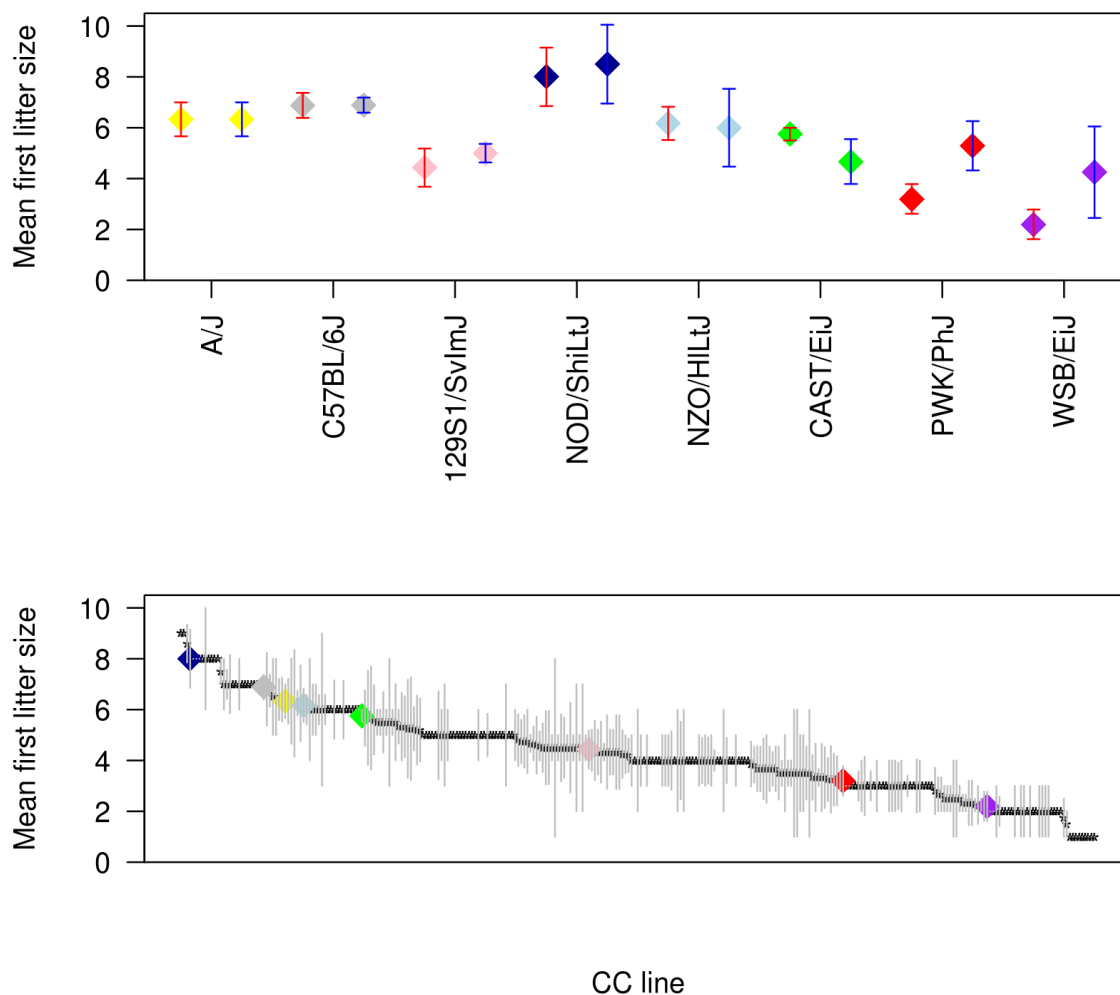


Figure 4: First litter size for CC mice. The upper panel (a) shows first litter sizes produced by females and males (females shown first in each pair). The calculated male and female means share some data, but only A/J uses exactly the same samples for both sexes. Any observed differences between sexes in first litter sizes are likely to result from the small sample sizes. Error bars depict SEM. The lower panel (b) shows first litter size of incipient Collaborative Cross lines. The incipient inbred lines are arranged in descending order along the x axis. The range of mean first litter sizes for females in the incipient lines of the Collaborative Cross exceeds that of the founders, and the variation between lines is greater than that within lines indicating litter size is amenable to quantitative trait locus mapping. Light gray bars depict SEM. Data from the Mouse Phenome Database: 43541.¹⁶²

CC breed reasonably well and are relatively easy to maintain, when crossed and inbred they show significant variation and frequent problems in fertility. Decreasing fertility attributable to inbreeding depression is expected with inbred lines, but the systematic variation between lines and the existence of founder combinations that breed poorly suggest that specific genetic factors also affect breeding performance.

The existence of records of the CC breeding process offers the opportunity to investigate this phenomenon, and to potentially identify genetic mechanisms by which breeding performance is enhanced or impaired. The observed difficulties in breeding performance might result from several overlapping factors including inbreeding depression caused by accumulation of deleterious mutations, and specific allelic incompatibilities such as those causing hybrid sterility between *M. m. musculus* and *M. m. domesticus*. In addition, some of the observed variation between inbred lines is likely to result from additive variation that is amenable to quantitative trait mapping approaches. In the next chapters, I analyse breeding records from one of the CC populations to investigate genetic factors associated with reproductive performance in these mice.

Chapter 3 | Collaborative Cross breeding records

The idea that the data *must certainly*, if properly massaged, be able to answer the burning question is most seductive and dangerous

—John W Tukey (1993)¹⁶⁴

The CC at Tel Aviv University

In order to investigate the genetics of reproductive traits in the CC, it is necessary to have phenotypic data on these traits. Having been advised that the breeding records for the CC populations would contain suitable phenotype data, I obtained these records from the CC population maintained at Tel Aviv University, Israel (TAU). These records exist in the form of three extensive interrelated .xls workbooks, which respectively recorded those CC lines initiated in Kenya (two workbooks) and those subsequently started in Tel Aviv. It is immediately obvious that these breeding records are not in fact true pedigrees in which individuals are assigned unique identifiers and family relationships can be tracked. Instead, the records are organized to record littering events and their associated information.

The workbooks each comprise multiple data sheets, and have been maintained by hand. As a result they contain numerous duplicate records and empty cells, which has required careful curation and cleaning before analysis to minimise the risk of systematic error. In particular, the spreadsheets are arranged such that information from a particular breeding funnel is spread over several rows, often with labelling only on the first line; consequently they contain many empty cells that might or might not be informative. I invested significant time into translating these spreadsheets into a format that was suitable for computational analysis. As the structure of the breeding records affects the conclusions drawn from the data I illustrate the three workbooks in Figures 5–7.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
	Line #	Line #	Line genotype	Keyway cross	G2 litter size/f	Comments	G3-M-weans	G3-F-weans	N-G3 line	G4 litter size/fem	G4-M-weans	G4-F-weans	N-G4												
2	IL1A	IL1	BAFGEHDC	cage 33	8.7		1.3	2	IN-21	10	5	5	IN-21												
3	IL2A	IL2	GBECDAHAF	cage 160	9		2.5	3	IN-88	8															
4	IL5A	IL5	GBHAFEC	cage 34	6.5		1.5	2	IN-22	9	6	0													
5	IL6A	IL6	BHAGECFD	cage 35	4		1.5	2	IN-23																
6	IL7A	IL7	CEHFBDA	cage 14	6.5		2	3.5	IN-38	4.5	1.5	2	IN-38												
7	IL9A	IL9	FAHGBEDC	cage 15	9		3	3	IN-124	6															
8	IL11A	IL11	HBACDEFG	cage 212	6		2	4	IN-136	7															
9	IL12A	IL12	FBACDAEHG	cage 103	10		5	4	IN-125																
10	IL13A	IL13	ABDEGFHC	cage 104	10		2	4	IN-61	9															
11	IL15A	IL15	GBFACEDH	cage 42	7		2.5	2.5	IN-32	11	6	2	IN-32												
12	IL17A	IL17	FEAGDBHC	cage 16	7		2.5	3	IN-11	7	4	3	IN-11												
13	IL18A	IL18	GAEBCHDF	cage 142	7.5		1	5	IN-90	6															
14	IL19A	IL19	DCBHFHEA	cage 143	5		2.5	3	IN-72	7															
15	IL20A	IL20	GBFHACDE	cage 144	6		3	1.5	IN-64																
16	IL21A	IL21	DABHFEGC	cage 237	11		8	3	IN-177																
17	IL22A	IL22	CBGHFEA	cage 4	7		3.8	1	IN-2	9	3	4	IN-2												
18	IL23A	IL23	BEHGFADC	cage 18	4		2.5	1.5	IN-26	6.5	2	4	IN-26												
19	IL26A	IL26	DCGFHEAB	cage 147	7.5		4	1	IN-147																
20	IL27A	IL27	QHADCFE	cage 2	10		4.5	4.5	IN-3	7.5	3.5	2.5	IN-3												
21	IL28A	IL28	BHGFADG	cage 224																					
22	IL29A	IL29	DGFCABH	cage 87	9		2	3	IN-54																
23	IL30A	IL30	GBFHGAC	cage 88	4		3	1	IN-40																
24	IL31A	IL31	DBAQHECF	cage 146	7		2	3	IN-73	6															
25	IL33A	IL33	DHFGCAB	cage 90	5		2	2	IN-65	9															
26	IL34A	IL34	GHFCDBAE	cage 145	7		5	1	IN-65																
27	IL35A	IL35	ABDEGFHC	cage 146																					
28	IL37A	IL37	HBACDFD	cage 148	6		5	3	IN-171																
29	IL38A	IL38	BGFEDCAH	cage 161	5		2.5	2.5	IN-69	3															
30	IL39A	IL39	HEABDCGF	cage 43	12		4	6	IN-33	8.5	4	4	IN-33												
31	IL40A	IL40	DBEHGFAC	cage 92	11		7	3	IN-47	5															
32	IL41A	IL41	CGAFBEDH	cage 162	9		3	3	IN-175																
33	IL45A	IL45	HCFDGAEB	cage 213	8.5		2.5	3	IN-106																
34	IL47A	IL47	CBHFGEDA	cage 13	10.5		3.5	2	IN-163																
35	IL48A	IL48	GADFCENB	cage 94	8		6	1	IN-57	9															
36	IL49A	IL49	HACDFGFE	cage 163	5		2	4	IN-180	6															
37	IL50A	IL50	CBGEADFH	cage 211	6		1	1	IN-105																
38	IL51A	IL51	GCHABFDE	cage 44	6		3	3	IN-130																
39	IL53A	IL53	FHGDABCG	cage 3	8.5		3.5	3	IN-1	5	2	2	IN-1												
40	IL54A	IL54	AEDBGCHF	cage 242	10.5		2	4	IN-161																
41	IL55A	IL55	EDBCAGFH	cage 20	13.5		3.5	2.5	IN-17		3	3	IN-17												
42	IL57A	IL57	DACFCBPH	cage 149	14		7	5	IN-74	11															
43	IL58A	IL58	EAQFBHCD	cage 45	8		2	4	IN-99	12															
44	IL59A	IL59	AHCEGDF	cage 105	11		3	6	IN-132																
45	IL60A	IL60	EDFGHBAC	cage 151	10		2.5	2.5	IN-66	7															
46	IL62A	IL62	CGAFBDHE	cage 164	8		4	4	IN-151																
47	IL63A	IL63	ECDFBHAC	cage 36	10		2	4	IN-36	6															
48	IL67A	IL67	CBDHGEFA	cage 21	5.5		3	2.5	IN-16	8	3	3	IN-16												
49	IL68A	IL68	DHAFBGCE	cage 165	14		4	5	IN-75																
50	IL69A	IL69	BFHGDCA	cage 5	9		4	4.5	IN-4	6	2	3	IN-4												
51	IL71A	IL71	AEFGDBCH	cage 153	7		2	5	IN-133																
52	IL72A	IL72	CADBFHGE	cage 46	10		7	3	IN-115	4															
53	IL73A	IL73	DOCEFBAB	cage 214	13		4	8	IN-141																
54	IL76A	IL76	HCFDGAEB	cage 95	8		2	2	IN-60	8															
55	IL77A	IL77	CEHBDGAF	cage 96	7		4	3	IN-49	7															
56	IL78A	IL78	EHDFBCGA	cage 97	6		5	1	IN-41	5															
57	IL80A	IL80	MBEAFHED	cage 98		not preg (late mating-13 weeks)																			
58	IL81A	IL81	AGEBHFC	cage 180	8		3.5	2.5	IN-77	8															
59	IL82A	IL82	FBGDAEHD	cage 105	9		5	3	IN-62	8															
60	IL83A	IL83	FBGDAEHC	cage 152	8		2.5	3.5	IN-110																
61	IL84A	IL84	GBFHACDE	cage 47	13		7	5	IN-53																
62	IL85A	IL85	BHCEADGF	cage 9	10		3.5	3.5	IN-14	12	5	6	IN-14												
63	IL88A	IL88	EHCAEDGF	cage 6	10.5		2	2	IN-5	10	5	5	IN-5												
64	IL89A	IL89	BFDACEHG	cage 166	10		3.5	2.5	IN-152																
65	IL90A	IL90	ABFHGECD	cage 22	6.3		1	3.7	IN-39	10															

Figure 5: Breeding records for Kenyan lines before move to TAU. CC_in_Kenya_8-07-06.xls contains records from inbred lines initiated in Kenya. Successive generations of inbreeding are recorded across three sheets. Some of the numbers of offspring given seem to be summaries—most likely means—which limits their use for subsequent mapping. The workbook is dated July 2006, which means a gap of around 1 year exists between the period covered by this workbook and the resumption of record keeping in the subsequent workbooks.

I	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG				
1	IL	Parent Generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL	generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL	generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL	generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL	generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL	generation	L Size D.B	L Size W Male	L Size W Female	Date of Mating	IL
2																																					
3	IL-7	8	7/3/07	5	4	1	26/3/07	IL-7	9	14/9/07	7	5	2	13/9/07	IL-7	10	26/7/07	1	1	0	nm	IL-7	11	20/12/07	2	1	1	9/1/08							IL-7		
4	IL-7	8	29/3/07	8	3	5	19/4/07	IL-7	9	5/9/07	7	5	0	nm	IL-7	10	15/9/07	1	0	0	nm	IL-7	11	23/12/07	3	1	1	16/1/08							IL-7		
5	IL-7	8	22/4/07	8	5	3		IL-7	9	17/6/07	1	0	0	nm	IL-7	10	16/9/07	1	0	1	nm	IL-7	11	9/1/08	6	3	3	nm							IL-7		
6	IL-7	8	16/5/07	8	4	4	7/6/07	IL-7	9	18/6/07	3	2	1	17/7/07	IL-7	10	8/10/07	2	1	0	nm	IL-7	11	12/1/08	3	2	1	19/2/08							IL-7		
7	IL-7	8	9/6/07	9	7	1	nm	IL-7	9	26/6/07	6	1	3		IL-7	10	18/10/07	8	4	4	17/11/07	IL-7	11	2/2/08	4	2	2	nm							IL-7		
8	IL-7	8	3/7/07	8	4	4	nm	IL-7	9	17/7/07	4	0	0	nm	IL-7	10	18/11/07	2	0	0	nm	IL-7	11	23/2/08	5	3	1	nm							IL-7		
9	IL-7	8	11/8/07	6	0	0	terminated	IL-7	9	7/8/07	6	0	0	nm	IL-7	10	31/12/07	6	1	3	nm	IL-7	11	4/5/08	3	0	0	terminated							IL-7		
10	IL-7	8						IL-7	9	18/9/07	5	4	1		IL-7	10	9/10/08	5	3	0	nm	IL-7	11											IL-7			
11	IL-7	8						IL-7	9	14/9/07	5	4	1	nm	IL-7	10	terminated					IL-7	11	terminated										IL-7			
12	IL-7	8						IL-7	9	29/9/07	3	0	2	nm	IL-7	10					IL-7	11											IL-7				
13	IL-7	8						IL-7	9	18/9/07	4	1	2	nm	IL-7	10					IL-7	11											IL-7				
14	IL-7	8						IL-7	9	5/9/07	3	0	0	nm	IL-7	10					IL-7	11											IL-7				
15	IL-7	8						IL-7	9	19/9/07	6	5	0	nm	IL-7	10					IL-7	11											IL-7				
16	IL-7	8						IL-7	9	11/10/07	10	4	2	nm	IL-7	10					IL-7	11											IL-7				
17	IL-7	8						IL-7	9	11/10/07	5	0	0	nm	IL-7	10					IL-7	11											IL-7				
18	IL-7	8						IL-7	9	23/10/07	7	5	2	nm	IL-7	10					IL-7	11											IL-7				
19	IL-7	8						IL-7	9	24/10/07	9	2	7	nm	IL-7	10					IL-7	11											IL-7				
20	IL-7	8						IL-7	9	20/11/07	3	0	0	nm	IL-7	10					IL-7	11											IL-7				
21	IL-7	8						IL-7	9	terminated					IL-7	10					IL-7	11											IL-7				
22	IL-11	6	25/3/07	5	2	3	19/4/07	IL-11	7	3/6/07	5	1	5	27/6/07	IL-11	8	9/8/07	3	0	0	nm	IL-11	9	13/11/07	3	0	2	nm							IL-11		
23	IL-11	6	16/4/07	6	0	0	nm	IL-11	7	26/6/07	8	4	3	17/7/07	IL-11	8	3/9/07	6	3	3	25/9/07	IL-11	9	24/11/07	4	2	2	nm							IL-11		
24	IL-11	6	19/4/07	3	1	1	17/7/07	IL-11	7	22/7/07	7	5	2	nm	IL-11	8	27/9/07	9	4	6	nm	IL-11	9	21/9/08	6	3	1	11/6/08							IL-11		
25	IL-11	6						IL-11	7	12/8/07	8	4	4	nm	IL-11	8	24/10/07	5	2	2	nm	IL-11	9	10/6/08	2	0	2	nm							IL-11		
26	IL-11	6						IL-11	7	3/9/07	5	2	3	25/9/07	IL-11	8	10/11/07	3	2	0	nm	IL-11	9	16/6/08	4	0	3	nm							IL-11		
27	IL-11	6						IL-11	7	5/9/07	8	0	0	terminated	IL-11	8	7/12/07	5	3	2	nm	IL-11	9	12/7/08	4	1	2	6/8/08							IL-11		
28	IL-11	6						IL-11	7	24/9/07	6	3	3	nm	IL-11	8	4/2/08	3	0	0	nm	IL-11	9	4/8/08	4	3	0	nm							IL-11		
29	IL-11	6						IL-11	7	20/10/07	4	1	2	nm	IL-11	8	11/2/08	3	0	0	nm	IL-11	9	6/8/08	4	0	3	nm							IL-11		
30	IL-11	6						IL-11	7	14/11/07	4	1	2	nm	IL-11	8	21/2/08	6	1	4	18/3/08	IL-11	9	26/8/08	1	1	0	nm							IL-11		
31	IL-11	6						IL-11	7	7/12/07	8	2	5	21/1/08	IL-11	8	29/2/08	3	0	1	nm	IL-11	9	16/3/08	7	2	5	3/7/08							IL-11		
32	IL-11	6						IL-11	7	terminated					IL-11	8	1/3/08	6	2	3	9/4/08	IL-11	9	28/10/08	4	1	2	18/11/08							IL-11		
33	IL-11	6						IL-11	7						IL-11	8	16/3/08	4	1	0	nm	IL-11	9	terminated										IL-11			
34	IL-11	6						IL-11	7						IL-11	8	terminated					IL-11	9											IL-11			
35	IL-11	6						IL-11	7						IL-11	8	6/4/08	5	3	2	nm	IL-11	9											IL-11			
36	IL-15	5	22/3/07	5	2	3	19/4/07	IL-15	6	30/7/07	4	3	0	nm	IL-15	7	26/12/07	5	2	3	16/1/08	IL-15	8	5/8/08	2	1	0	nm							IL-15		
37	IL-15	5	17/4/07	4	0	2	nm	IL-15	6	28/8/07	4	2	1	18/9/07	IL-15	7	8/2/08	7	4	4	5/3/08	IL-15	8	11/9/08	5	1	0	nm							IL-15		
38	IL-15	5	10/5/07	6	3	2	6/6/07	IL-15	6	19/9/07	9	8	0	nm	IL-15	7	8/3/08	1	1	0	24/8/08	IL-15	8	14/9/08	1	1	0	nm							IL-15		
39	IL-15	5	5/6/07	5	2	3	27/9/07	IL-15	6	7/10/07	2	0	0	nm	IL-15	7	14/10/08	3	0	3	nm	IL-15	8	25/1/09	1	0	0	nm							IL-15		
40	IL-15	5	17/7/07	8	6	2	18/9/07	IL-15	6	12/10/07	8	4	2	7/11/07	IL-15	7	10/4/08	5	3	1	nm	IL-15	8	28/3/09	2	0	0	nm							IL-15		
41	IL-15	5	29/7/07	7	4	2	nm	IL-15	6	18/10/07	2	1	1	7/11/07	IL-15	7	22/4/08	4	3	1	nm	IL-15	8	15/4/09	2	0	0	nm							IL-15		
42	IL-15	5	21/8/07	6	3	2	nm	IL-15	6	6/11/07	8	2	5	28/11/07	IL-15	7	15/5/08	8	4	1	nm	IL-15	8	17/7/09	3	0	0	nm							IL-15		
43	IL-15	5						IL-15	6	4/12/07	7	0	1	nm	IL-15	7	17/9/08	4	1	1	11/6/08	IL-15	8										IL-15				
44	IL-15	5						IL-15	6	11/2/07	3	1	2	nm	IL-15	7	6/8/08	6	1	1	nm	IL-15	8										IL-15				
45	IL-15	5						IL-15	6	6/1/08	4	2	2	30/1/08	IL-15	7	11/6/08	3	2	1	nm	IL-15	8										IL-15				
46	IL-15	5						IL-15	6	9/3/08	5	2	2	nm	IL-15	7	13/6/08	8	2	5	8/7/08	IL-15	8										IL-15				
47	IL-15	5						IL-15	6	terminated					IL-15	7	3/7/08	7	3	4	nm	IL-15	8										IL-15				
48	IL-15	5						IL-15	6						IL-15	7	12/7/08	5	3	1	nm	IL-15	8										IL-15				
49	IL-15	5						IL-15	6						IL-15	7	5/8/08	2	2	0	nm	IL-15	8										IL-15				
50	IL-15	5						IL-15	6						IL-15	7	21/9/08	4	1	1	15/9/08	IL-15	8										IL-15				
51	IL-15	5						IL-15	6						IL-15	7	21/9/08	4	1	1	15/10/08	IL-15	8										IL-15				
52	IL-15	5						IL-15	6						IL-15	7	21/9/08	2	0	0	nm	IL-15	8										IL-15				
53	IL-15	5						IL-15	6						IL-15	7	15/10/08	7	1	2	9/11/08	IL-15	8										IL-15				
54	IL-15	5						IL-15	6						IL-15	7	11/11/08	5	3	1	nm	IL-15	8										IL-15				

Figure 6: Breeding records for Kenyan lines after move to TAU. CC_from_RIHS_24-6-2012.xls contains two sheets of records from these Kenyan lines after the move to TAU. The first sheet contains some brief notes on death and termination of inbred lines (not shown); the second sheet contains records of parent generation, date of birth, litter size, number of weaning males and females, and date of subsequent mating. Data are arranged consistently by column, but the spreadsheet rows are only informative within a given colour block (i.e. for a given row, the mating listed in the blue columns [I:P] does not necessarily occur between the offspring of the mating listed in the white columns [A:H]). Thus time between mating and birth cannot be easily tracked; neither can

J	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		A-(M-AKF-B)																						
2	# Funnel	Reciprocal	Substage	Generation	D.Mating	D.Birth	litter	W-M	W-F	Generation	Substage	D.Mating	D.Birth	litter	W-M	W-F								
3	1011	ABECGFDFH	A	G3	01/04/2009	02/05/2009	7	2	4	G4	A	27/05/2009												
4	1011	ABECGFDFH	A	G3		26/05/2009	7	4	3	G4	B	28/06/2009												
5	1011	ABECGFDFH	A	G3		21/06/2009	9	4	4	G4	C	16/07/2009												
6	1011	ABECGFDFH	A	G3		13/07/2009	8	4	4	XXX	XXX													
7	1011	ABECGFDFH	B	G3	01/04/2009	13/05/2009	5	3	2	XXX														
8	1011	ABECGFDFH	C	G3																				
9	1032	ACBEGHFD	A	G3	17/06/2009	21/07/2009	15	4	4	XXX	XXX													
10	1032	ACBEGHFD	A	G3		18/08/2009	8	4	4	XXX	XXX													
11	1032	ACBEGHFD	A	G3		09/09/2009	8	3	5	XXX	XXX													
12	1032	ACBEGHFD	B	G3	17/06/2009	16/07/2009	9	2	5	G4	A	29/07/2009												
13	1032	ACBEGHFD	B	G3						G4	B	29/07/2009												
14	1032	ACBEGHFD	B	G3		29/07/2009	10																	
15	1032	ACBEGHFD	B	G3		18/08/2009	9	4	3	XXX	XXX													
16	1032	ACBEGHFD	B	G3		29/09/2009	10	4	6	XXX	X													
17	1052	ACGDEHBF	A	G3	27/07/2009	10/09/2009	16	5	11	G4	A	01/10/2009												
18	1052	ACGDEHBF	A	G3	27/07/2009	10/09/2009				G4	B	01/10/2009												
19	1052	ACGDEHBF	A	G3	27/07/2009	10/09/2009				G4	C	01/10/2009												
20	1052	ACGDEHBF	A	G3		15/11/2009	17	8	9	XXX	XXX													
21	1052	ACGDEHBF	A	G3		07/12/2009	14	6	7	XXX	XXX													
22	1052	ACGDEHBF	B	G3	27/07/2009					XXX	XXX													
23	1052	ACGDEHBF	B	G3		04/10/2009	12	0	10	XXX	XXX													
24	1061	ADBCFGEH	A	G3	28/06/2009	01/08/2009	6	4	1	XXX	XXX													
25	1061	ADBCFGEH	A	G3	28/06/2009	01/08/2009																		
26	1061	ADBCFGEH	A	G3		20/08/2009	7	5	2	XXX	XXX													
27	1061	ADBCFGEH	A	G3		13/09/2009	10	7	2	XXX	XXX													
28	1061	ADBCFGEH	A	G3		09/10/2009	9	5	4	XXX	XXX													
29	1061	ADBCFGEH	B	G3	28/06/2009	02/08/2009	7	3	3	G4	A	08/09/2009												
30	1061	ADBCFGEH	B	G3	28/06/2009	02/08/2009				G4	B	08/09/2009												
31	1061	ADBCFGEH	B	G3		23/08/2009	9	6	3	XXX	XXX													
32	1061	ADBCFGEH	B	G3		24/09/2009	9	4	4	XXX	XXX													
33	1072	ADECFBGH	A	G3	15/07/2009																			
34	1072	ADECFBGH	B	G3	15/07/2009	31/08/2009	11	4	6	G4	A	24/09/2009												
35	1072	ADECFBGH	B	G3						G4	B	24/09/2009												
36	1072	ADECFBGH	B	G3						G4	C	24/09/2009												
37	1072	ADECFBGH	B	G3		22/09/2009	3	1	2	XXX	XXX													
38	1072	ADECFBGH	B	G3		15/11/2009	10																	
39	1076	ADFBGHCE	A	G3	01/04/2009	05/05/2009	9	3	3	G4	A	28/06/2009												
40	1076	ADFBGHCE	A	G3		09/06/2009	11	4	6	G4	B	28/06/2009												
41	1076	ADFBGHCE	A	G3		03/07/2009	10	7	3	XXX	XXX													
42	1076	ADFBGHCE	A	G3		20/08/2009	11	5	6	XXX	XXX													
43	1076	ADFBGHCE	B	G3	01/04/2009	19/05/2009	8	6	2	XXX	XXX													
44	1076	ADFBGHCE	B	G3		10/06/2009	7	4	3	XXX	XXX													
45	1076	ADFBGHCE	B	G3		30/06/2009	10	7	2	XXX	XXX													
46	1076	ADFBGHCE	B	G3		04/08/2009	10	3	7	XXX	XXX													
47	1076	ADFBGHCE	B	G3		25/08/2009	6	3	3	XXX	XXX													
48	1086	ADHBFGCE	A	G3	13/04/2009	11/05/2009	8	1	4	G4	A	28/06/2009												
49	1086	ADHBFGCE	A	G3		07/06/2009	10	2	5	G4	B	28/06/2009												
50	1086	ADHBFGCE	A	G3		02/07/2009	8	4	2	XXX	XXX													
51	1086	ADHBFGCE	A	G3		16/08/2009	10	4	6	XXX	XXX													
52	1086	ADHBFGCE	B	G3																				
53	1102	AEDCFBGH	A	G3	10/07/2009																			
54	1102	AEDCFBGH	B	G3	10/07/2009																			
55	1107	AEBGBCDH	G3																					
56	1116	AEBGBCDF	A	G3	13/04/2009	19/05/2009	8	5	3	XXX	XXX													

Figure 7: Breeding records for CC lines initiated at TAU. CC_from_ORNL_24-6-2012.xls contains 21 sheets of records for the CC lines initiated at TAU. Each sheet represents a successive stage of the inbreeding process, with some duplication and redundancy. Mice arrived from ORNL as G1 founder strain crosses (e.g. A x B). Several sheets then document the formation and progress of the G2 generation. Sheet '!8-WAY-YELLOW-3' documents the outcomes of the G2 matings, with the litter sizes reported representing the G2:F1 generation—the first generation with genetic contributions from all founders. Finally, the remainder of the sheets document successive generations of inbreeding. These records are stored in a reasonably sensible manner, with both columns and rows of the spreadsheet informative, and details of successive generations of inbreeding kept in separate sheets. Nonetheless, the use of rows is not consistent, with most representing the outcomes of particular mating or littering events, but some detailing subsequent breedings (e.g. rows 17:19) or apparent duplicates (e.g. rows 24:25).

Data structure

Before acting to clean the data for analysis, it is useful to consider the structure of the breeding records. Unlike a conventional pedigree, the dataset does not record the existence of individual mice. Instead the breeding records are organized to record littering events, although these events are not given unique identifiers; unique events can be inferred by the unique combinations of variables. The data are structured so that each informative row documents a discrete littering event (additional information is logged in incomplete rows and columns, requiring careful curation). The net result is that the dataset logs the characteristics of each existing littering event, along with associated data such as a CC inbred line identifier and generation number. It offers limited direct information on nonexistent littering events (i.e. those in which mating took place, but litters were not brought to term).

While no data exist on individual mice within the CC population, this is not necessarily an obstacle to further investigation. A core assumption of genetic reference populations such as the CC is that individuals within a given inbred line are genetically identical. As each littering event is associated with a CC inbred line identifier, the data can be interpreted as providing multiple measurements from a small number of genetically identical individuals. Knowledge of the exact individuals involved is therefore not necessarily important for analysis of the dataset. One important caveat is that the data log the breeding of the *incipient* CC lines, so the individuals in question are by definition not fully inbred and therefore not genetically identical (although they will be increasingly similar with increasing inbreeding). A second caveat is that the dataset interpreted in this way contains multiple litters from within a CC line (as well as multiple litters from a given mating pair), with the possibility that these measurements are not independent of each other.

Although litter size is likely to be influenced by both parental and offspring factors it is most commonly treated as a characteristic of the parental generation. Some other

phenotypes recorded in the breeding records, such as the time from birth to first mating, are by contrast better defined as characteristics of the fillial generation. In reality, the organization of the records by litter means that some of the traits logged in a given row of the dataset actually represent the characteristics of different individuals (i.e. the parents versus their offspring), which restricts the intuitive interpretation of these data and limits the use of certain parameters as covariates. Owing to the lack of unique identifiers and the absence of a true pedigree, it is difficult to accurately link phenotype values across generations.

The final point to note is that labelling of the generation of inbreeding seems inconsistent among the three workbooks. For the CC lines initiated at TAU the litters that result from the G2 cross are designated as ‘G3’ whereas biologically they make more sense as a characteristic of the G2 cross (and technically the offspring are G2:F1 rather than G3). By contrast, the records for the lines initiated in Kenya seem to label the litters by parental generation rather than fillial generation (although the column headers are inconsistent). I discuss the implications of this possible inconsistency of labelling in detail below.

Cousin lines

The breeding records for the lines initiated at TAU also hold data on a number of ‘cousin’ lines (Figure 8). These cousin lines are pairs of lines that originated from the same founder crosses (e.g. ABCDEFGH founder order) and were subsequently inbred for a variable number of generations before being separated and inbred in parallel. Cousin lines are labeled using the format IL-3XXX (for cousins of lines IL-1XXX) and IL-4XXX (for cousins of lines IL-2XXX). Lines IL-2XXX are the reciprocals of lines IL-1XXX (i.e. if IL-1XXX is ABCDEFGH, IL-2XXX is EFGHABCD).

These cousin lines therefore all appear in the breeding records at generation 4 (G2:F2)

or later. However, the breeding records for the earlier generations are relevant (as cousins were indistinguishable before parallel inbreeding began). I therefore duplicate the breeding records for prior generations to ensure that cousins have full data. The existence of additional cousin lines from the lines initiated in Kenya became apparent when investigating genotypes (see below). The reasoning behind generating and maintaining these cousin lines is somewhat unclear, although they are briefly described in the published CC literature.¹⁰¹

Data cleaning

The three workbooks have been maintained by hand over a considerable period of time, and as a result they contain numerous obvious errors and inaccuracies in data entry. I have invested significant time in systematically identifying and correcting any apparent errors. As the structure of the workbooks does not facilitate computational manipulation, much of the cleaning has been done by hand.

The most notable recurring error is in the recording of dates. The breeding records contain a mix of little-endian and middle-endian (i.e. DD/MM/YYYY vs MM/DD/YYYY) date formats. The sheets are formatted to accept little-endian entries; however, much data entry seems to have been in middle-endian format with Excel automatically changing the format of any entry where the day (DD) entry is greater than 12, which keeping the remainder of entries as they were input. Not all input styles are, however, consistent; for example 3 May is variously recorded as 03/05 and 05/03. As a result the data contain numerous cases of mice that breed before they are born, with subsequent generations born before the prior breeding has occurred. Such cases make little biological sense.

Fortunately, knowledge of the breeding pedigree and the biology of the mouse has enabled correction of most of these errors. Specifically, young mice should be removed from their home cage before the age of sexual maturity, to ensure control over the next generation's

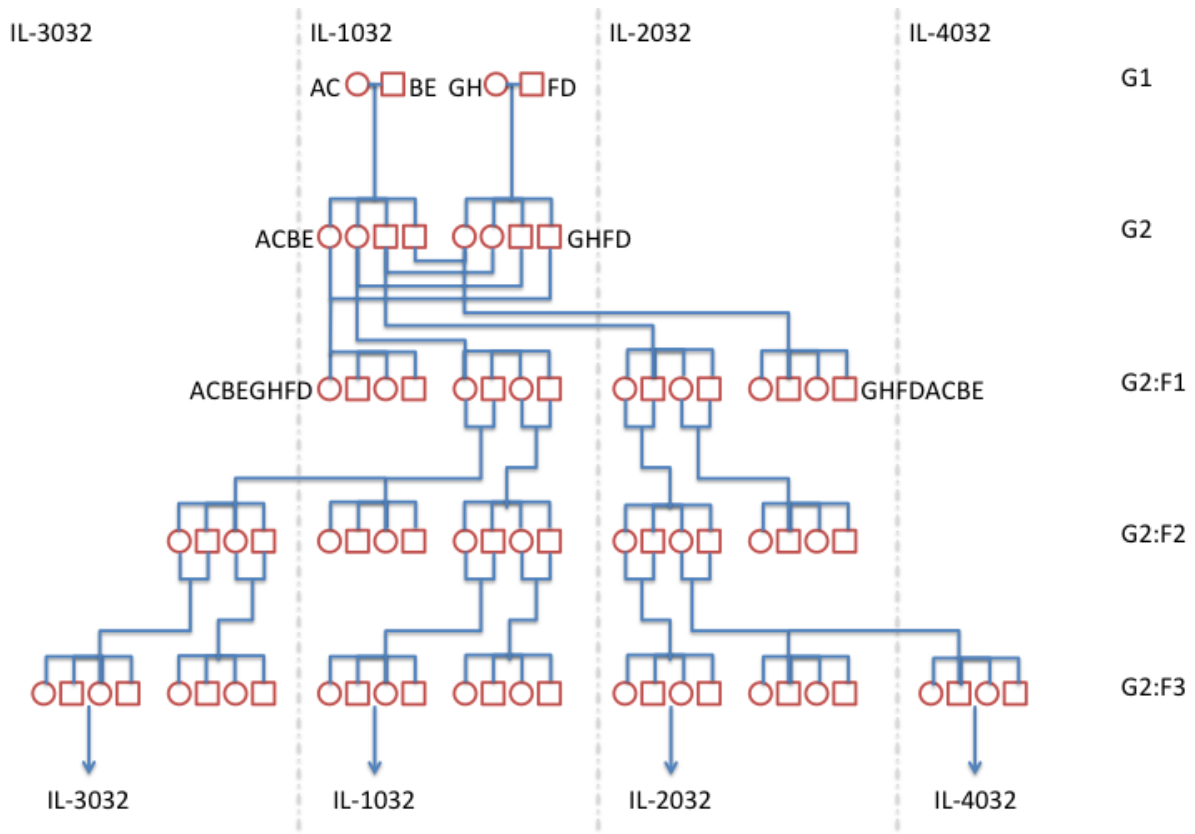


Figure 8: Collaborative Cross breeding and cousin lines. In the CC breeding scheme, pairs of mice are crossed to produce the two-founder G1 (e.g. AC), which are then crossed to produce the four-founder G2 (e.g. ACBE). G2 mice are crossed to produce the eight-founder G2:F1, which are then sibling-mated in order to generate inbred lines. In this example, G2 mating of ACBE with GHFD mice creates two different G2:F1 lines, designated IL-1032 and IL-2032, depending on whether the ACBE mice were female or male, respectively. From within the G2:F1 litters, two or more new sibling matings are set up, with one carried forward into the subsequent generation of inbreeding. Sometimes, G2:F1 sibling matings are used to initiate a cousin line, which is inbred in parallel with the original line. For example, inbred line IL-3032 is a cousin of line IL-1032. Siblings from among the G2:F1 were crossed to produce the G2:F2 of IL-1032. Further offspring of the G2:F1 siblings (it is not clear whether from the same mating pair or different, or whether this is consistent across lines) were labelled IL-3032 and inbred in parallel. Thus IL-1032 and IL-3032 are cousins that share a G2:F1 and I therefore duplicate this generation in our records. IL-2032 and IL-4032 are also cousins, with IL-4032 initiated one generation later than was IL-3032.

breeding. This removal happens within approximately 30 days of birth. By systematically searching for all cases where the recorded time interval is too long or too short I have been able to identify suspicious date entries. In addition the records are largely ordered chronologically, which provided extra information to identify date entry errors. Ultimately, swapping the date and month on these suspicious date entries resolved the vast majority of implausible birth–mating and mating–birth intervals.

Finally, the workbooks contain numerous incomplete rows and empty cells. It is not immediately clear whether these gaps are informative (perhaps representing unsuccessful breeding attempts) or whether they represent individual styles of data entry among personnel. During data cleaning I have retained these empty rows for further investigation. These incomplete rows do highlight a particular issue of working with reproductive phenotypes, which is that unsuccessful and difficult matings are likely to be informative. At this stage it is unclear how much data exists on unsuccessful attempted matings, and whether mating producing no offspring from nonattempted mating (i.e. 0 vs NA values) can be differentiated.

Data cleaning took place via several cumulative steps (file and version names given in square brackets):

1. Data from the individual generation sheets were combined into one large sheet. Unique record numbers were added to allow safe reconstruction of original data if necessary, and potentially to facilitate pedigree building at a later date [v1].
2. General data tidying was done, including removal of obviously unnecessary entries and deletion of uninformative ‘XXX’ from breeding records [v2].
3. The ‘A’ rows were carefully combined with the spurious ‘B’ rows. Now, rather than having A and B subcages noted in separate rows, where necessary these have been combined into a single record, with the subcage designation ‘AB’. Extra rows were deleted [v3].

4. Empty and duplicate rows were identified and highlighted, along with anything else that needed attention [v4].
5. Empty lines with birth dates matching the previous record were marked as ‘DUP’, as likely to be a duplicate record. Any empty records past the last breeding generation +1 were deleted [v5].
6. A final file was saved as a .csv for analysis with R [CCdataclear]. This file was used for all initial analyses.
7. During preparation of this thesis a further final file was generated for archiving with the thesis and sharing with subsequent publication. Some uninformative columns were deleted, as were duplicate rows. Missing reciprocals (i.e. ABCDEFGH designation) were input from cousin lines. Dates of mating were duplicated across generations to facilitate analysis of time to first littering. Dates were corrected when input incorrectly, or when reformatting from MM/DD/YYYY to DD/MM/YYYY format was consistent with the chronology of the pedigree. Additional ‘0’ columns replace NAs with 0 values in cases where litters have birth dates but no pup counts. First matings are labelled and date subtractions done. NAs were added to all empty cells [ccdata_final].
8. Additional breeding records for the TAU-NEW data (up to summer 2013) were later added to the file [ccdata13_final].

The final dataset contains 16,550 record rows, documenting littering events and associated information. Its structure is illustrated below, along with an explanation of the variables (Table 4). From the available data, the reproductive outcomes of interest are litter size, number of weaning males and females, and time to first litter. Secondary outcomes are time to subsequent litter, average time to litter, maximum generation number reached, time per generation, and total number of offspring produced. I initially focus on litter size, as this a straightforward and common phenotype investigated in mice. The final dataset will be corrected to the appropriate YYYY-MM-DD format as per ISO 8601 and

archived in an online repository.

/data/NGK/SEED/data/2015-10-14_ccdata13.csv

/data/NGK/SEED/data/ccdata.rds

##	Funnel	Reciprocal	Subcages	Gen	DMating	DBirth	Days_mate2birth
## 1	IL-1011	ABECGFDH	A	3	01/04/09	02/05/09	31
## 2	IL-1011	ABECGFDH	A	3	01/04/09	26/05/09	55
## 3	IL-1011	ABECGFDH	A	3	01/04/09	21/06/09	81
## 4	IL-1011	ABECGFDH	A	3	01/04/09	13/07/09	103
## 5	IL-1011	ABECGFDH	B	3	01/04/09	13/05/09	42
## 6	IL-1011	ABECGFDH	C	3	<NA>	<NA>	NA

##	First_litter	Litter	Wmale	Wfemale	Litter0	Wmale0	Wfemale0	NxGen	
## 1		1	7	2	4	7	2	4	G4
## 2		0	7	4	3	7	4	3	G4
## 3		0	9	4	4	9	4	4	G4
## 4		0	8	4	4	8	4	4	<NA>
## 5		1	5	3	2	5	3	2	<NA>
## 6		NA	NA	NA	NA	NA	NA	NA	<NA>

##	NxSubcages	NxDMating	Days_birth2mate	Source	Dup_cousin	Cousin	ID
## 1	A	27/05/09		25 TAU-NEW	0	0	1
## 2	B	28/06/09		33 TAU-NEW	0	0	2
## 3	C	16/07/09		25 TAU-NEW	0	0	3
## 4	<NA>	<NA>		NA TAU-NEW	0	0	4
## 5	<NA>	<NA>		NA TAU-NEW	0	0	5
## 6	<NA>	<NA>		NA TAU-NEW	0	0	6

Table 4: Descriptions of variables in the CC TAU dataset

Variable name	Description
Funnel	CC inbred line identifier
Reciprocal	Description of mating scheme
Subcages	Indication of mating pair (i.e. the cage they were assigned); A-E
Gen	Generation of inbreeding ("G3" litters are in fact the offspring of G2 mice and thus not inbred)
DMating	Date mating was initiated
DBirth	Date of birth of litter
Days_mate2birth	Time in days from initiation of mating to birth
First_litter	First litter status; 1 = yes, 0 = no
Litter	Litter size
Wmale	Number of weaning males
Wfemale	Number of weaning females
Litter0	Litter size with 'NA' replaced by '0' (unused)
Wmale0	Number of weaning males with 'NA' replaced by '0' (unused)
Wfemale0	Number of weaning females with 'NA' replaced by '0' (unused)
NxGen	Next generation of inbreeding; only exists for mice that bred onwards
NxSubcages	Indication of subsequent mating pair (i.e. the cage they were assigned); A-E
NxDMating	Date next mating was initiated
Days_birth2mate	Time in days from birth to initiations of next mating
Source	Workbook origin of original data; KEN-OLD, TAU-OLD, TAU-NEW (as in Figures 5-7, respectively)
Dup_cousin	Early cousin generations duplicated in the dataset; 1 = yes, 0 = no
Cousin	Is a cousin; 1 = yes, 0 = no
ID	Unique litter identified assigned during data cleaning

Exploratory data analysis

Before attempting to map reproductive traits, I first examine the data using extensive analysis and visualisation. I use exploratory data analysis (EDA) techniques to look for obvious errors in data entry, assess assumptions about the data, determine relationships among explanatory variables, and assess the direction and size of any such relationships. I also use this exploratory process to identify appropriate hypotheses and models for subsequent genetic mapping. All data analysis was done using R version 3.1.3,¹⁶⁵ distributed as Revolution R Open version 8.0.3 (Revolution Analytics, Redmond, WA, USA).

EDA is an approach championed by the US mathematician John W Tukey (1915–2000) that uses numerical and graphical techniques to interrogate data with minimal prior assumptions. Tukey himself describes EDA as “detective work”, analagous to the search for evidence in a criminal investigation.¹⁶⁶ More-generally, EDA typically refers to a set of techniques used in the relatively unbiased investigation of distributions and relationships within datasets, with a strong emphasis on visualisation rather than formal significance tests. A number of resources, both historic^{164,166,167} and contemporary^{168–172} outline the principles and practices of EDA.

The term EDA has become increasingly fashionable in the past few years, and the field has undergone something of a renaissance driven by the growth of the data analytics and visualisation industries and the ever-growing availability of large datasets. Most researchers employ some form of EDA in their work; however, reports of these processes are very rarely included in published papers, with the result that the structure and management of datasets are rarely discussed. As this project illustrates, however, understanding these features of a dataset is important for drawing correct conclusions and offering biological insight.

Initial observations

In total, I obtained breeding records from 460 CC lines initiated or maintained at TAU, albeit some of which have very limited data. Each line has a mean of 36 litters recorded. The distribution of litter records is skew, with many lines having few recorded litters, and only a small number of lines having very large numbers of recorded litters (Figure 9). The breeding data comprise 16,550 litter records, of which 14,712 (88.9%) contain obviously meaningful data. The remaining records seem largely empty, but have been retained in case they represent failed matings or are otherwise informative. As previously noted, the breeding records have been maintained in separate workbooks, representing different breeding histories and record-keeping practices. In total I have 1,219 records from the lines initiated and maintained in Kenya; an additional 5,559 records cover these lines after their transfer to TAU. 9,772 records are from the lines initiated and maintained at TAU.

While some CC lines have reached very high numbers of generations of inbreeding, the majority of lines in the dataset are less than 10 generations inbred (mean 9.04 generations; Figure 10). Some of these lines will have died out, whereas others will have continued the inbreeding process since the dataset was censored in 2013. Thus the majority of litters in the breeding records are from the early stages of inbreeding (Figure 11). Note, I have not adjusted the generation numbering, so G3 actually denotes the results of the outbred G2 cross, rather than the third generation of inbreeding; that is, the CC lines are less inbred than the generation labelling suggests.

Litter size

As litter size is the most obvious trait detailed in the breeding records, I start the examination of phenotypes with this trait. Litter size in the complete TAU cohort ranges from 0 to 19, with a mean of 5.69 and an interquartile range of 3–8 (Figure 12). These

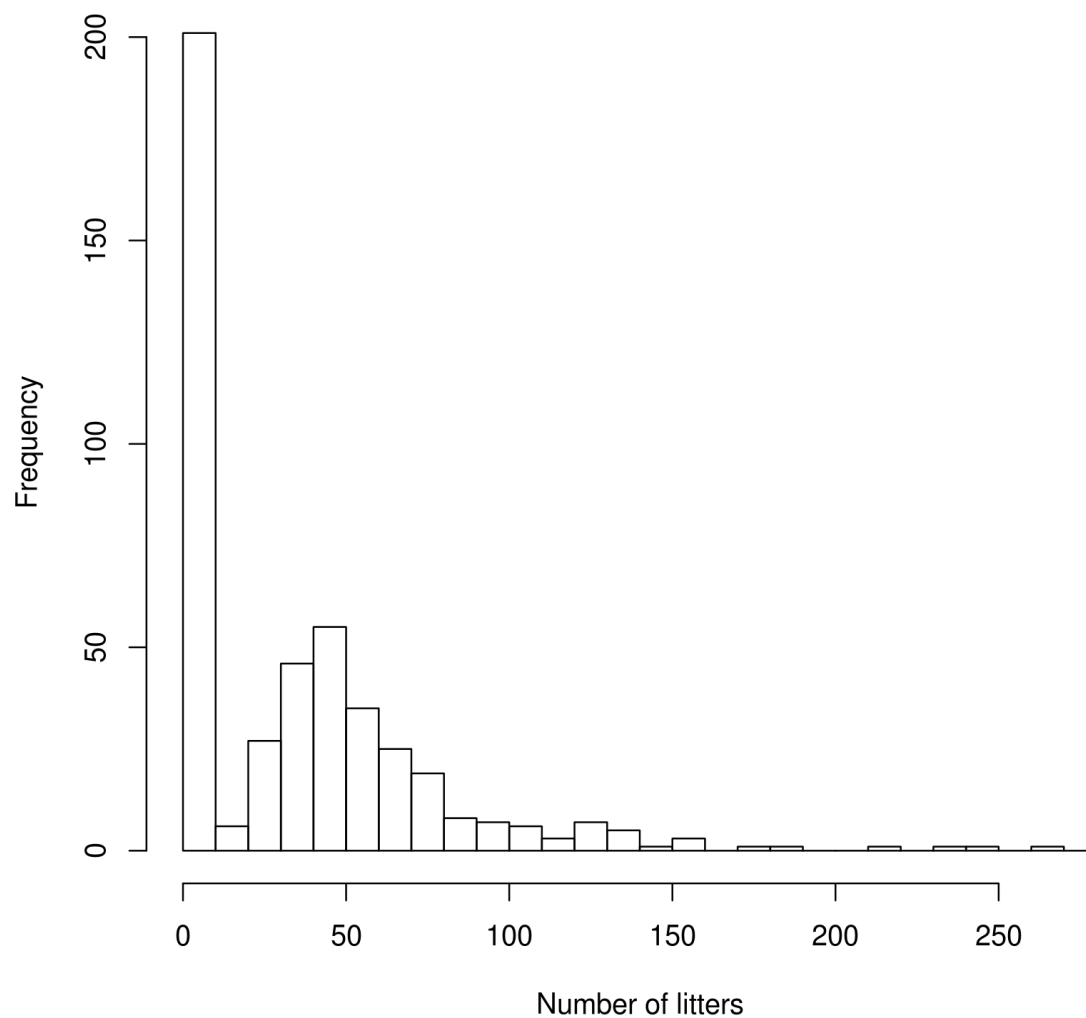


Figure 9: Number of litters per inbred CC line. A large number of inbred lines have fewer than 10 recorded litters, and few inbred lines have more than 100 recorded litters.

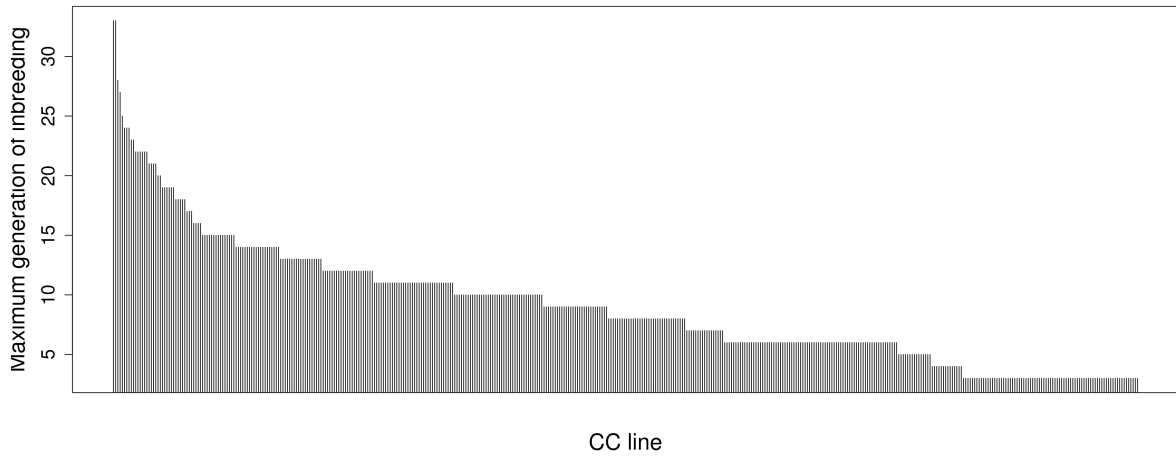


Figure 10: Maximum generation of inbreeding per CC line. The majority of CC lines are less than 10 generation inbred.

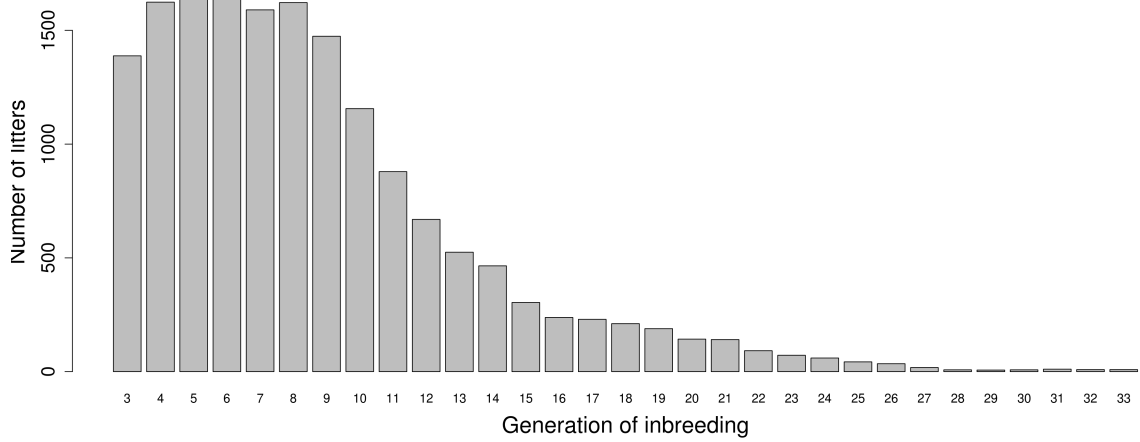


Figure 11: Number of breeding records per generation. Most of the litters in the dataset are from the early generations of the inbreeding process.

observed litter size parameters are consistent with our expectations based on other mouse breeding populations. The majority of recorded litter sizes are integers as expected from count data. However, a small minority of litter sizes are reported as fractions (seen in the rug under the histogram on Figure 12). On examination, all the relevant records are from the first generation of the lines initiated and maintained in Kenya, where some litter sizes have potentially been recorded as means or other summary statistics. Noninteger data are also seen for the recorded number of weaning males and females in these records.

In general, the numbers of weaning males and females are consistent with the reported total litter sizes, and no excess of one sex is seen. Nonetheless, in several cases the total number of weans is reported to be greater than the birth litter size. Figure 13 shows the reported litter size plotted against the sum of male and female weans. As expected, the number of reported weans is generally the same as or smaller than the total reported litter size. A small number of datapoints ($n = 136$) are plotted above the $x = y$ line, indicating errors in the records. Interestingly, a number of large litters report no weans, suggesting a possible phenotype for follow-up.

Calculated dates

Similarly, the two calculated date columns (representing days from mating to birth, and days from birth to subsequent mating) contain values that are biologically implausible (Figure 14). While manipulations of date formats corrected many such errors, a number of recorded date values offered no plausible correction. The onset of puberty varies between animals and among inbred strains, but typically occurs 6–8 weeks after birth in females and 34–38 days after birth in males.⁸⁰ Thus young mice in breeding populations are typically separated by sex and new matings initiated before these ages are reached. I conservatively mark a boundary of 34 days on Figure 14a. While the majority of matings are recorded as having been initiated within this time, a number of records show a substantially longer period. Similarly, a number of records show very short and even

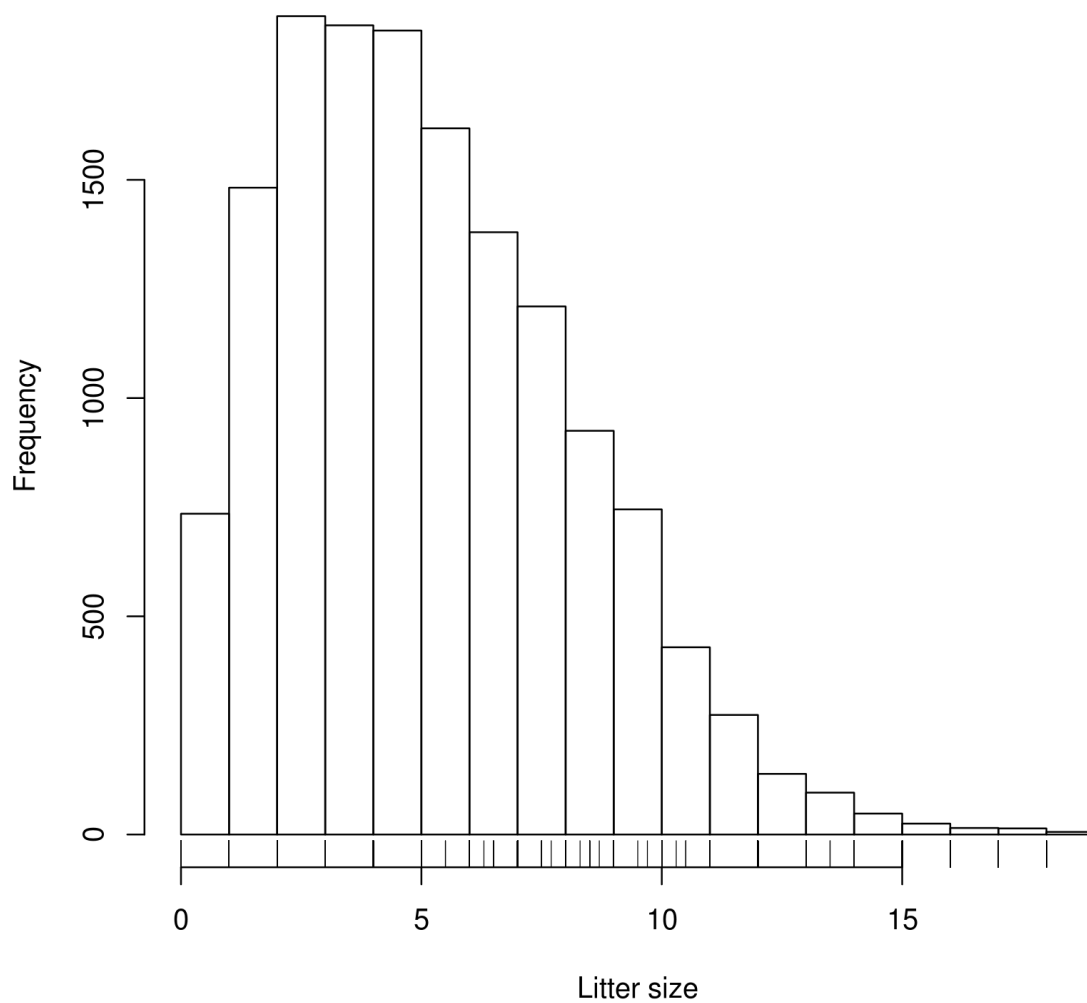


Figure 12: Litter size in the CC breeding records. The Figure shows the distribution of litter sizes in the complete dataset. The rug underneath the histogram shows the individual datapoints and indicates a number of non-integer values, which are biologically implausible.

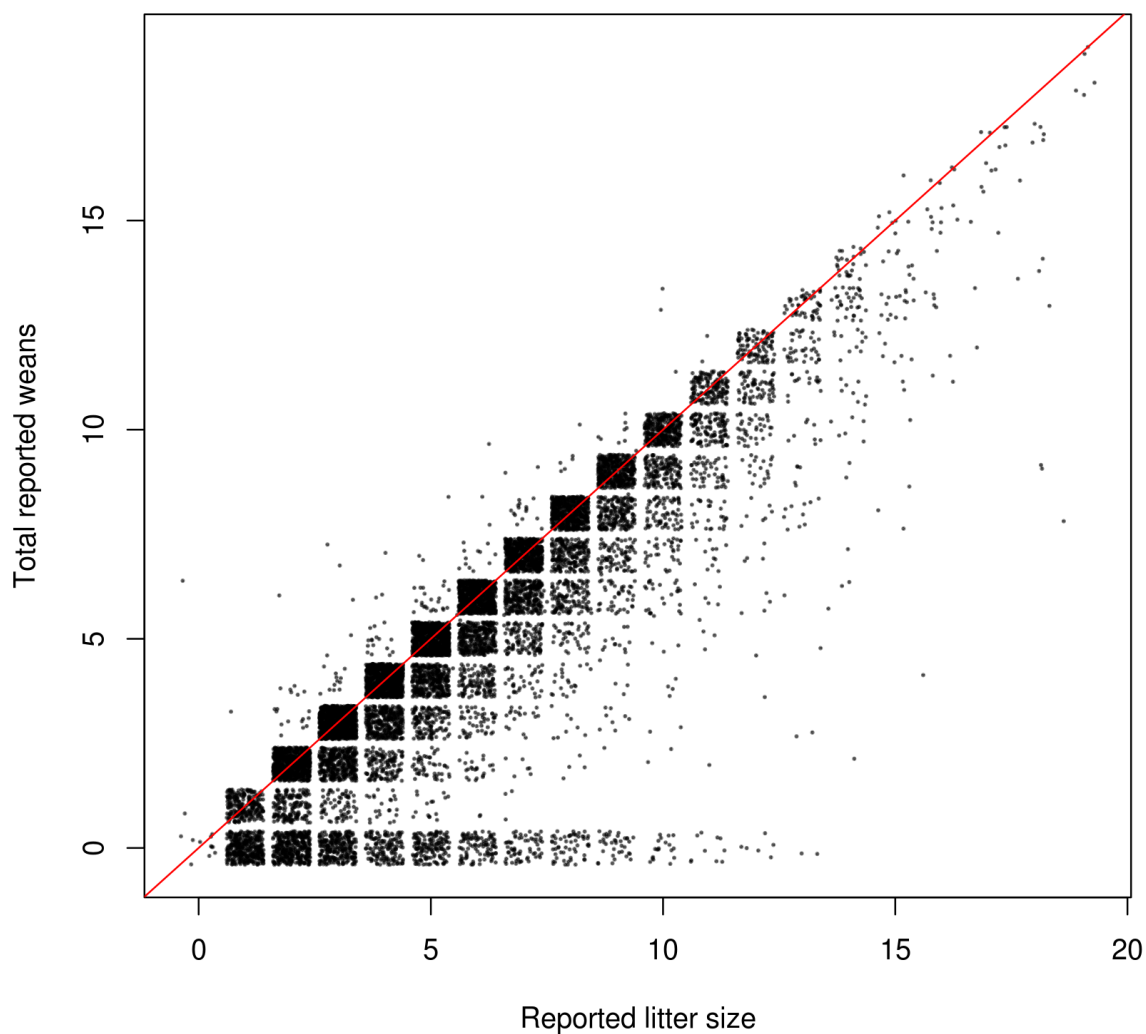


Figure 13: Total litter size vs weans in the CC breeding records. The reported litter size from the CC breeding records is plotted against the sum of the reported numbers of male and female weans. The data are jittered to facilitate visualisation of overlapping points. The $x = y$ line is marked in red. The total number of weans is expected to lie on or beneath this line in all cases. In fact, a number of points are seen above this line, indicating errors in the breeding records.

negative durations. Incidentally, the median and interquartile statistics indicate that these data denote the date that mating was initiated (i.e. well within 34 days), rather than when pregnancy was achieved.

The recorded time from mating to birth also contains implausible values. The gestational period in mice is 18–22 days.⁸⁰ I mark the lower bound of this range on Figure 14b, and see that a number of recorded durations are shorter than 18 days (again including negative values). Among the litter records, I have noted which are first litters (i.e. earliest birth dates for a given mating), and which are subsequent litters. This information is only generally available for the lines initiated and maintained at TAU. In total, 2,724 litters are noted as first litters, 5,701 are subsequent litters, and 8,125 have no information. As matings were set up before mice reached puberty, the mating-to-birth parameter for first litters includes an unspecified period of time until the mice became sexually mature, in addition to a period of time before successful mating occurred. Thus, while time to birth of first litter seems longer than for subsequent litters (Figure 15), some of this difference is an artifact of the data recording process.

Suitability for mapping

Mean litter size varies by CC line, generally within the range of the founder strains (Figure 16; parent strains not shown). Within-strain variance (shown as SEM) is generally low, indicating that litter size is amenable to QTL mapping. I see an indication that within-strain variance increases slightly with increasing litter size ($r = 0.126$, $P = 0.0184$) These observations are consistent with the first litter data from the UNC lines.¹⁶² Fitting a simple linear model suggests that CC inbred line (denoted as ‘Funnel’) is a significant predictor of litter size ($F_{(371, 14,340)} = 7.39$, $P < 2.2\text{e-}16$), indicating a line-specific (i.e. genetic) component that can be investigated further. Similar parameters for other simple reproductive phenotypes are given in Table 5, suggesting that many of these traits are also amenable to mapping.

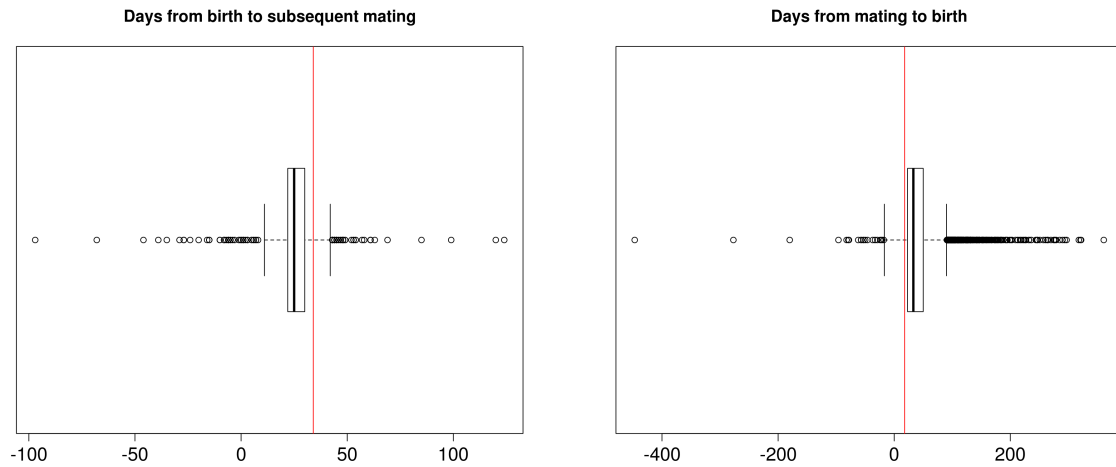


Figure 14: Calculated dates in the CC breeding records. The left panel (a) shows the number of days from birth of a litter to the time the subsequent sibling mating was set up. Some of the values are reported as <0 days, which is biologically implausible. The red line at 34 days indicates the onset of male puberty. Theoretically, new matings should have been initiated before this point. The right panel (b) similarly shows the number of days from the initiation of mating to subsequent birth. The red line at 18 days marks the lower bound of the reported gestational period. Some data points are <18 days, and these include negative values.

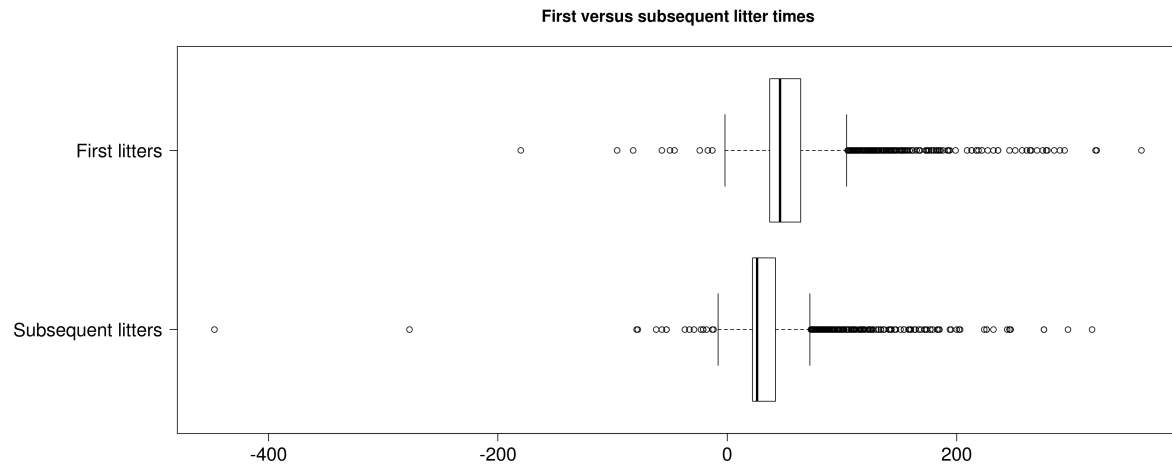


Figure 15: Days from mating to birth for first versus subsequent litters. The number of days from mating to birth is significantly greater for first litters. This difference is attributable to the fact that mating pairs are generally set up before the mice are sexually mature, and the measured duration includes a period of time before the first mating can take place.

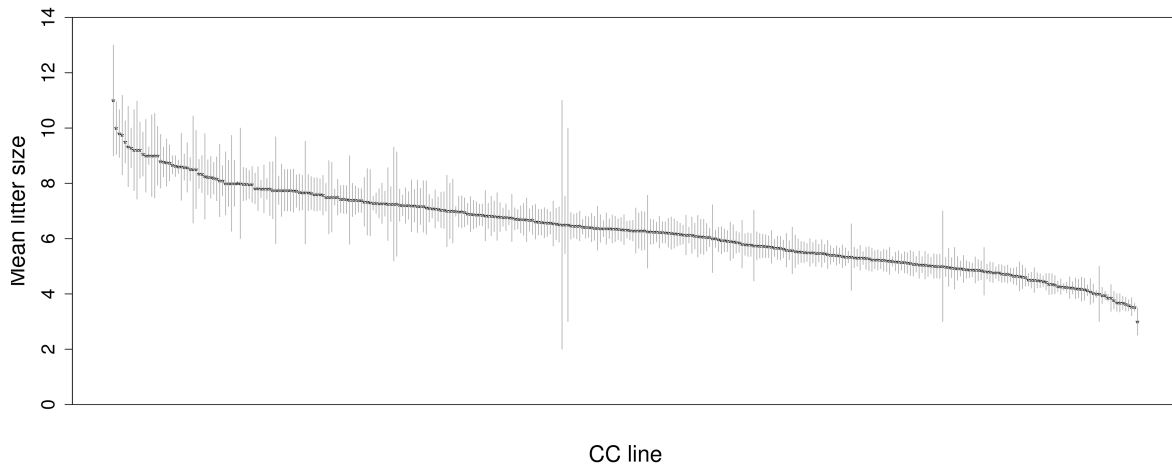


Figure 16: Litter size of incipient TAU Collaborative Cross lines. The lines are ordered by decreasing mean litter size. Within-strain variance (shown as SEM) is generally low, indicating that litter size is amenable to QTL mapping.

A number of other phenotypes can be derived from the data. For example, the numbers of weaning male and female offspring could be considered as a ratio or proportion of total litter size rather than a difference. Similarly, the number of total weans could be considered as a proportion of the reported litter size, or could be otherwise analysed in the context of overall litter size. Litter size itself might be better expressed as a litter size index that takes into account information from relatives,⁴⁰ or by measuring total offspring production or the rate of offspring production for a line. Failed matings could be accounted for by inserting zero litter size values at ~30-day intervals (or another appropriate distance) between successful matings.

Table 5: **Reproductive phenotypes in CC breeding records**

Phenotype	Mean	Median	Range	IQR	Df	Proportion SS	F-value	P-value
Litter size	5.69	5	0,19	3,8	371, 14,340	0.161	7.39	<2.2e-16
Weaning males	2.39	2	0,12	1,4	358, 12,834	0.143	6.01	<2.2e-16
Weaning females	2.4	2	0,12	1,4	358, 12,824	0.163	6.96	<2.2e-16
Sex difference	-0.0104	0	-10,9	-1,1	358, 12,824	0.0356	1.32	5.14e-05
Total weans	4.79	5	0,19	2,7	358, 12,824	0.194	8.62	<2.2e-16
Not weaning	0.926	0	-6,13	0,1	358, 12,824	0.0817	3.19	<2.2e-16
First litter size	6.19	6	1,19	4,8	180, 2,532	0.0847	1.3	0.00558
Subsequent litter size	6.33	6	0,19	4,9	171, 5,478	0.107	3.82	<2.2e-16
Days to litter	56.2	46	-180,361	37,64	169, 8,148	0.0414	2.08	1.77e-14
Days to mating	26.1	25	-97,124	22,30	264, 3,725	0.0838	1.29	1.51e-03

The recorded litter sizes currently comprise multiple litters from the same parents, denoted by subcage. Subcage-level phenotypes might therefore be appropriate for mapping. These phenotypes include total offspring production; number of litters produced; time from mating to next generation mating (time per generation); line survival (a binary measure of at least one viable litter produced); and variation in litter size. It might also be possible to analyse line-specific traits, such as average time to produce a litter, total offspring production, total number of litters; variation in litter size; and maximum generation achieved.

To facilitate analysis of the calculated phenotypes (i.e. sex difference, total weans, not weans), I add these parameters to the data file. Other secondary parameters can be calculated for specific analyses. I also remove all calculated day values below 0 days from the dataset. For the moment, other outlier values are retained and adjusted as necessary on a per-analysis basis.

Covariates

I next explore associations between the variables in the dataset, to check overlap in outcome variables and to identify covariates for inclusion in the mapping models. I continue to treat litter size as the main outcome variable.

First, as expected, the numbers of male and female weans are each associated with total litter size ($r = 0.755$ and $r = 0.753$, respectively). The numbers of male and female weans are less well associated with each other ($r = 0.485$; Figure 17a). The difference between the numbers of male and female weans is well centered on 0 (Figure 17b); however, this difference also understandably tends to increase with litter size (Figure 17c). The number of reported weans is closely associated with the total reported litter size. However, as previously identified a notable cluster of records has no reported weaning despite large litter size (Figure 13).

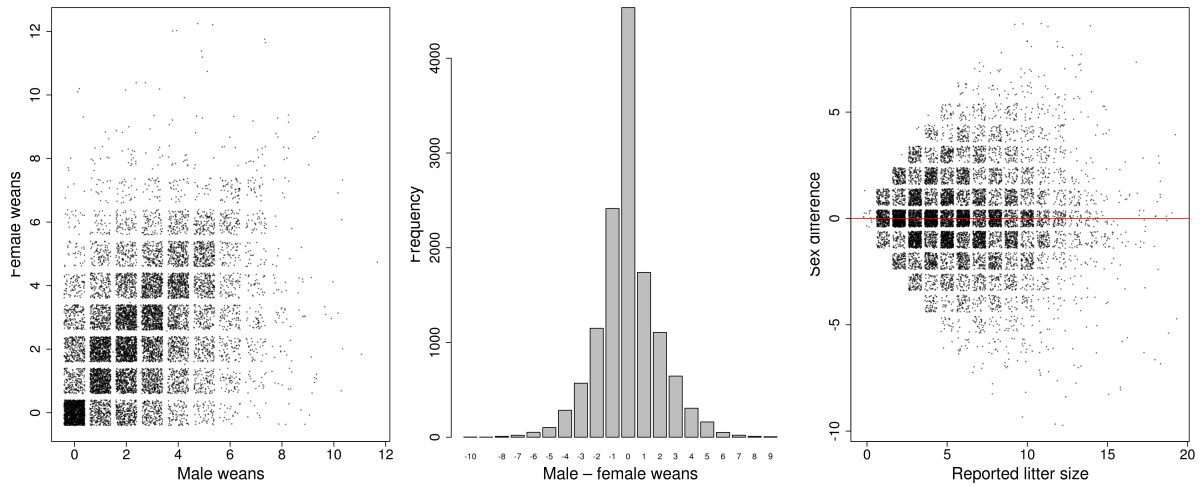


Figure 17: Male vs female weans. The left panel (a) shows the number of male weans plotted against the number of female weans for each litter. The data are jittered to facilitate visualization of overlapping data points. The center panel (b) shows the difference between the numbers of male and female weans for each litter. This difference is well centered on 0, but extreme values exist. The right panel (c) shows the difference between numbers of male and female weans plotted against total litter size. This sex difference is capped at low litter size, constrained by the total number of offspring. The data are again jittered to facilitate visualization of overlapping data points.

The time from birth to mating shows no notable association with litter size (Figure 18a). By contrast, increasing time from mating to birth seems associated with smaller litter size, suggesting a tendency for the later litters of a mating to be smaller (Figure 18b). Earlier litters show more bias to one sex and slightly more offspring that do not wean (not shown); however, these effects are confounded by the association of early litters with higher litter size (i.e. larger litters show more extreme sex differences and more offspring not weaning). Finally, increasing time between litters (measured as the time between litters born to the same mother) also shows an association with decreasing litter size, suggesting a poor-performance phenotype characterized by difficulty breeding and small resulting litters (Figure 18c).

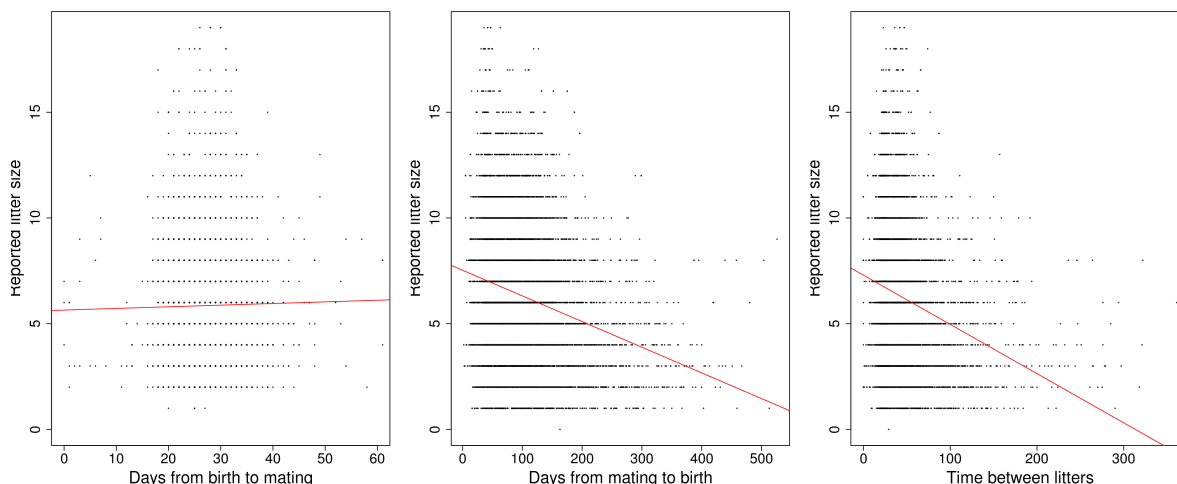


Figure 18: Associations between time and litter size. The left panel (a) plots time from birth to mating against litter size. As expected no significant relationship is seen between these variables. The center panel (b) plots time from mating to birth against litter size. A negative relationship is seen, suggesting that litter size decreases throughout the lifetime of a mating. The right panel (c) plots the time between litters against litter size. A similar negative relationship is seen, suggesting that litters that take a long time to produce are also small. The red lines depict linear regression fits in each case.

Inbreeding

Among possible explanatory variables, I expect the number of generations of inbreeding to be strongly associated with reproductive outcomes. As they inbreed, mice (similar to most species) typically suffer from problems of inbreeding depression, and each CC breeding center has reported such problems. Thus, as the CC lines inbreed, I expect the average litter size, along with other markers of reproductive success, to decrease. The underlying genetic cause of inbreeding depression is thought to be related to accumulation of homozygosity—probably via accumulation of deleterious alleles with recessive effects, or perhaps through a loss of overdominance effects.⁸¹

The data do indeed show a characteristic decrease in litter size with successive generation number, followed by a stabilization at around generation 10 of inbreeding (Figure 19). This decline matches very closely that seen in the UNC lines of the CC.¹⁰⁷ Litter size seems to rise past generation 24 of inbreeding, but this effect is driven by a small number of

lines with good fertility, which bred fast. A similar effect has been previously reported.⁸² Among other outcomes, the numbers of male, female and total weans show a similar relationship with generation, and the difference between numbers of male and female weans remains constant with increasing generation (not shown). Similarly, the number of offspring not weaning decreases only slightly with increasing generation. Both these outcomes will, however, be confounded by the general decline in litter size with increasing generation of inbreeding. Finally, time between litters seems to increase with increasing generation, consistent with increasing problems of fertility with inbreeding (Figure 20). This duration seems to plateau at high generations; however, similar to the late-generation rise in litter size, this effect is probably also driven by a small number of more-fertile lines that breed well.

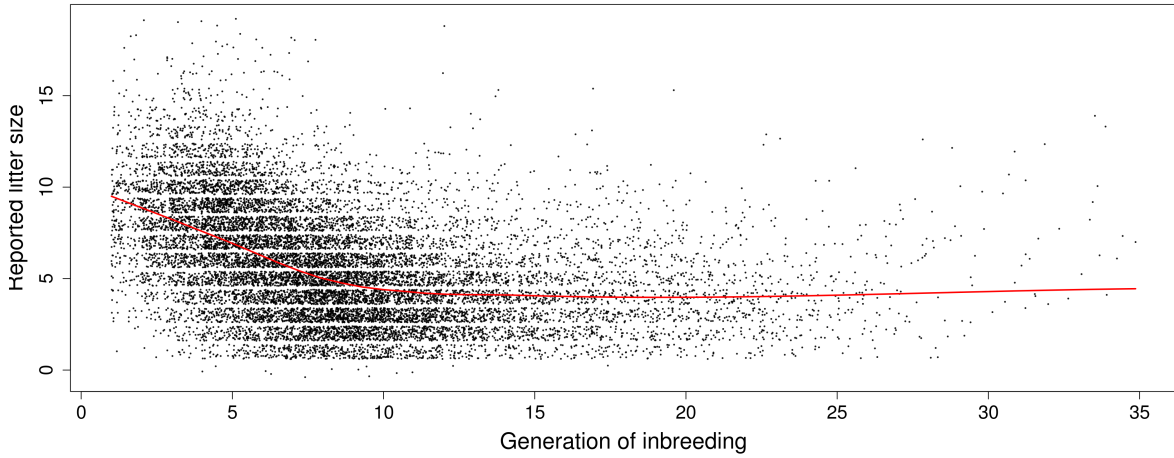


Figure 19: Decrease in litter size with inbreeding. Litter size declines with increasing inbreeding before reaching a plateau at a level around 4 offspring. The slight rise in litter size seen in later generations is driven by a small number of inbred lines that bred well. The red line depicts a loess fit.

Breeding histories

The TAU lines have varied breeding histories. 190 lines were initiated in Kenya and subsequently transferred to Tel Aviv at approximately generation 6 of inbreeding (i.e. around

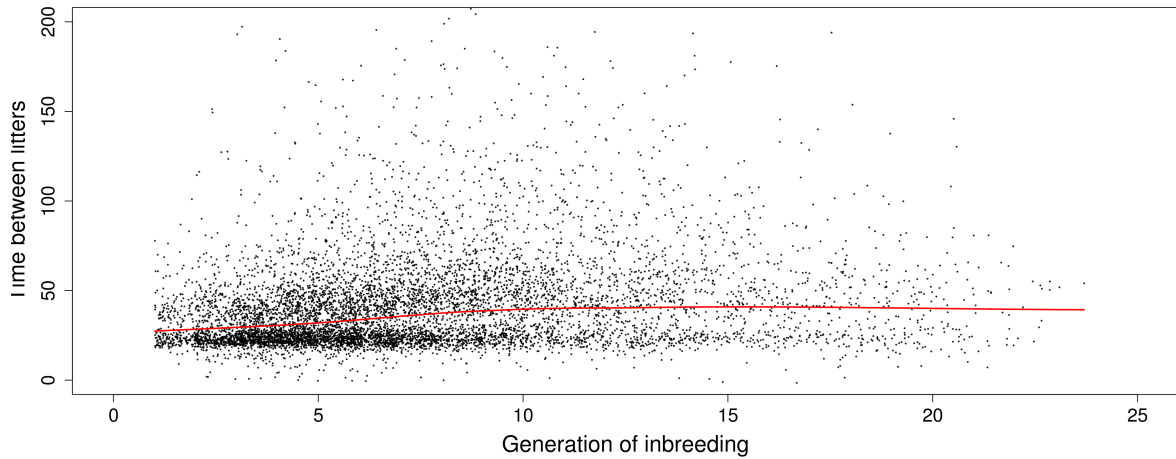


Figure 20: Increase in time between litters with inbreeding. Overall, time between litters increases with increasing inbreeding. The plot indicates that while many times between litters remain around 25 days, some are considerably higher. These higher values might be the result of poor fertility in the inbred lines, or poor record keeping. The red line depicts a loess fit.

G2:F3). The remaining 270 lines were initiated in Tel Aviv in response to the poor breeding of the initial CC lines. Plotting the breeding history (i.e. the variable ‘Source’) against litter size shows a notable difference between groups (Figure 21). However, this variable is confounded with generation: the data from Kenya represent only early generations; data from these lines after the move to TAU represent only later generations; data from lines initiated at TAU represent all generations (but not extreme late generations; Figure 22).

Plotting litter size by breeding history and generation suggests the sets of CC lines have similar breeding performance; however, the lines initiated at TAU produce systematically larger litters than do those transferred to TAU (Figure 22). In fact, place of origin shows a significant association with litter size, with the lines transferred to TAU producing ~ 0.65 fewer pups per litter at each generation compared with the lines initiated at TAU. If anything, the difference is more marked when the number of weaning offspring is considered, with the TAU-initiated lines consistently displaying more weans, and fewer

offspring not weaning.

Such an effect could easily be the result of environmental differences such as feeding or handling regime,¹⁷³ or intestinal biota. Similar environmental effects have been suggested to affect litter size in other long-term mouse breeding experiments.¹⁷⁴ Alternatively, the differences in litter size might reflect differences in generation numbering between the data workbooks, as previously discussed. This possibility is supported by the observation that the within-generation difference in litter size in between populations is of approximately the same magnitude as the per-generation decrease in litter size estimated from a linear model in individual populations.

First litters

Across the whole dataset, I see some evidence that first litters are smaller than subsequent litters (by ~ -0.138 offspring, $P = 0.0689$; Figure 23a). This observation is consistent with our expectations, but the association is not significant when also plotted by generation (Figure 24), indicating that the data contains proportionally more first litters in later generations. By contrast, first litters are associated with significantly longer time between litters than are subsequent litters (by 21.5 days, $P = < 2.2e-16$; Figure 23b). This association is consistent across generations (Figure 25). It results from the fact that the time between litters variable is estimated in a different way for first versus subsequent litters.

Cousins

The data also contain inferred cousin status, and some duplicate records that represent the early breeding generations of cousin lines (among the lines initiated at TAU). Surprisingly, cousin status is associated with higher litter size across generations, presumably because productive lines were more likely to be selected to generate cousin lines (Figure 26). These

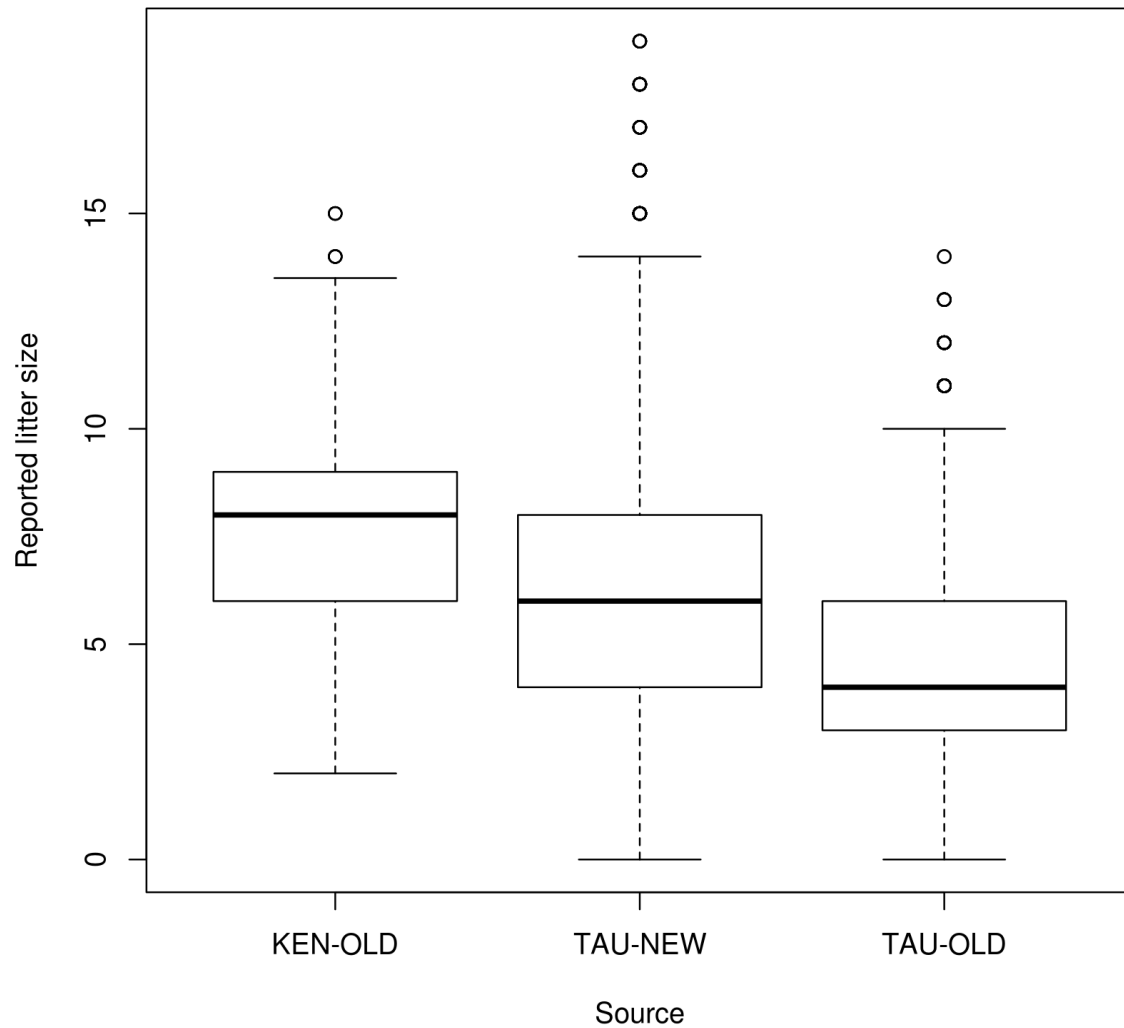


Figure 21: CC litter size by breeding history. The CC inbred lines have different breeding histories in different geographic locations. 190 lines were initiated in Kenya (KEN-OLD), and subsequently transferred to TAU (TAU-OLD). 270 lines were subsequently initiated at TAU (TAU-NEW). Litter sizes vary significantly among these populations, but are confounded with the varying level of inbreeding.

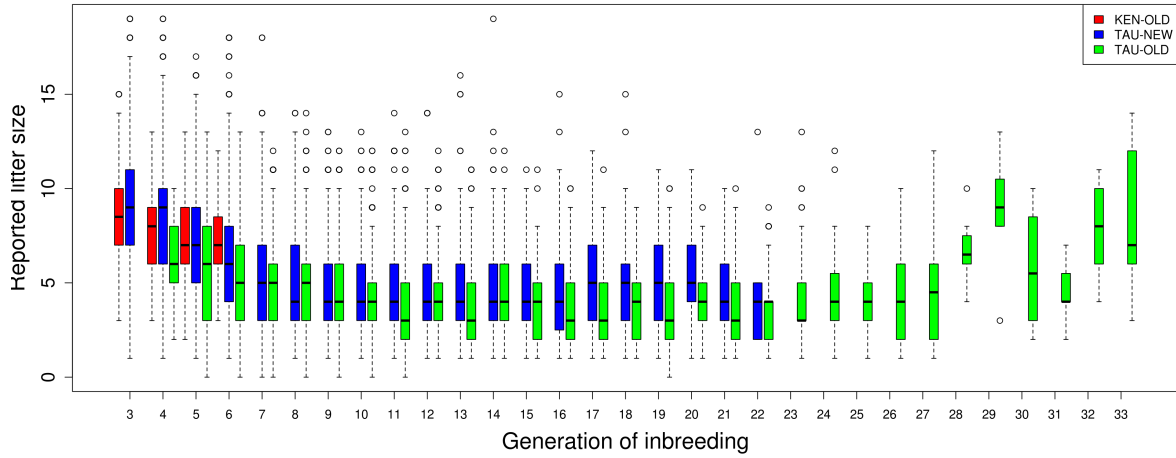


Figure 22: CC litter sizes plotted by generation of inbreeding and breeding history. The boxplot shows that the lines initiated and maintained in Kenya (KEN-OLD) are only represented in the early generations; these lines were transferred to TAU (TAU-OLD). Additional lines were initiated at TAU (TAU-NEW), spanning all generations from 3–22. For any given generation, the lines initiated at TAU have larger litter sizes than those initiated in Kenya, indicating systematic differences between these populations in either generation labeling or biology.

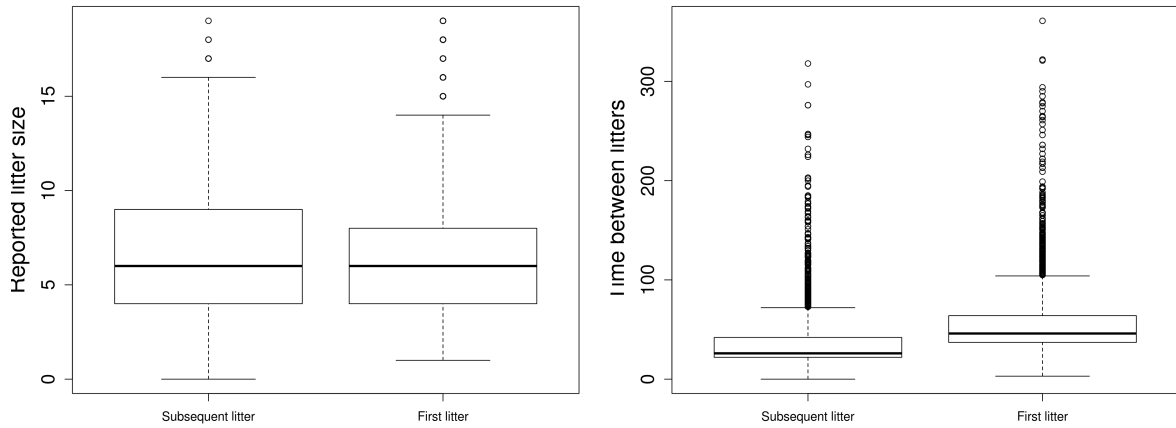


Figure 23: Litter sizes and times between litters for first versus subsequent litters. The left panel (a) indicates first litters are typically smaller than are subsequent litters. The right panel (b) shows that first litters have a significantly longer time between litters than do subsequent litters.

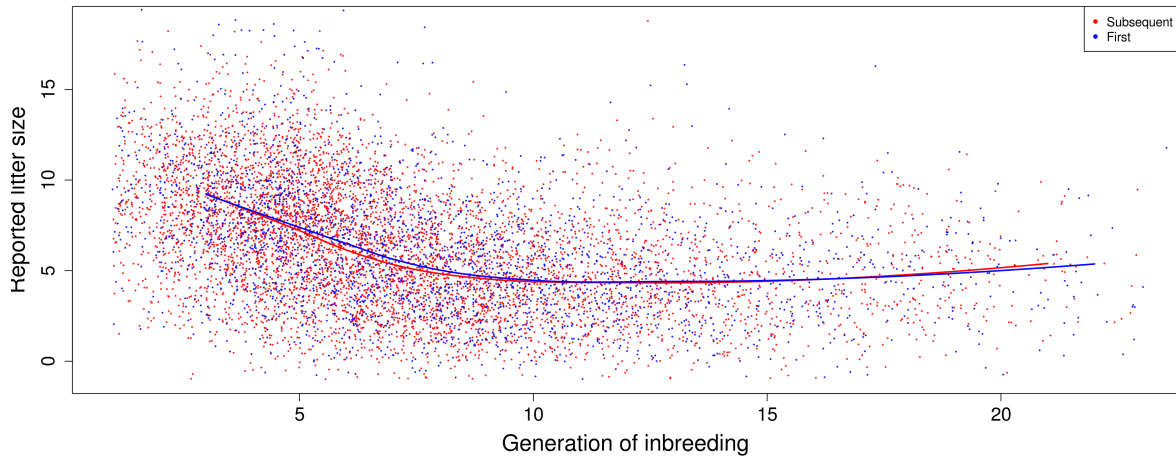


Figure 24: Litter sizes for first versus subsequent litters by generation. When generation is taken into account, first litter status has no association with litter size. The line depicts a loess fit for first and subsequent litters.

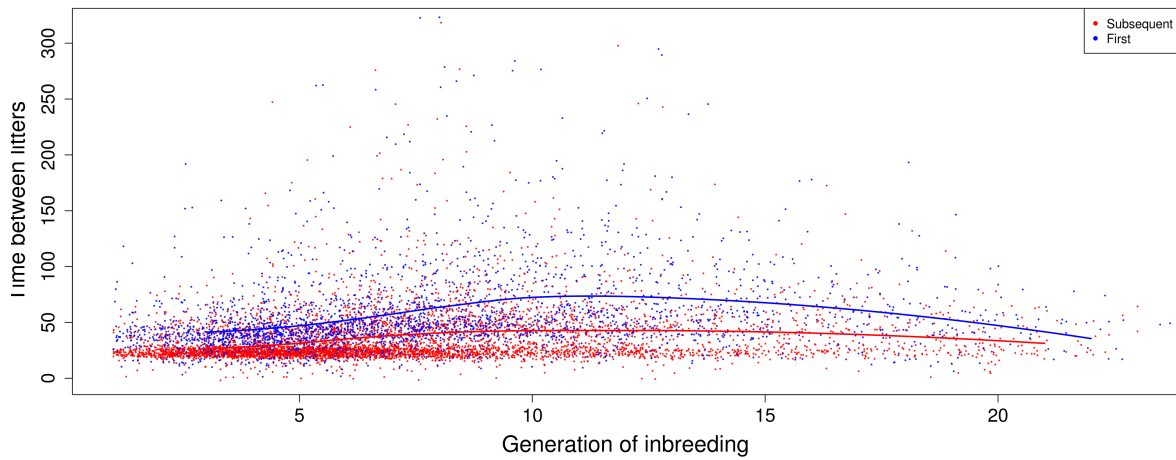


Figure 25: Time between litters for first versus subsequent litters by generation. When generation is taken into account, first litter status retains a significant association with time between litters. The line depicts a loess fit for first and subsequent litters.

cousin lines potentially represent a sample of high-fertility inbred lines. Similarly, the breeding subcage is also associated with litter size, presumably because the need to use C,D and E subcages linked to poor breeding performance (not shown).

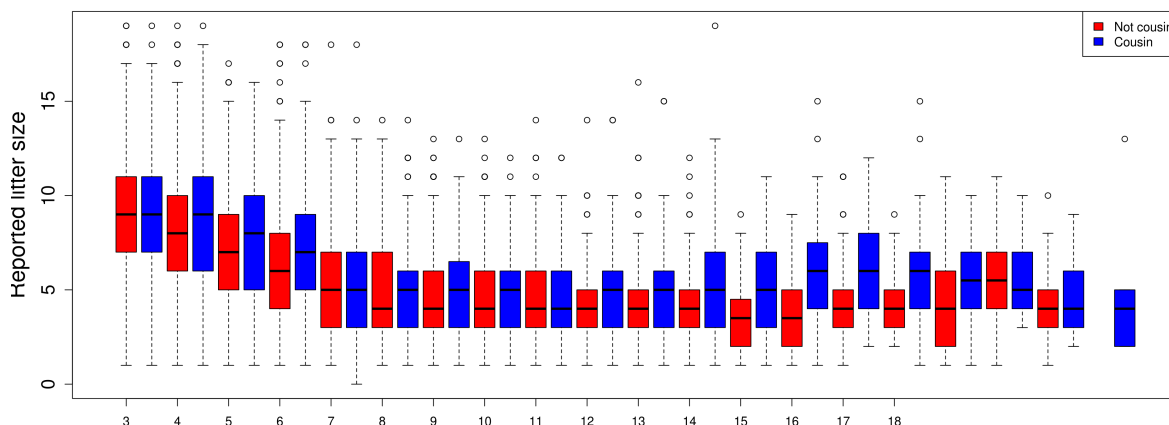


Figure 26: Litter size by cousin status and generation. Cousin lines have higher litter sizes than do non cousins, across generations.

Summary

To check our findings on possible covariates, I compute correlation between all pairs of numeric variables (Figure 27). As expected, the various litter size phenotypes are correlated with each other in the expected direction, and litter size is negatively correlated with generation, and less strongly with time from mating to birth and time between litters. The plot reveals no additional relationships of interest.

After checking all combinations of covariates, I attempt to draw some conclusions about the data. On the basis of extensive plotting and analysis, litter size remains a reasonable outcome variable for assessment. The time from mating to birth (re-expressed as time between litters, i.e. the variable `Gdays_Adj`) is also a relevant outcome. Other outcomes, particularly the relationship between male and female weans, and the number of mice not weaning might also merit further examination.

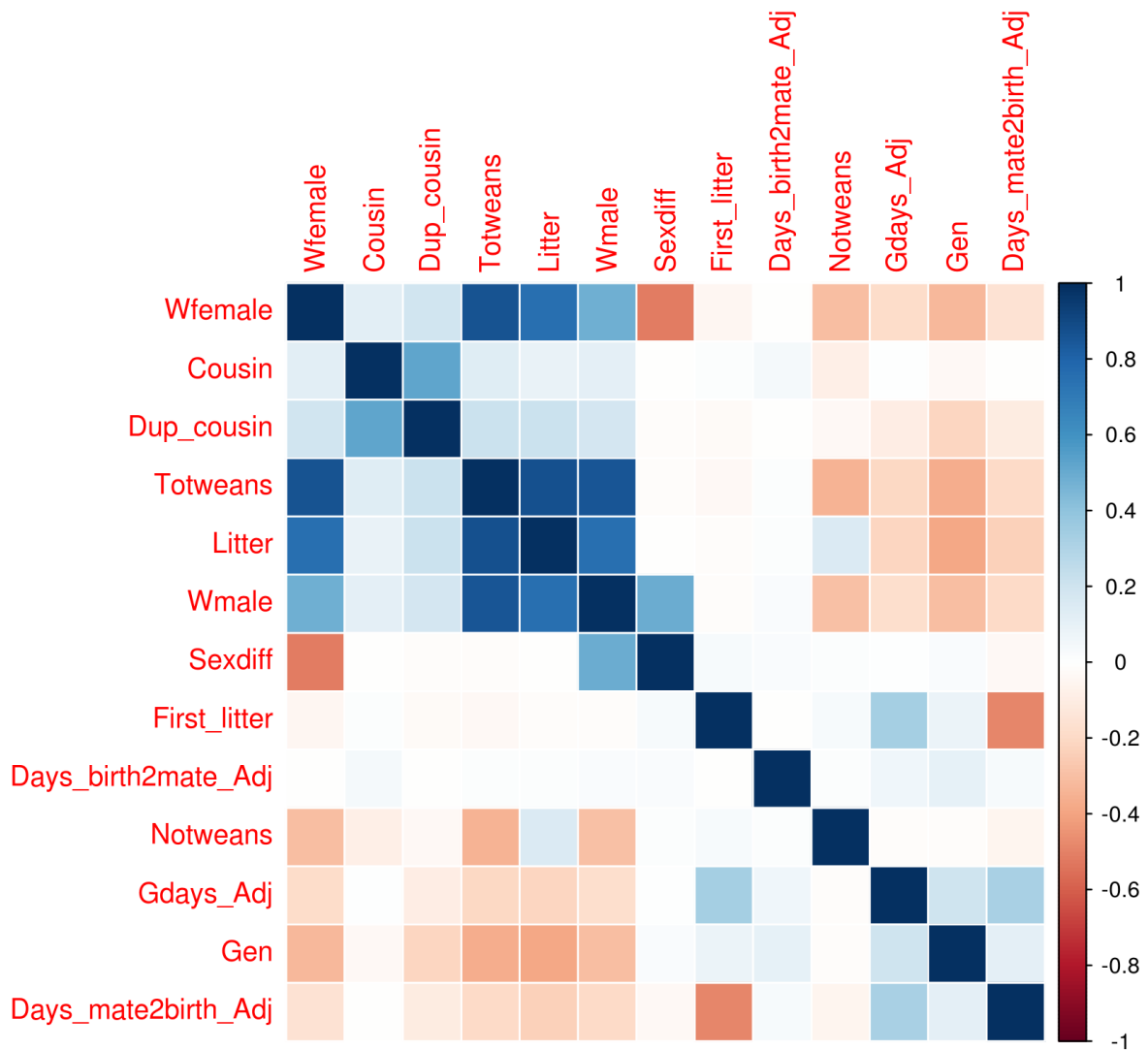


Figure 27: Correlations between variables in the CC TAU data. Litter size shows strong correlations with the number of weans and with generation of inbreeding. Time between litters (Gdays_Adj) is correlated with first litter status.

In terms of explanatory variables, generation of inbreeding is an important predictor of litter size and should be included in any model for mapping. As discussed above, Funnel (as a surrogate for a heritable genetic effect) is also a predictor of litter size in a linear model. On close examination of the linear model terms, this effect is initially driven by the lines initiated in Kenya, and is therefore confounded with the breeding history of the mice (the ‘Source’ variable). Generation and Funnel are also significant predictors of time between litters.

Source is a significant predictor of litter size, and will need to be accounted for either by inclusion in a model, or by adjusting (possibly by altering the generation number, which might be systematically mislabelled) or discarding the affected data. In a linear model with generation and Source, Funnel remains significant, however, with the effect driven by both old and new lines. While first litter status might have been expected to predict smaller litter size, I see only a small effect of borderline significance. First litter is, however, a significant predictor of time between litters. Table 6 shows the proportion of the observed variance in litter size and time between litters explained by these variables in a simple linear model.

Table 6: **Variance explained by model variables**

Variable	Litter size	Time between litters
Gen	0.147	0.0404
Funnel	0.161	0.0472
Source	0.107	NA
First_litter	0.000396	0.111

A number of other factors are typically associated with litter size and reproductive function in mice. For example, increased body weight is typically associated with larger

litters, probably because increased uterine capacity allows more offspring to develop to term. Similarly, litter size in the maternal generation is negatively associated with litter size in the filial generation, probably as a result of body size variations owing to competition for resources. Thus litter size selection experiments typically standardize litter sizes. Unfortunately, as the the TAU data do not identify individuals, they do not provide information on these possible covariates of reproductive traits.

Genotypes

I obtained genotypes for the inbred CC lines from the TAU pre-CC genome cache, maintained at the Wellcome Trust Centre for Human Genetics, Oxford, UK. This cache contains genotypes of pre-CC lines previously used for mapping experiments^{113,114} and interim analysis,¹⁰¹ plus additional lines subsequently genotyped on the mouse-specific Mouse Diversity Array (MDA), Mouse Universal Genotyping Array (MUGA) and MegaMUGA (approximately 550,000, 7,500 and 77,800 SNPs, respectively).^{101,110,150,175,176} Genotype data are stored as a probabilistic reconstruction of the genome mosaic, generated using the HAPPY package.^{177,178}

The TAU CC lines have been genotyped at a variety of generations of inbreeding, data on which are not included in the genome cache. In general, one individual has been genotyped from each line; the assumption is that mice from a given inbred line can be treated as genetically identical to the genotyped individual. Unlike previous CC mapping studies, however, our data include phenotypic measurements made across generations, including those from very early generations of inbreeding before genotyping was done. While the genotyped pre-CC individuals are not completely inbred, a large portion of the genome will be homozygous in any given individual, and a smaller portion of the genome will be fixed in the population. The precise quantities will depend on the generation of inbreeding at the time of genotyping (Figure 28).

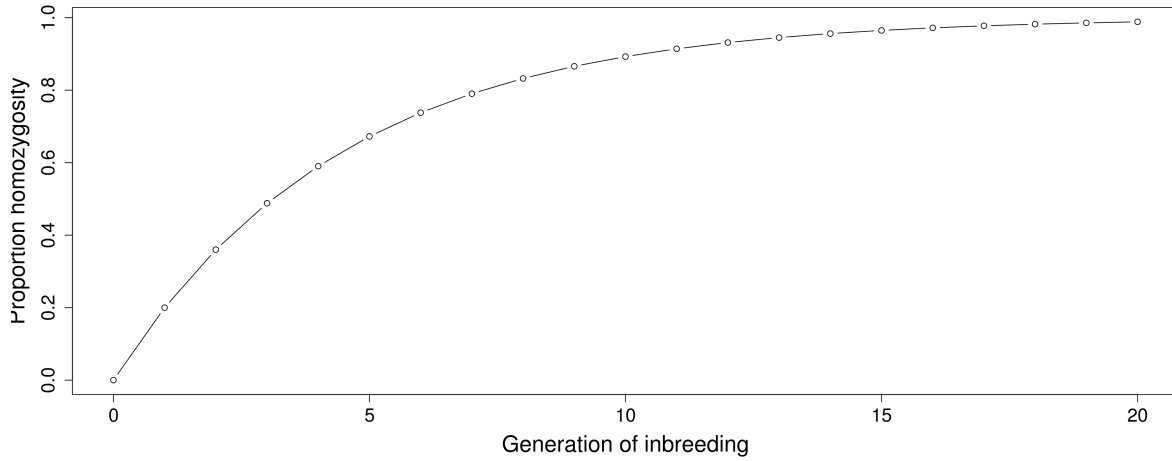


Figure 28: Theoretical homozygosity in each CC line with increasing inbreeding. The proportion of loci homozygous in each individual increases with increasing generation of inbreeding. The proportion of loci in which homozygosity is fixed in the population will be somewhat lower.

Genotyped CC lines

Of the 460 inbred lines in the CC TAU breeding records, 223 had genotype information available for initial analyses (Table 7); these lines represent the vast majority of the total records in the dataset. Each genotyped CC line has reached a minimum of generation 6 of inbreeding (in the notation of the TAU breeding records; which corresponds to around G2:F3). The lines that are not genotyped have typically died out early in the inbreeding process (Figure 29). The genotype mosaics report CC founder status at 5,788 locations.

A number of additional lines were subsequently genotyped, which provided a larger dataset for analysis during preparation of this thesis. The additional genotypes increase the number of genotyped lines to 256, including IL-170 and IL-2513, which are labeled IL-170A and IL-2513A in the original genotypes and are therefore not correctly matched to our data. This genotype set reports CC founder status at 11,339 locations.

As the genome cache contains no information of the generation of inbreeding at which the CC lines were genotyped, I estimate the amount of homozygosity from the haplotype

probabilities of each of the CC lines in the TAU dataset (Figure 30). The genotypes are a mean of 0.655 homozygous (equivalent to 5 generations of inbreeding; range 0.361–0.938). Thus half the genotyped lines are at least 5 generations inbred, and most lines seem to have been genotyped at less than 10 generations inbred. 32 of the genotypes were done on MDA, 186 on MUGA and 38 on MegaMUGA.

When only the data from lines with available genotypes are considered, some interesting differences can be observed. Total litter size is greater in the ungenotyped lines, perhaps reflecting their lower amount of inbreeding (Figure 31a). However, the total number of weans seems higher in the genotyped lines, perhaps indicating a reason for the noncontinuation of the ungenotyped lines (Figure 31b). In general though, the identified trends hold when only the records with associated genotypes (which in any case comprise the majority of the full dataset) are considered. To facilitate analysis, I have added information on genotype availability to the data file.

Table 7: **CC lines and available genotypes**

Source	Lines	Genotypes	No genotypes
KEN-OLD	189	59	130
TAU-OLD	91	60	31
TAU-NEW	270	163	107
Other	65	65	0

Unrecorded cousins

The initial version of the cache contained genotype data from 288 inbred lines. Surprisingly, no breeding records existed for 65 of these lines (all numbered IL-5XX, IL-6XX, IL-7XX and IL-8XX [plus IL-170A and IL-2513A]). After extensive investigation, it seems that

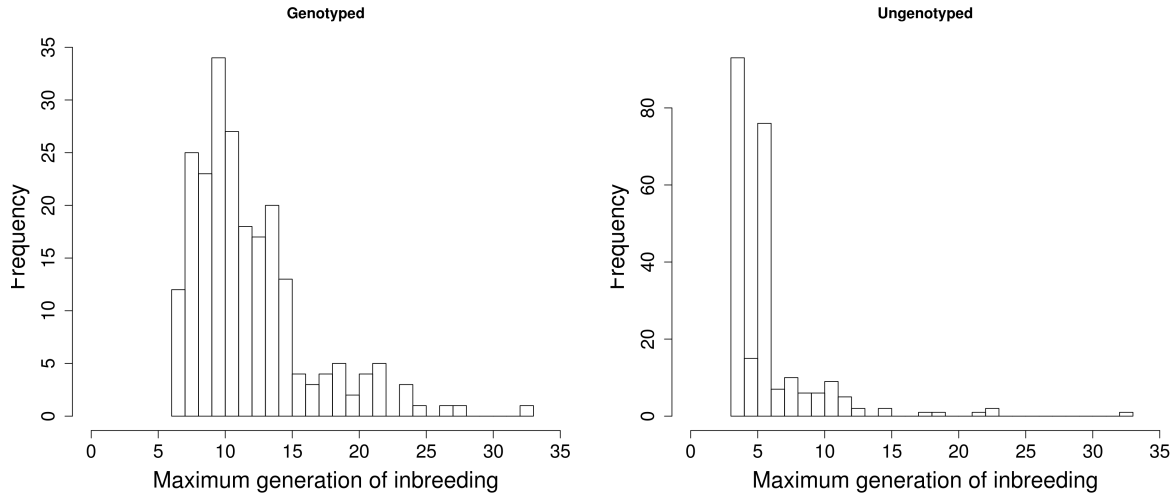


Figure 29: Maximum generation of inbreeding achieved by genotyped versus ungenotyped CC lines. All genotyped CC lines reached at least generation 6 of inbreeding. By contrast, very few of the ungenotyped lines survived past generation 6. The majority of litters in the TAU breeding records are therefore associated with genotypes.

these lines are also cousin lines (of IL-XX, IL-1XX, IL-2XX and IL-3XX, respectively), established at relatively late generations (earliest recorded is G10, although they in fact seem to have begun before this point). They are partially noted in the breeding records using a complex notation comprising letters and numbers; unfortunately the records do not contain sufficient information to make these CC lines useful.

From what I have established, ‘A’ denotes the original breeding line, and ‘B’ denotes the newly numbered cousin line. Approximately 60% of these lines for which I have genotypes, however, have no obvious breeding records (e.g. IL-511). Any notation indicating cousin status appears late (generation 9–18 of inbreeding), and it is not clear that this is when cousining occurred. In many cases, only one subline (A or B) continues past the first noted occurrence. Data from a partial list of available tail clips suggest that the cousining process (i.e. division into A and B sublines) occurs much earlier than this time (e.g. by G7 for IL-72, versus G16 in the breeding records). I therefore have no accurate breeding records for the cousin lines (IL-500+) and no indication of which of the recorded litters should actually be regarded as these cousins.

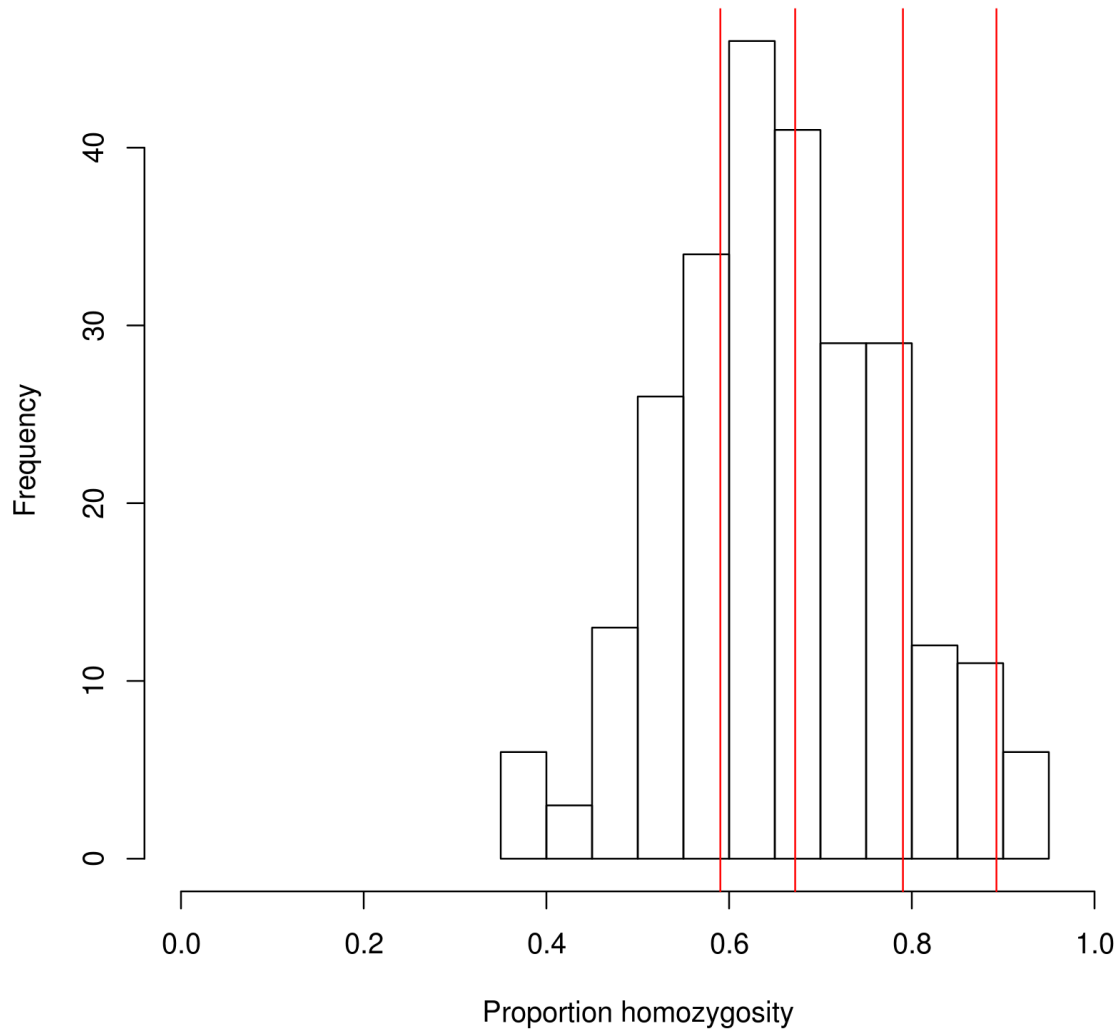


Figure 30: Proportion of homozygosity among the genotyped CC lines. Homozygosity is estimated from the hapotype probabilities for each CC line. The lines are a mean of 65.5% homozygous at the time of genotyping. Some lines are almost completely homozygous across all loci, whereas others have low levels of homozygosity. The vertical red lines denote the expected proportions of homozygosity at 4, 5, 7 and 10 generations of inbreeding, from left to right.

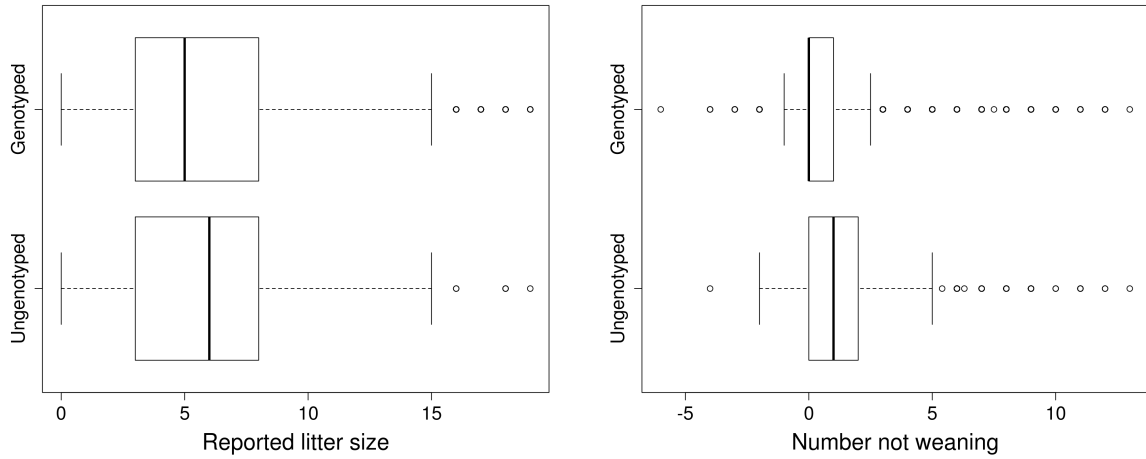


Figure 31: Properties of genotyped versus ungenotyped CC lines. The left panel (a) shows that litter size is in fact higher in the ungenotyped lines, probably because they represent the earliest generations of inbreeding. The right panel (b) shows that the ungenotyped lines have a greater number of offspring not weaning.

All in all, the lack of a clear breeding history for the lines and cousins initiated in Kenya, coupled with their uncertain generation labelling and missing data during the move from Kenya to TAU call into question the reliability of these data. I investigate further during the mapping process, but one tempting option is to discard these lines from the analysis and rely solely on the lines initiated at TAU, for which the breeding records seem more complete (if not without their own quirks and errors).

Chapter 4 | Genetic analysis of reproductive traits

Litter size is a quantitative character of some considerable complexity, and the title of this paper should more properly have been “Some aspects of the genetics of litter size.”

—Douglas S Falconer (1960)⁸²

Genetic analysis of complex traits

In Chapter 3 I explore the TAU CC breeding records and present exploratory analysis of the breeding data, which identifies reproductive traits suitable for genetic mapping as well as significant covariates. Here I analyse of the heritability of reproductive traits in the TAU CC mice, and map genetic loci associated with litter size and time between litters (Table 8). Both these traits show notable variation among the eight CC founders and among the incipient CC lines.

Table 8: **Reproductive traits for QTL mapping**

Trait	<i>n</i>	lines	mean	median	SD
Litter size	8,374	190	6.29	6	3.25
Time between litters	8,289	170	42	33	30.1

: Data are from only the CC lines initiated at TAU, which were used for most of the heritability estimates and QTL mapping in this Chapter.

The data used in this analysis are described in detail in Chapter 3. The reproductive trait data are extracted from the breeding records of the incipient (pre-CC) lines initiated at TAU, and the genotypes represent multiple generations of inbreeding (mostly less than 10 generations inbred).

Heritability

In Chapter 3 I highlight the extent of variation in litter size and other reproductive traits in the CC population, and demonstrate that CC inbred line is a significant linear predictor of many of these traits, indicating a line-specific (i.e. genetic) component. Here I provide a more-formal estimation of the heritability of such traits.

Heritability is a summary measure of the proportion of the variation in a trait that is attributable to variation in genetic factors.¹⁷⁹ This parameter provides a measure of the importance of genetic factors in explaining the differences between individuals. Several methods exist for estimating heritability, which typically involve estimating the genetic and environmental components of variance based on sample characteristics.

Previous work has observed that litter size is a heritable trait within CC lines,^{108,124} and litter size and other reproductive traits are heritable in other mouse and mammal populations.

Broad-sense heritability

I estimate broad-sense heritability (H^2) of the litter size trait in each generation of the TAU CC population from one-way analysis of variance (ANOVA). I use a method previously applied to other traits in the CC TAU lines.^{114,145,146} This method determines the fixed effects of CC line on litter size by least squares in a standard normal linear model.

It estimates H^2 as the proportion of the phenotypic variation that is attributable to genetic factors:

$$H^2 = \frac{V_G}{V_G + V_E}$$

Where V_E is the environmental variance within CC lines (estimated as the mean squares

within lines, taken from the ANOVA table), and V_G is the genetic variance among CC lines. V_G is also estimated from the ANOVA table as:

$$V_G = \frac{MS_b - V_E}{n}$$

MS_b is the mean squares between lines, taken from the ANOVA table, and n is the mean number of mice per line. As the CC mice are assumed to be inbred and therefore homozygous at all loci, this estimate includes additive (V_A) and epistatic (V_I), but not dominance (V_D) effects.

The H^2 estimates of litter size for each generation of the CC lines initiated at TAU are shown in Table 9. In general, these estimates are in the range 0.15–0.30, compatible with estimates for other traits in the TAU population. 95% confidence intervals indicate that the true heritability is nonzero. The implication is that 15–30% of the observed variation in litter size in each generation is attributable to genetic factors. H^2 estimates for the lines initiated in Kenya are also within this range (not shown).

I also calculate H^2 of litter size for each generation from the variance components of a linear mixed effect model in which litter size is the response to the random effect of CC line:

$$y_{iL} = \mu + u_L + e_{iL}$$

y_{iL} is the litter size of litter i in CC line L , μ is the overall mean, u_L is a random effect that models the differences between lines, and e_{iL} is the residual error. The estimate is in each case near identical to the values calculated by ANOVA (Table 9). This mixed-model approach estimates V_G by maximum likelihood rather than by least squares. The estimates from the two methods will converge if the line effects do resemble a sample of random deviates from the normal distribution, as assumed by the mixed model.

Across all generations of inbreeding, a mixed model that includes generation of inbreeding as a fixed effect estimates broad-sense heritability of litter size in the lines initiated at TAU as 0.104, in the lines initiated in Kenya as 0.2 and in the combined population as 0.115. These values are consistent with previous estimates for the heritability of litter size, which is generally reported to be low.

Table 9: **Heritability of litter size among TAU and UNC CC lines**

Generation	H^2	95% CI	H^2_{mixed}	CV_G	h^2 (UNC)	CV_A (UNC)	h^2
4	0.149	0.107, 0.198	0.152	0.139	0.02 (G2:F1)	0.0486	0.0325
5	0.281	0.228, 0.34	0.288	0.226	0.14 (G2:F2)	0.148	0.202
6	0.169	0.118, 0.228	0.164	0.186	0.10 (G2:F3)	0.143	0.0338
7	0.25	0.192, 0.315	0.239	0.251	0.31 (G2:F4)	0.268	0.275
8	0.273	0.211, 0.343	0.281	0.275	0.26 (G2:F5)	0.252	0.229
9	0.265	0.194, 0.345	0.278	0.274	0.22 (G2: F6)	0.246	0.137
10	0.313	0.232, 0.402	0.309	0.284	0.22 (G2:F7)	0.269	0.273
11	0.356	0.266, 0.456	0.369	0.337	0.07 (G2:F8)	0.15	0.148
12	0.193	0.109, 0.298	0.2	0.236	NA	NA	0.146
13	0.172	0.0796, 0.291	0.176	0.243	NA	NA	0.0462
14	0.184	0.0712, 0.329	0.196	0.257	NA	NA	0.288
15	0.223	0.0813, 0.405	0.229	0.269	NA	NA	0.442

Genetic coefficient of variation

Also following previous methods,¹⁴⁵ I calculate the genetic coefficient of variation (CV_G) for each generation of the TAU CC lines (Table 9). I use the following method:

$$\frac{SD_G}{Mean}$$

which is equivalent to:

$$\frac{V_G^{0.5}}{Mean}$$

where:

Mean = mean trait value across all CC lines

Theoretically, this parameter provides an estimate of the absolute amount of genetic variation that is generated by the genetic factors responsible for the heritability estimate.^{180,181} Our moderate heritability estimates are compatible with low CV_G if total phenotypic variation is low, and high CV_G if total phenotypic variation is high. It is best thought of as the unit-free ratio of standard deviation to the mean (interpreted as the mean of the individual line means, although in this case using the population mean makes no notable difference to the parameter estimates). Iraqi and colleagues suggest comparing CV_G values with a benchmark of 0.071 to assess how the genetic variation among CC lines compares to that of typical outcrossing populations.¹⁴⁵

The estimates of CV_G for each generation are in the range of 0.14–0.33, again comparable with those for other traits in the TAU population. Such values theoretically show that the absolute magnitude of genetic variation surpasses that seen in outcrossing populations and that large-effect QTLs might segregate in the CC lines.¹⁴⁵ Iraqi and colleagues attribute this effect to the inclusion of the three wild-derived strains in the CC breeding scheme. In fact, traits closely related to fitness (which I assume include reproductive phenotypes) are expected to have relatively high CV_G values (in contrast to their often low reported heritability).¹⁸⁰ Our estimated CV_G values suggest that the absolute magnitude of this genetic variation for litter size is of reasonable size, and thus this trait is amenable to mapping.

Narrow-sense heritability

Narrow-sense heritability (h^2) is an estimate of the proportion of the phenotypic variation that is attributable to additive (V_A), but not epistatic (V_I) or dominance (V_D) effects. h^2 of litter size and other reproductive traits has previously been estimated in the UNC arm of the CC.¹²⁴ After accounting for differences in labeling of generation (i.e. G2 in the UNC data refers to the offspring produced by G2 parents, whereas the TAU data records these offspring as G3 litters) I present this parameter in Table 9 along with additive coefficient of variation (CV_A) parameters calculated from the published narrow-sense heritability.

I also calculate narrow-sense heritability from the TAU CC lines by parent–offspring regression. In theory the difference between this parameter and our calculated broad-sense heritability would reveal V_I (although in practice the estimates are likely to be both similar and affected by noise). I use the Subcages field of our data (which can be used to link litters to their parent litter) to identify the corresponding parental litter for all offspring litters. I treat this parental litter size as the midparent value (parental mean) and estimate h^2 by the slope of the regression of offspring litter size on this value (Table 9). I report the regression of Generation 4 offspring on Generation 3 parents as a property of Generation 3 (technically G2:F1, with parents representing G2), although this assumption is not necessarily correct.

The h^2 estimates show a similar range to our estimates of H^2 , although with greater variation; for some early and late generations the estimate of h^2 is essentially zero. Our estimates match those from the UNC arm of the CC reasonably well.¹²⁴ However, on the basis of the per-litter structure of our dataset, and the fact that litters represent discrete events rather than parent or offspring traits, I urge caution in interpretation of this statistic. The broad-sense heritability calculations by ANOVA and mixed model are likely to be more reliable.

Estimates of broad sense heritability and coefficient of genetic variation for time between

litters are presented in Table 10. These parameter estimates are of approximately the same magnitude as are those for litter size, indicating that a portion of the phenotypic variation between lines is attributable to genetic variation and that this trait is also amenable to genetic mapping. By contrast, broad-sense heritability estimates for the difference between the number of male and female weans (Sexdiff) and the number of offspring not weaning (Notweans) are consistently <0.05 for each phenotype across generations indicating a lack of a mappable genetic component (not shown). These phenotypes might however make some sense re-expressed as ratios or proportions of total litter size, respectively.

Table 10: **Heritability of time between litters**

Generation	H^2	95% CI	H^2_{mixed}	CV_G
4	0.215	0.169, 0.269	0.11	0.234
5	0.16	0.115, 0.213	0.0582	0.249
6	0.0716	0.0286, 0.123	0.0531	0.164
7	0.191	0.136, 0.254	0.17	0.314
8	0.269	0.206, 0.34	0.282	0.322
9	0.25	0.178, 0.33	0.181	0.314
10	0.111	0.0435, 0.194	0.0971	0.227
11	0.271	0.184, 0.373	0.306	0.358
12	0.169	0.0881, 0.273	0.145	0.251
13	0.126	0.0382, 0.242	0.134	0.253
14	0.194	0.0794, 0.342	0.159	0.258
15	0.0795	-0.0343, 0.248	0.00427	0.167

Genetic mapping of reproductive traits

Mapping of quantitative trait loci (QTLs) is a common approach to investigating the genetics of complex traits. Mapping is the process of identifying the physical regions of the genome that are associated with trait variation.

A QTL is a section of the genome that contains genetic variation that is associated with

variation in phenotype. A number of methods have been used for identifying QTLs in different populations. At the time of writing, 5,050 QTLs for a wide range of traits are listed in the Mouse Genome Informatics database (<http://www.informatics.jax.org/>), with many more identified in other species including humans. QTL mapping can indicate the genetic architecture of a particular trait and the number of causative genes. QTLs can subsequently be investigated further to identify causal variants and genes.

QTL mapping strategies have traditionally been developed for diallelic (i.e. two-parent) populations. The development of multiparental populations such as the CC has necessitated alternative mapping approaches, as single marker and interval mapping approaches can fail to identify QTLs if the marker do not segregate between founders in the same manner as the causative variants. In general, the recombinant inbred lines produced from such populations are represented as mosaics of their founders and traits are mapped using these mosaics.¹⁵⁰ Such approaches are well established in multiparental populations and have been shown to be effective for mapping QTLs.¹⁸²

One of the stated aims of developing the CC was to create a resource for QTL mapping, with good power and resolution. As discussed previously, this goal has seen some success with identification of QTLs associated with a variety of traits using the incipient lines of the cross. Trait mapping in the CC has typically used regression of the trait of interest on the haplotype founder doses generated from the haplotype mosaics to calculate a LOD score, with significance assessed by permutation. I use a variation on this approach to map reproductive traits in the TAU CC population.

Linear mixed effect models

I map QTLs associated with reproductive traits in the pre-CC lines, using a linear mixed modelling (mixed model) approach with the lme4 R package.^{183–185} Mixed models comprise both fixed and random effects, and are suitable for analysis in settings where

repeated measures are made on related statistical units (i.e. the same individual, or individuals with different degrees of relatedness) and when the repeated measures are made on individuals who are representative of a larger population (i.e. when variation between subjects is more relevant than are subject-specific effects). In essence, the mixed model approaches used for genetic mapping act to decorrelate the phenotypes from related individuals (including repeated measures of a given individual or within a CC line), which can then be analysed as if independent.

As with the exploratory analyses I begin by examining the total recorded litter size phenotype. Given our observation of the differences between the CC lines initiated in Kenya and those initiated at TAU, I initially investigate only the latter cohort. I fit a model in which generation of inbreeding is included as a fixed effect, and CC inbred line is included as a random effect:

$$y_{iL} = \mu + \beta x_{iL} + u_L + e_{iL}$$

y_{iL} is the litter size of litter i in CC line L , μ is the overall mean, x_{iL} is the generation number for the combination i, L , β is the regression coefficient relating litter size to generation, u_L is a random effect that models the differences between lines, and e_{iL} is the residual error. u_L is sampled from a normal distribution with between-line variance estimated from the data, and the residuals are sampled from a normal distribution with variance also estimated from the data.

Within a given generation there is variability in litter size, but also considerable variability between inbred lines. This model fits a random intercept with a fixed mean (i.e. it estimates the mean and standard deviation of the intercepts, with the slope of the fixed effect generation of inbreeding being constant among funnels). I include generation of inbreeding as a fixed-effect covariate because of the strong effect of inbreeding depression on litter size. I expect this effect to result from accumulation of homozygosity during the inbreeding process, and to not be easily mapped using fixed inbred genotypes that assume

homozygosity. Thus I include generation as a fixed effect and aim to map reproductive loci independent of the inbreeding depression process. Including generation of inbreeding as a fixed effect significantly improves the model fit over the random effect only (i.e. versus the model $y_{iL} = \mu + u_L + e_{iL}$; $\chi^2=2,167$, $P=0e+00$).

Profile zeta plots show balanced nonzero confidence intervals for the model terms (Figure 32), and profile pairs plots shows that the conditional estimates of most parameters are constant; however, the conditional effect of Generation (given the intercept) depends on the intercept and vice versa (Figure 33). A plot of the best linear predictors of the random effects against standard normal quantiles show that many prediction intervals for the random effects do not overlap zero (not shown). The distribution of random effects seems reasonably balanced, potentially with a few extreme values at the high end of the distribution.

QTL mapping

I evaluate genetic effects on reproductive phenotypes (litter size in this initial case) by including a genetic effect for each locus in turn as a fixed effect in the mixed model, and comparing the model fit with that of a reduced model by likelihood ratio test.

This is the comparison of our previous model:

$$y_{iL} = \mu + \beta x_{iL} + u_L + e_{iL}$$

versus the genetic model:

$$y_{iL} = \mu + a_L + \beta x_{iL} + u_L + e_{iL}$$

Where y_{iL} is the littersize of litter i in CC line L , μ is the overall mean, x_{iL} is the generation number for the combination i, L , β is the regression coefficient relating litter

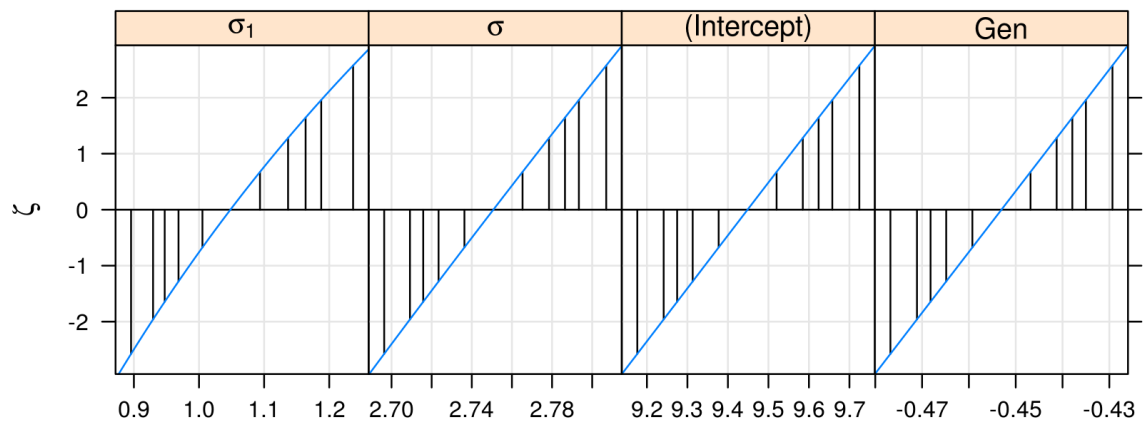


Figure 32: Profile zeta plot for the fitted model fm02. The confidence intervals for the parameter estimates are nonzero, and are well balanced (shown by the reasonably linear nature of the blue lines).

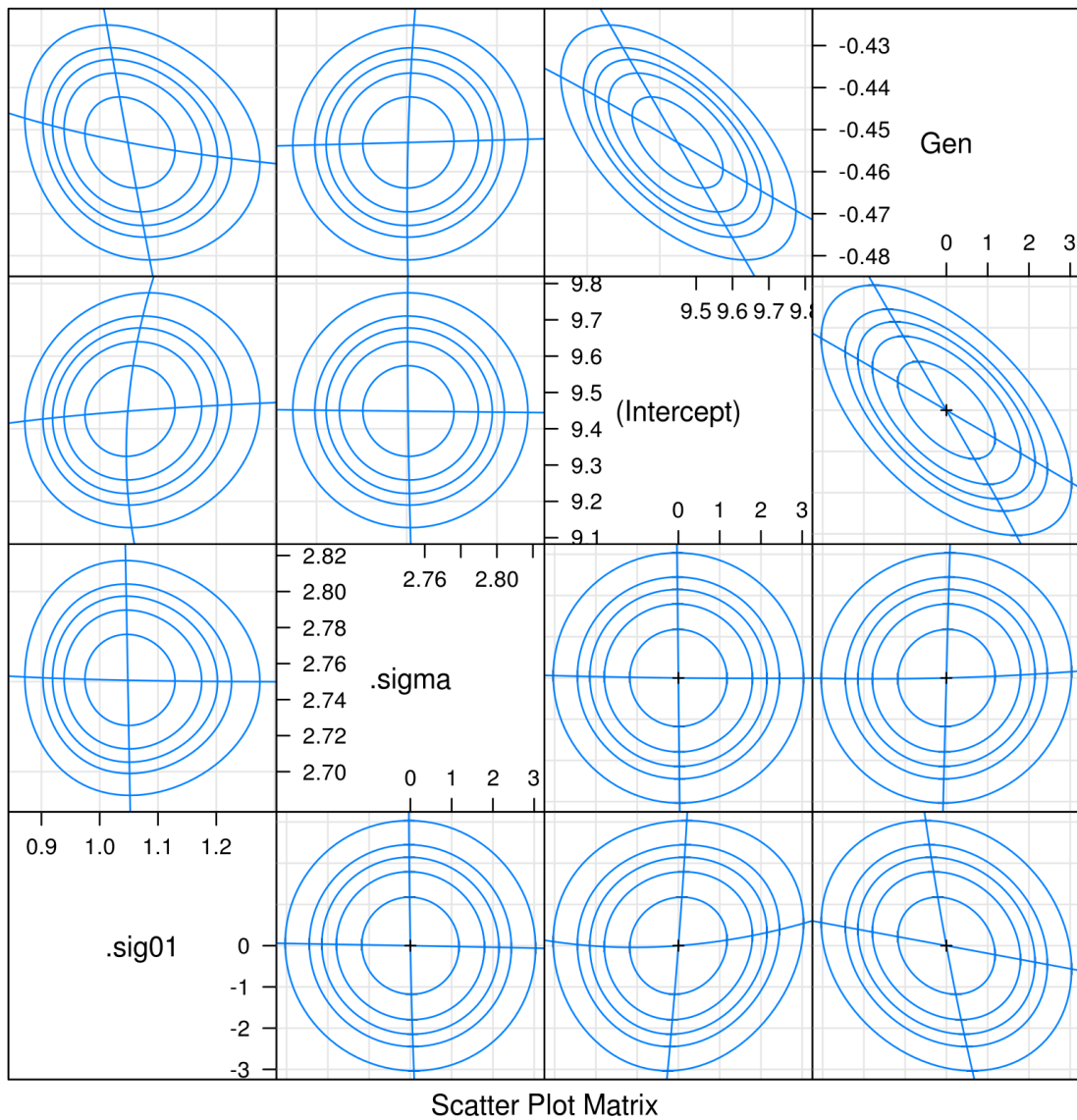


Figure 33: Profile pairs plot for the fitted model fm02. The conditional estimates of most parameters are constant, and do not depend on the other estimates. However, the conditional effect of Generation depends on the estimate of the intercept, and vice versa.

size to generation, u_L is a random effect that models the differences between lines, e_{iL} is the residual error, and a_L is the genetic effect at the locus assessed as haplotype doses.

This genetic model has an implicit intercept for the locus, plus a fixed effect of generation of inbreeding. Funnel is fitted as a random effect to account for the different levels of relatedness between CC lines, the fact that the CC lines represent a sample of all possible lines, and the repeated measures taken from each line. The comparison tests whether the inclusion of a genetic effect at a locus (in this case represented by 7 fixed effects comprising the uncorrelated principal components of a matrix of 8 correlated haplotype doses) significantly improves the model fit at the cost of 7 degrees of freedom.

In the terminology of the lme4 R package this method is executed as:

```
fm02 <- lmer( Litter ~ Gen + ( 1 | Funnel ))
fm02a <- lmer( Litter ~ d + Gen + ( 1 | Funnel ))
anova( fm02, fm02a )
```

Here lmer() is the R function that fits the specified mixed model, ‘Litter’ denotes litter size, ‘Gen’ denotes generation number, ‘Funnel’ denotes CC line, ‘d’ denotes the genetic effect at the locus, and anova() is the R function used to compare the model fits.

I use permutation tests to calculate genome-wide thresholds of significance for each scan. Such tests are described in detail elsewhere.¹⁸⁶ Briefly, the CC line identifiers are permuted and the genomewide maximum association (in this case reported as $-\log_{10}P$) across all loci is recorded for each permutation, to generate the empirical distribution of test statistics under the null condition of no QTL. Permutation is done by reallocating the observed litter sizes among CC lines, while preserving the variation within lines, to simulate a null hypothesis of no effect of genotype on phenotype at any locus.

Following the procedures used in previous CC mapping studies I test QTLs by running 500–1,000 permutations. I modify scripts used in previous studies to optimize computational

efficiency, and to correct an error in genotype assignment (in which genotypes from outside the original dataset are included in the permutations). I calculate 95% genomewide thresholds of significance for each model, and 50% thresholds for suggestive loci. I also investigate pointwise (or comparisonwise) thresholds when appropriate.

QTL mapping for litter size

QTL mapping using a model with a fixed effect of generation and a random effect of CC line identifies two QTLs on chromosomes 1 and 5 that are genomewide significant at the 95% level by permutation testing (Figure 35a). A number of other QTLs (on chromosomes 2, 4, 6, 7, 11, 17 and 18) are suggestive at the 50% level.

The model treats all litters within a given funnel as independent events. In fact among the TAU lines two or more distinct matings (Subcages) are initiated at each generation, with litters then representing repeated measures over time. I generate a new variable (Mating) from a combination of Gen, Funnel and Subcages, to represent each unique mating pair. Mating is nested within the combination of Funnel and Generation. The inclusion of the variable ‘Mating’ (denoting litters produced by a specific mating pair within a CC line) as a random effect variable makes no material difference to these results, with additional suggestive loci on chromosomes 9 and X (Figure 35b).

I then fit a model in which the slope of the generation effect is allowed to vary by CC line, which potentially better reflects the fact that the decline in litter size associated with increasing generation of inbreeding is not uniform across CC lines (Figure X). Importantly, the slope (and to a lesser extent intercept) is generally a reflection of how well each line has bred, and how many generations it reaches. I therefore refit the model allowing the slope of the fit to vary according to funnel. I see a strong negative correlation (-0.87) between the random effects for intercept and slope, suggesting that litter sizes in the first generations of inbreeding give information on litter size in subsequent generations.

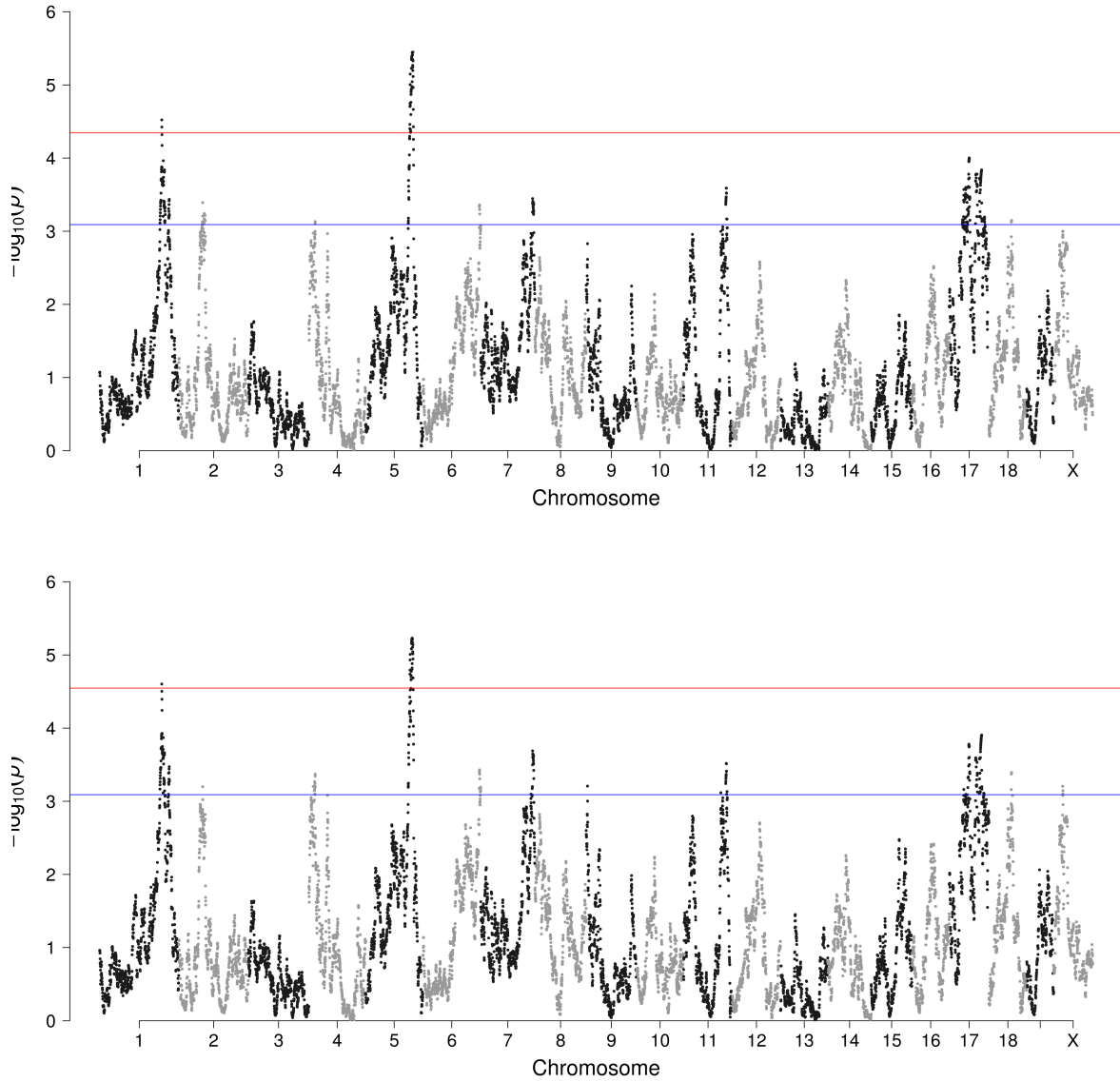


Figure 34: QTL mapping for litter size. The upper panel (a) shows two QTLs significant at the genomewide 95% level (red line), and several suggestive QTLs (blue line). The fixed effects of generation and genetics are uncorrelated based on a sample of loci. lme4 model: $\text{Litter} \sim \text{Gen} + (1 \mid \text{Funnel})$. The lower panel (b) shows the inclusion of a random effect of Mating. The mapping results are not materially different from the simpler model used in (a). lme4 model: $\text{Litter} \sim \text{Gen} + (1 \mid \text{Funnel}) + (1 \mid \text{Mating})$.



Figure 35: Litter size with increasing generation of inbreeding among the genotyped CC lines. The slope of the linear regression for each inbred line varies, and generally reflects how well the line has bred and the maximum generation of inbreeding reached. Each panel shows litter sizes for one inbred line. Intercepts are reasonably constant among lines, but some between-line variation is seen.

Mapping using this model identifies significant QTLs on chromosomes 5 and 17, with suggestive QTLs on chromosomes 1 (two loci), 6, 7, 8, 11, 15, 18 and X (Figure 36a). The inclusion of the Mating variable in the model does not materially affect the mapping results (Figure 36b). Importantly, allowing the slope of the fitted model to vary is computationally intensive, but does not materially affect mapping results compared with the simpler models used to produce Figure 35. I thus omit the variable-slope and Mating variable from subsequent analyses.

Together, these scans show overlapping significant peaks of interest on chromosomes 1, 5 and 17, with magnitudes similar to those reported as significant in prior CC mapping studies. Interestingly, the model with a variable slope (which well explains the phenotype being examined) produces a better resolution of the chromosome 17 peak than does either fixed slope model. Variable slope models are, however, notably more computationally intensive than are fixed slope models.

Modeling generation effects

I also investigate other features of the data: initially by truncating the dataset in various ways. Evidence indicates that lines that reach late generations of inbreeding are good breeders unrepresentative of the broader population; truncating the data limits the potential for these outliers to unduly influence the fit of the model. Removing later generations from the data removes the tail of high-littering lines seen in Figure 9 and retains the peaks on chromosomes 1 and 5, but increases the magnitude of suggestive peaks on chromosomes 6 and 7. I illustrate such a scan with the dataset censored after Generation 9 (Figure 37a).

By contrast, truncating the early generations of the dataset (in this case G3–G7) results in the progressive loss of the chromosome 5 peak (Figure 37b). A previous attempt to map litter size QTLs in the UNC arm of the incipient CC used only animals from

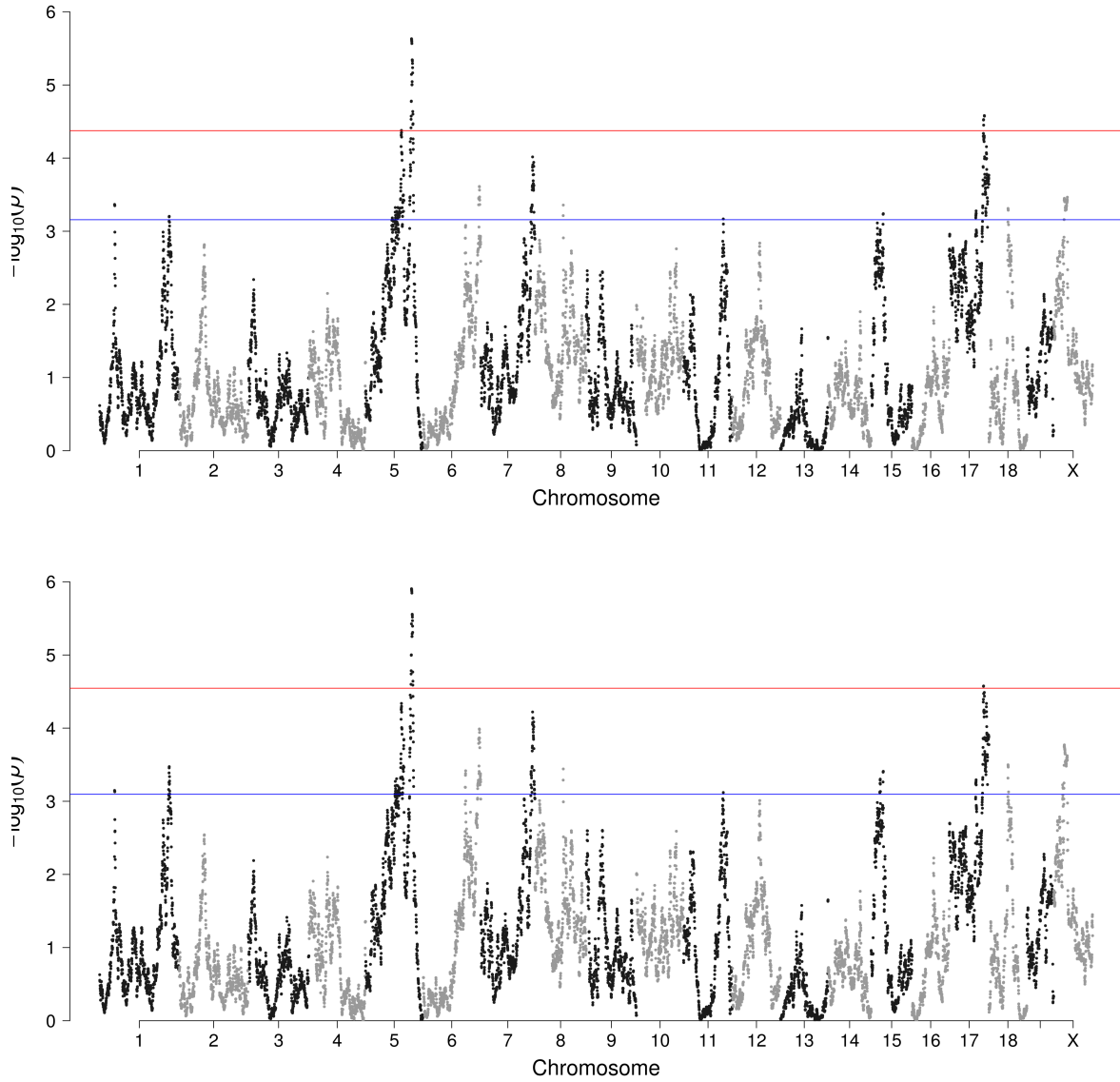


Figure 36: QTL mapping for litter size in which the slope of the generation effect is allowed to vary. The upper panel (a) identifies two QTLs significant at the genomewide 95% level (red line), and several suggestive QTLs (blue line). lme4 model: Litter $\sim$ Gen + (Gen | Funnel). The lower panel (b) shows the inclusion of a random effect of Mating. The mapping results are not materially different from the simpler model used in (a). lme4 model: Litter $\sim$ Gen + (Gen | Funnel) + (1 | Mating).

G2:F5 onwards (equivalent to Generation 8 in our data).¹²⁴ This prior study found only a suggestive QTL on chromosome 6 at approximately the same location as does our mapping. The lack of early generations in the UNC study might explain its limited success in identifying litter size QTLs. Notably, the size and balance of our dataset decreases with successive generations, and especially past Generation 10.

I also explore the effect of fitting generation of inbreeding as a nonlinear covariate. During exploratory scans I accidentally fitted Generation as a factor, and the resulting parameter estimates suggested its effect on litter size might not be linear. Fitting Generation as a factor again identifies peaks on chromosomes 1 and 5, as well as suggestive loci similar to our earlier scans (Figure 38a). Fitting generation as a nonlinear effect whereby the estimated heterozygosity is used rather than the generation number returns the same pattern of peaks (Figure 38b).

Effects of subpopulation

Exploratory data analysis demonstrates notable differences in litter size between the CC lines initiated in Kenya and those initiated at TAU. So far, I report genetic mapping from only the lines initiated at TAU. Being able to reliably include the Kenyan data would increase the size of our data set for mapping, possibly increasing power to detect QTLs. In fact, mapping litter size in just those lines initiated in Kenya produces no notable peaks (not shown). I see this result as reinforcing suspicions that problems of generation labeling and genotyping (including unspecified cousins, with different genotypes but undifferentiated breeding records) exist within these lines.

I map litter size using all data without accounting for Source, which identifies a novel pattern of peaks, including significant peaks on chromosomes 4 and 17 (Figure 39a). Interestingly a similar pattern emerges when the TAU data are truncated at early generations, and particularly when these data are mapped without including Generation

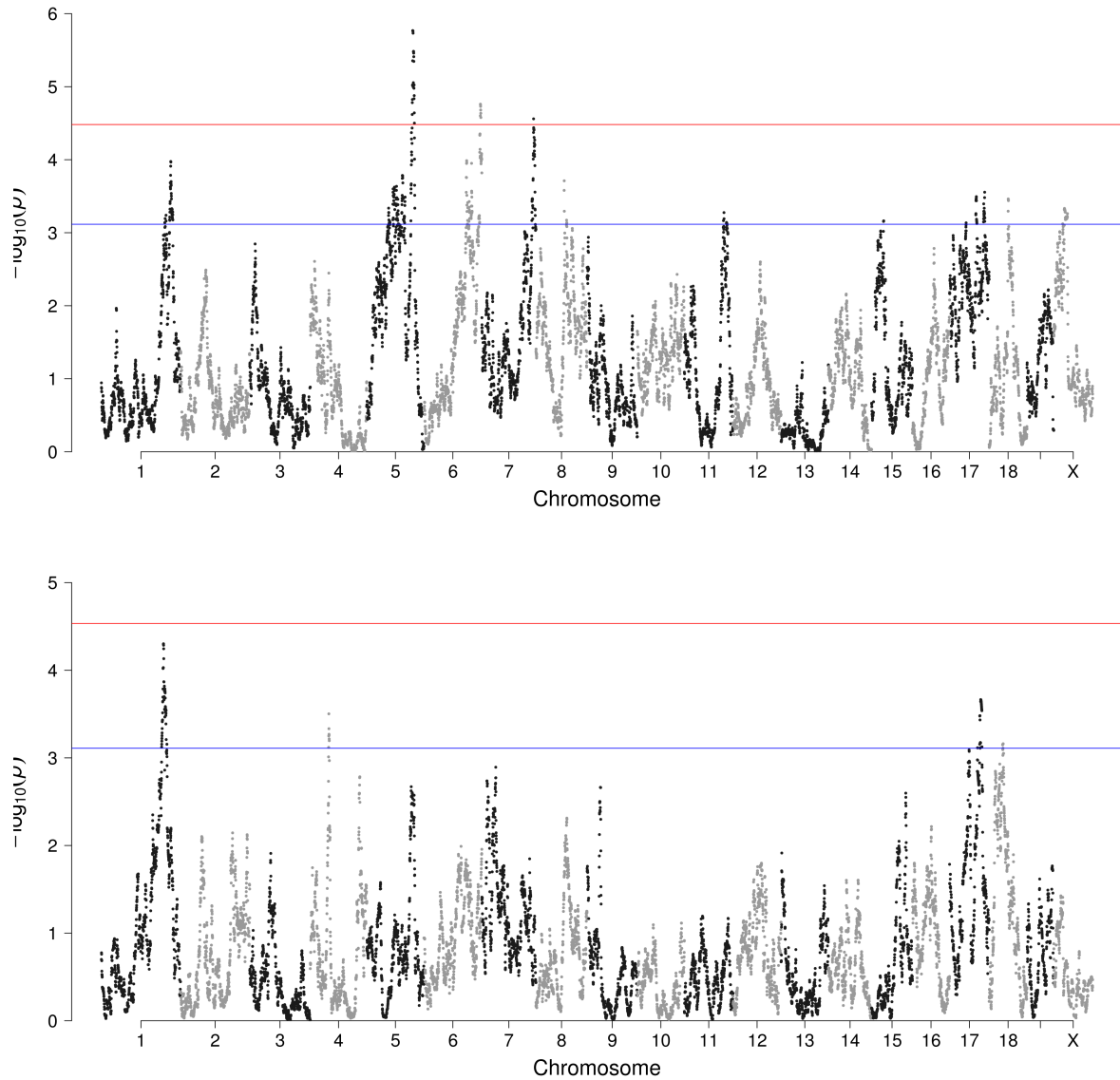


Figure 37: QTL mapping for litter size in early and late generations. The upper panel (a) shows the results of QTL mapping in the early generations of the CC (3–9) and identifies three QTLs significant at the genomewide 95% level (red line), and several suggestive QTLs (blue line). lme4 model: $\text{Litter} \sim \text{Gen} + (1 \mid \text{Funnel})$. By contrast the lower panel (b) shows QTL mapping in the late generations of the CC (>8), which results in the progressive loss of the peak on chromosome 5, and the disappearance of all significant QTLs. lme4 model: $\text{Litter} \sim \text{Gen} + (1 \mid \text{Funnel})$.

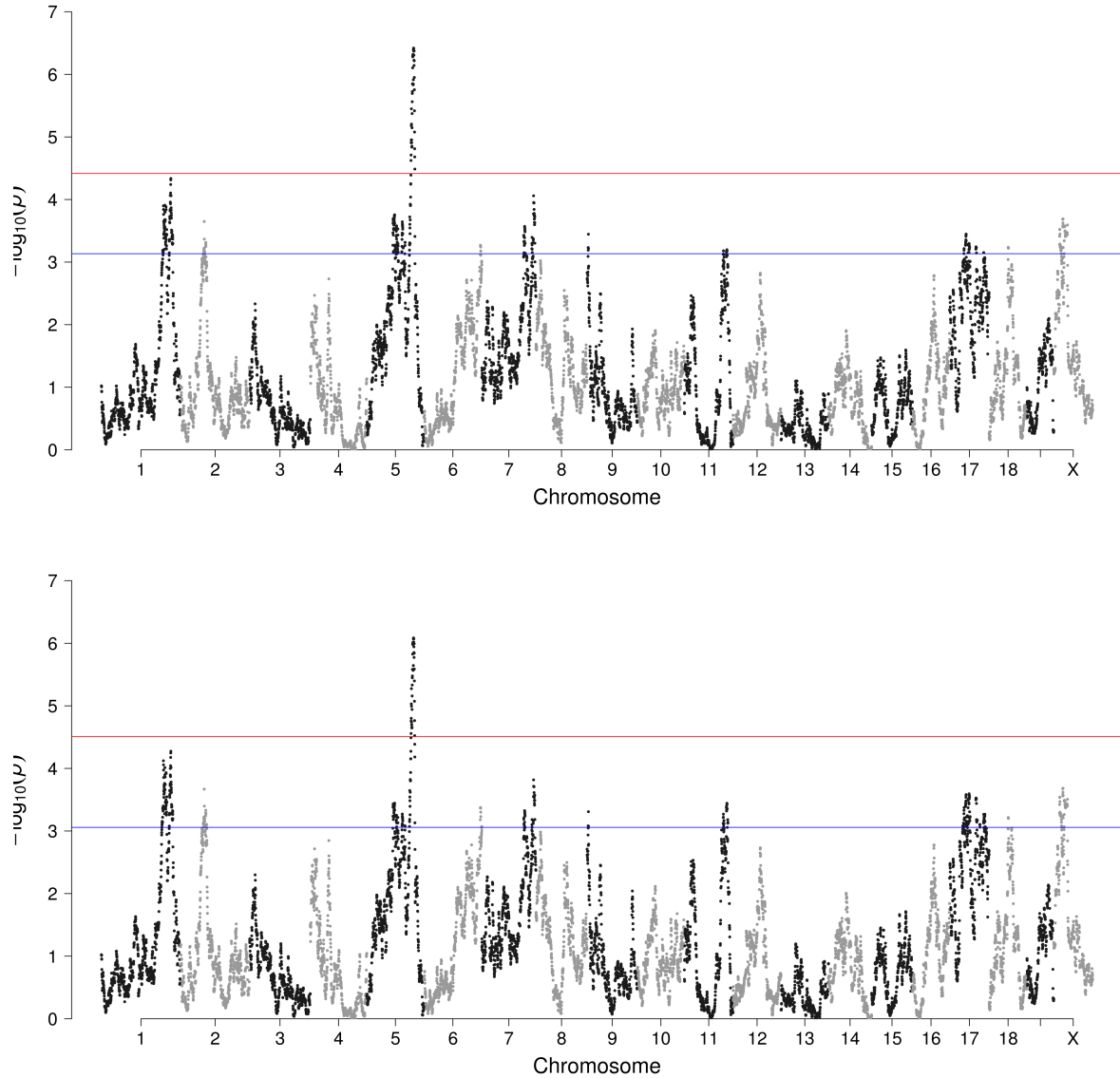


Figure 38: QTL mapping for litter size using different generation effects. In the upper panel (a) QTL mapping for litter size treating generation of inbreeding as a factor identifies one QTL significant at the genomewide 95% level (red line), and several suggestive QTLs (blue line). lme4 model: $\text{Litter} \sim \text{as.factor}(\text{Gen}) + (1 \mid \text{Funnel})$. The lower panel (b) fits the generation effect as the estimated heterozygosity at each generation of inbreeding. lme4 model: $\text{Litter} \sim I(0.8^{(\text{Gen}-3)}) + (1 \mid \text{Funnel})$; ccdataTAU

as a covariate. This observation provides some evidence that the generation labeling is inconsistent in the whole population, and that mapping in the whole population is similar to mapping in later generations where the effect of Generation is expected to be low. It also suggests that mapping in late generations of the CC might be possible without the inclusion of a generation covariate.

By contrast, mapping using the residuals of a linear model that regresses litter size on the interaction of Generation with Source shows the re-emergence of a suggestive peak on chromosome 5, and a pattern of peaks similar to mapping using just the lines initiated at TAU (Figure 39b). I conclude that regressing out this interaction is successfully accounting for some, but not all of the variation between data sources. I suspect that some within-population variation also exists in the recorded generation numbers in the dataset.

QTL mapping for time between litters

I also investigate genetic associations with time between litters. I include first litter status in the mapping model, as exploratory analysis revealed that first litter times from mating to birth included a period of time before the parents were sexually mature. Mapping revealed one QTL on chromosome 12 significant at the 95% level, and one QTL on chromosome 3 significant at the 90% level (Figure 40a). Additional inclusion of litter size as an explanatory variable in the mapping model results in 95% genomewide significant QTLs on both chromosomes 3 and 12 (Figure 40b). Litter size is known to be associated with gestation time, with larger litters being born sooner.¹⁶³

QTL mapping for other traits

Heritability analyses suggest that discovering significant genetic associations by QTL mapping for other reproductive traits in the TAU CC population is unlikely. Nonetheless,

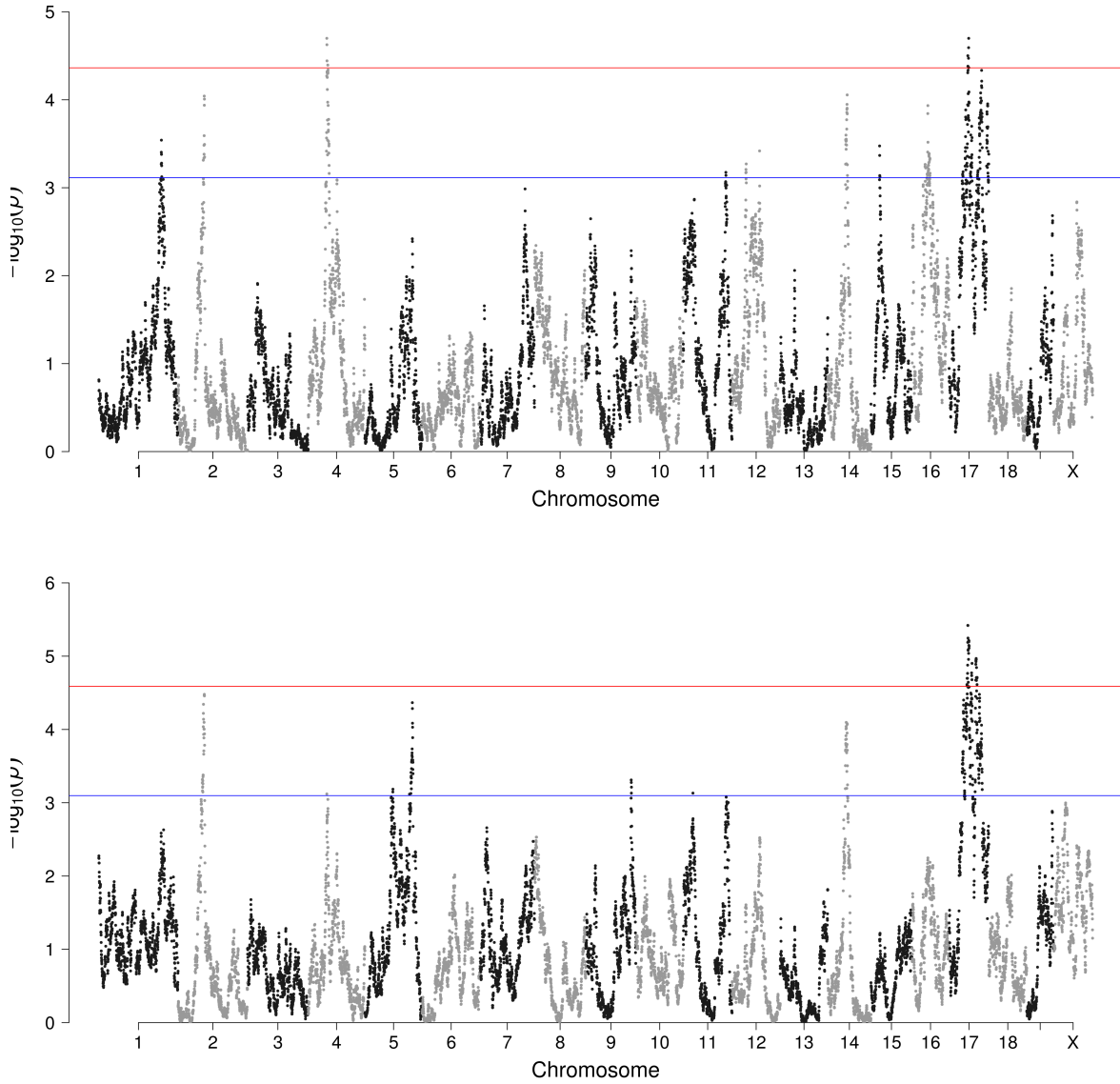


Figure 39: QTL mapping for litter size in the entire TAU CC population. The upper panel (a) shows two QTLs significant at the genomewide 95% level (red line), which are not identified in just the CC lines initiated and maintained at TAU. lme4 model: $\text{Litter} \sim \text{Gen} + (1 \mid \text{Funnel})$. The lower panel (b) shows mapping for litter size using the residuals of a linear model that regresses litter size on the interaction of Generation with Source. The scan shows the re-emergence of a suggestive peak on chromosome 5, and a pattern of peaks similar to mapping using just the lines initiated at TAU. lme4 model: $\text{Litter_res} \sim (1 \mid \text{Funnel})$.

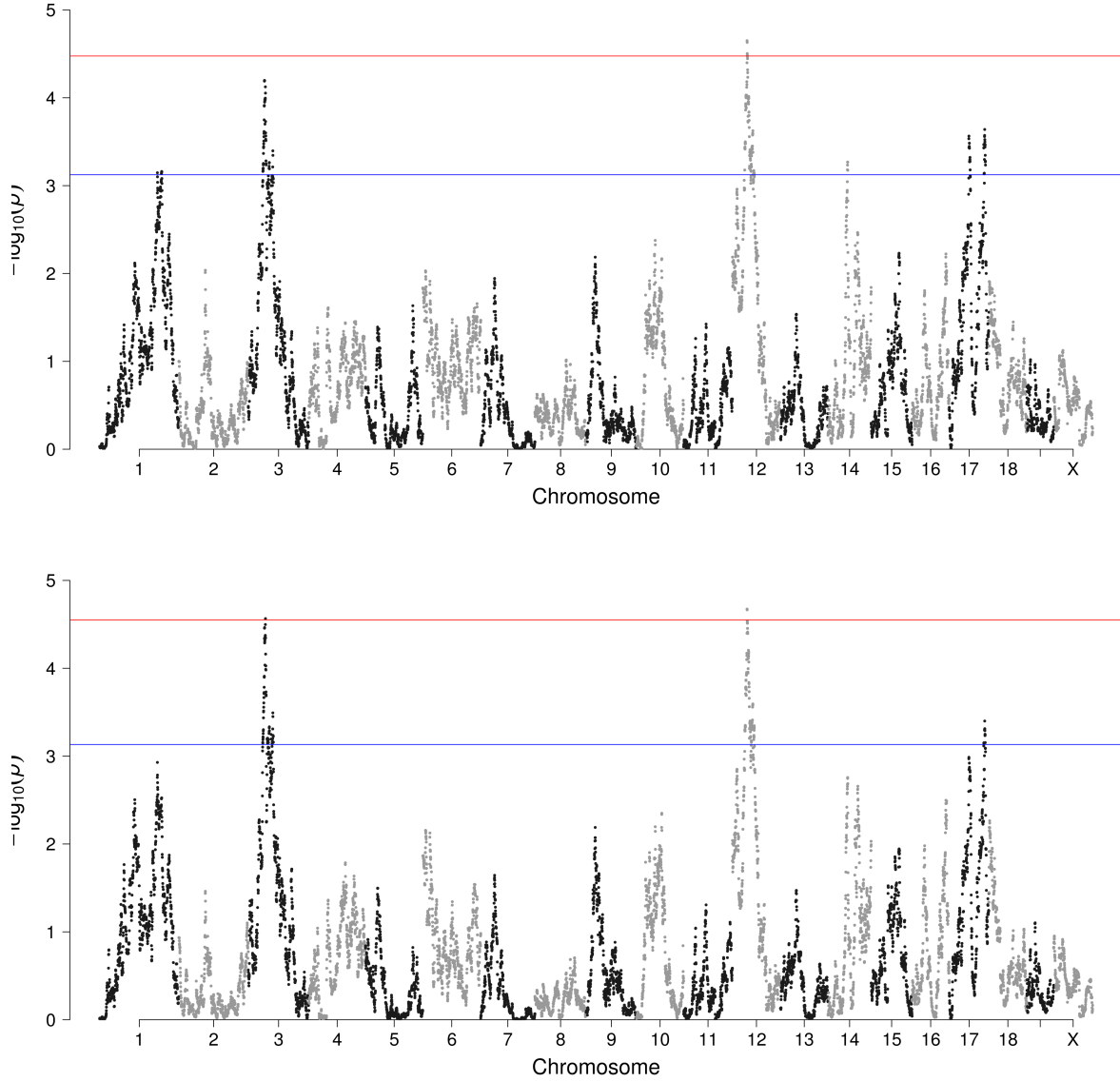


Figure 40: QTL mapping for time between litters. The upper panel (a) shows one significant QTL on chromosome 12. lme4 model: $Gdays_Adj \sim First_litter + Gen + (1 | Funnel)$. The lower panel (b) shows the results of mapping with the addition of litter size as an explanatory variable. Two significant QTLs are identified on chromosomes 3 and 12. lme4 model: $Gdays_Adj \sim Litter + First_litter + Gen + (1 | Funnel)$.

I scan traits for which there is an indication of possible line-specific differences, including number of offspring not weaning and difference between number of weaning males and females. I also test the total number of weans, in case any litter size QTLs emerge that were masked by the number of offspring not weaning; I see no material differences to the scan for reported litter size (not shown).

Scans of the difference between weaning males and females, and the proportions of male and female weans among the total weans reveal a plausible suggestive QTL on chromosome X (not shown). This male:female ratio phenotype might benefit from further examination in a generalised linear mixed model (GLMM) framework. Interestingly, a scan for number not weaning is suggestive for QTLs on chromosomes 6 and 17 indicating genetic associations with early-age survival (not shown). Similar to sex ratios, the number not weaning might benefit from analysis as a proportion within a GLMM framework.

Summary

The results of the various genome scans are summarized in Table 11 and Figure 41. Overall, I present evidence for significant QTLs associated with litter size on chromosomes 1, 5 and 17 in the CC TAU population, and a number of suggestive QTLs (including those in previously reported locations). Truncating the CC TAU data at Generation 10 reveals additional significant QTLs on chromosomes 6 and 7 at locations that are suggestive in the other analyses. The QTL on chromosome 5 is present across different scan types, and account for ~3% of total variation and 24% of genetic variation in litter size (Table 12). Other litter size QTLs each explain 1.4–2.6% of total variation.

Encouragingly, subsetting and otherwise re-expressing the dataset results in no material changes in the position of peaks, although it does result in some changes in magnitude. Mapping using the whole CC dataset identifies possible significant QTLs on chromosomes 4 and 17; however, mapping residuals after accounting for differences in the data sources

results in the disappearance of this peak on chromosome 4.

I see evidence for significant associations with litter size for QTLs on chromosome 17 across multiple mappings; however, in most scans much of this chromosome shows suggestive association with litter size and consequently resolution to dissect these effects is poor. I see a significant association of time between litters with QTLs on chromosomes 3 and 12, and suggestive associations for other reproductive phenotypes, including with chromosomes 17 and X, which have known roles in reproductive function. The QTLs for time between litters each explain $<1\%$ of total variation, but $>15\%$ of genetic variation. The identified QTLs are summarized in Table 12.

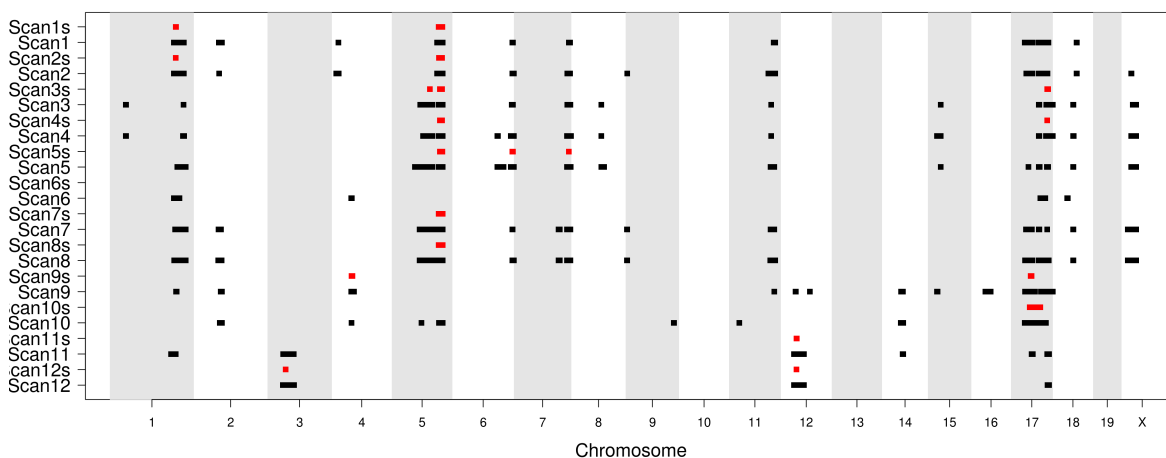


Figure 41: Summary of the various QTL scans. The QTL mapping scans described in the above section and Table 3 are illustrated in order. For each scan, the locations of 95% genomewide significant QTLs are marked in red. Suggestive QTLs (above 50% genomewide threshold) are shown in black. The litter size QTL on chromosome 5 is identified often among the various mapping analyses.

Table 11: **QTL scan results**

Scan	Figure	Filename	<i>n</i>	perms	Data	Outcome	Fixed effects	Random effects	95% QTLs	50% QTLs
Scan1	35a	x6.rds	1,000		ccdataTAU	Litter	Gen	Funnel	1,5	2,4,6,7,11,17,18
Scan2	35b	x7.rds	1,000		ccdataTAU	Litter	Gen	Funnel, Mating	1,5	2,4,6,7,9,11,17,18,X
Scan3	36a	x101.rds	500		ccdataTAU	Litter	Gen	Gen l Funnel	5,17	1,6,7,8,11,15,18,X
Scan4	36b	x102.rds	500		ccdataTAU	Litter	Gen	Gen l Funnel, Mating	5,17	1,6,7,8,11,15,18,X
Scan5	37a	x103.rds	500		ccdataTAU_t10	Litter	Gen	Funnel	5,6,7	1,8,11,15,17,18,X
Scan6	37b	x104.rds	500		ccdataTAU_t7	Litter	Gen	Funnel	NA	1,4,17,18
Scan7	38a	x105.rds	500		ccdataTAU	Litter	as.factor(Gen)	Funnel	5	1,2,6,7,9,11,17,18,X
Scan8	38b	x106b.rds	500		ccdataTAU	Litter	I(0.8^(Gen-3))	Funnel	5	1,2,6,7,9,11,17,18,X
Scan9	39a	x107.rds	500		ccdata	Litter	Gen	Funnel	4, 17	1,2,11,12,14,15,16
Scan10	39b	x108.rds	500		ccdata	Litter_res	NA	Funnel	17	2,4,5,9,11,14
Scan11	40a	x109a.rds	500		ccdataTAU	Gdays_Adj	First_litter, Gen	Funnel	12	1,3,14,17
Scan12	40b	x109c.rds	500		ccdataTAU	Gdays_Adj	First_litter, Litter, Gen	Funnel	3,12	17

Table 12: Significant QTLs

QTL	Scan	Chromosome	Start	End	MGSCv37	GRCm38	Genes	H2	H2g
Lsq1	1-5, 7-8	chr5	119562288	130823943	chr5:119562288-130823943	chr5:119112279-130348073	215	0.0294	0.237
Lsq2	1-2	chr1	158200094	169393758	chr1:158200094-169393758	chr1:156269964-167463627	163	0.0261	0.21
Lsq3	9	chr4	57127262	65863628	chr4:57127262-65863628	chr4:57114390-66202594	165	0.024	0.157
Lsq4	5	chr6	143910081	149341971	chr6:143910081-149341971	chr6:143961561-149393449	95	0.0195	0.189
Lsq5	5	chr7	143703172	151245395	chr7:143703172-151245395	chr7:136511489-144059490	193	0.0179	0.174
Lsq6	3-4	chr17	80521533	94886379	chr17:80521533-94886379	chr17:80122193-94487039	106	0.0189	0.153
Lsq7	9-10	chr17	35294539	61878890	chr17:35294539-61878890	chr17:35157594-61739467	541	0.0138	0.0937
Lsq8	9-10	chr17	67337768	80144472	chr17:67523181-80144472	chr17:67173841-79745132	106	NA	NA
Gtimq1	11-12	chr12	34890781	50697203	chr12:34890781-50697203	chr12:34205916-49596216	66	0.00972	0.175
Gtimq2	12	chr3	42878518	52714350	chr3:42878518-52714350	chr3:43074596-52910428	37	0.00459	0.156

: Lifter done using NCBI Remap (<http://www.ncbi.nlm.nih.gov/genome/tools/remap>). Intervals are reported as the region above the 50% significance line. Gene counts retrieved from NCBI37 (<http://may2012.archive.ensembl.org/index.html>).

Genetic effects

I investigate the genetic effects of the CC founders at the identified litter size loci on chromosomes 5 and 1. I first investigate the relative contributions of the CC founders at the identified litter size QTL on chromosome 5 (Figure 42). I expect to see 12.5% of each of the eight founders at each locus; deviations from this proportion might indicate selection has occurred at the locus. I find an excess of WSB/EiJ haplotypes at the chromosome 5 locus (17.2%; $P=0.0101$; measured at the QTL peak), and possible deficiency of CAST/EiJ (9.24%; $P=0.0714$). Previous work suggests that an excess of the WSB/EiJ haplotype spans half of chromosome 5 in the CC inbred lines, consistent with these findings.¹⁰¹ By contrast, the chromosome 1 locus shows an excess of PWK/PhJ (19.4%; $P=0.000147$).

Chromosome 5 locus

I next extract estimates of the effect of each CC haplotype from the fitted genetic model to investigate the effect of each CC founder haplotype on litter size across the chromosome 5 locus (Figure 43). PWK/PhJ founder is consistently associated with lower litter size across this locus, whereas NZO/HILtJ and 129S1/SvImJ founders are associated with larger litter sizes. WSB/EiJ alleles are not obviously associated with litter size, despite their apparent enrichment at this locus. The effect sizes are reasonably constant across the QTL. C57BL/6J founder seems associated with smaller litter size at this locus, despite these mice typically being good breeders.

Across the QTL, none of the poorest 30% of breeding lines (determined by mean litter size) carries 129S1/SvImJ haplotypes, indicating that presence of 129S1/SvImJ is typically associated with good breeding performance ($P=0.006$ for test of top 80 vs bottom 80 lines). CAST/EiJ and PWK/PhJ haplotypes at this locus are typically found in low performing lines, and in a few high performing lines, but rarely in intermediate lines.

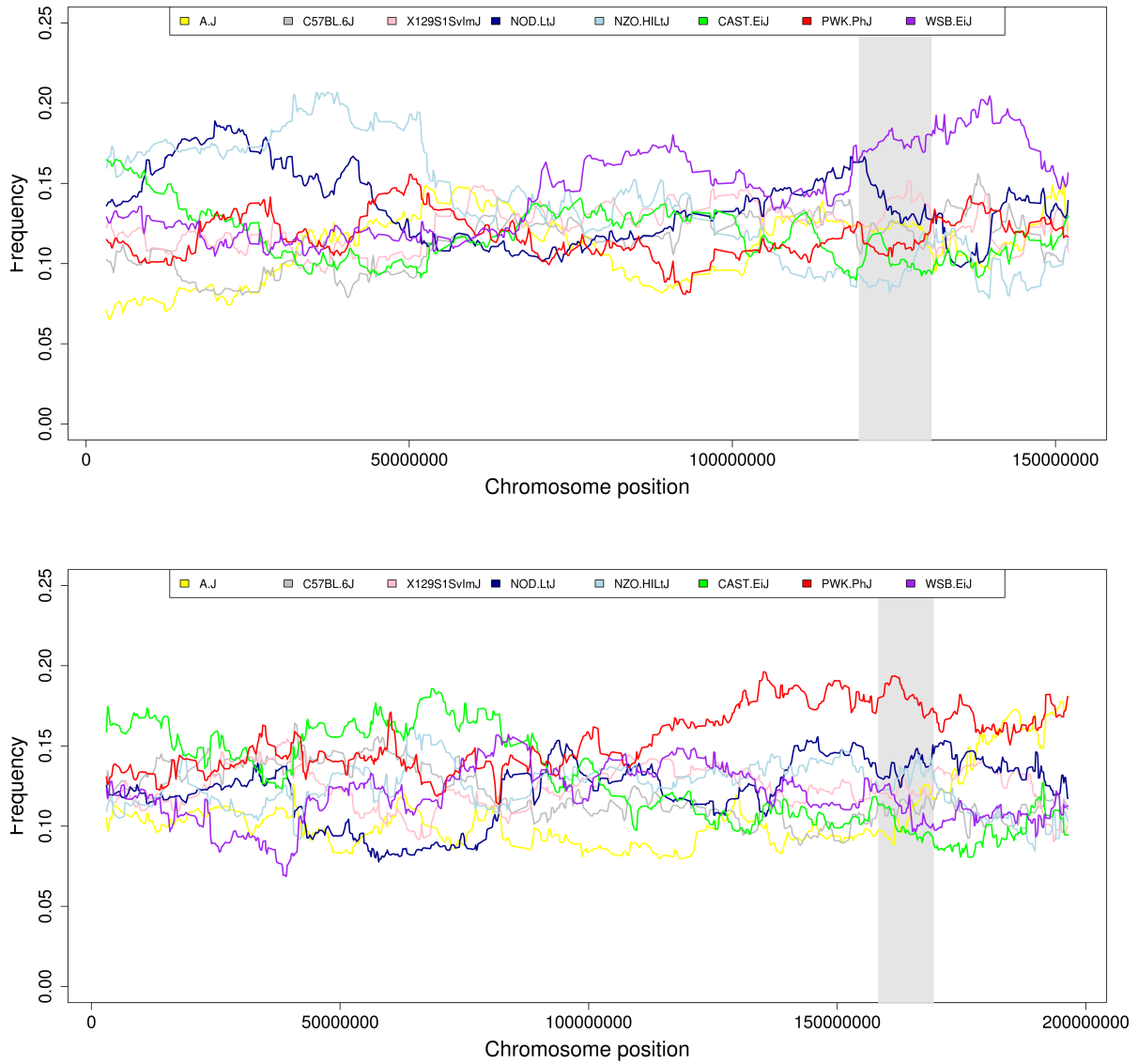


Figure 42: Chromosome 5 and 1 allele contributions. The upper panel (a) shows the proportion of each founder haplotype across chromosome 5. An excess of WSB/EiJ is seen at the QTL. The lower panel (b) shows the proportion of each founder haplotype across chromosome 1. An excess of PWK/PhJ is seen at the QTL.

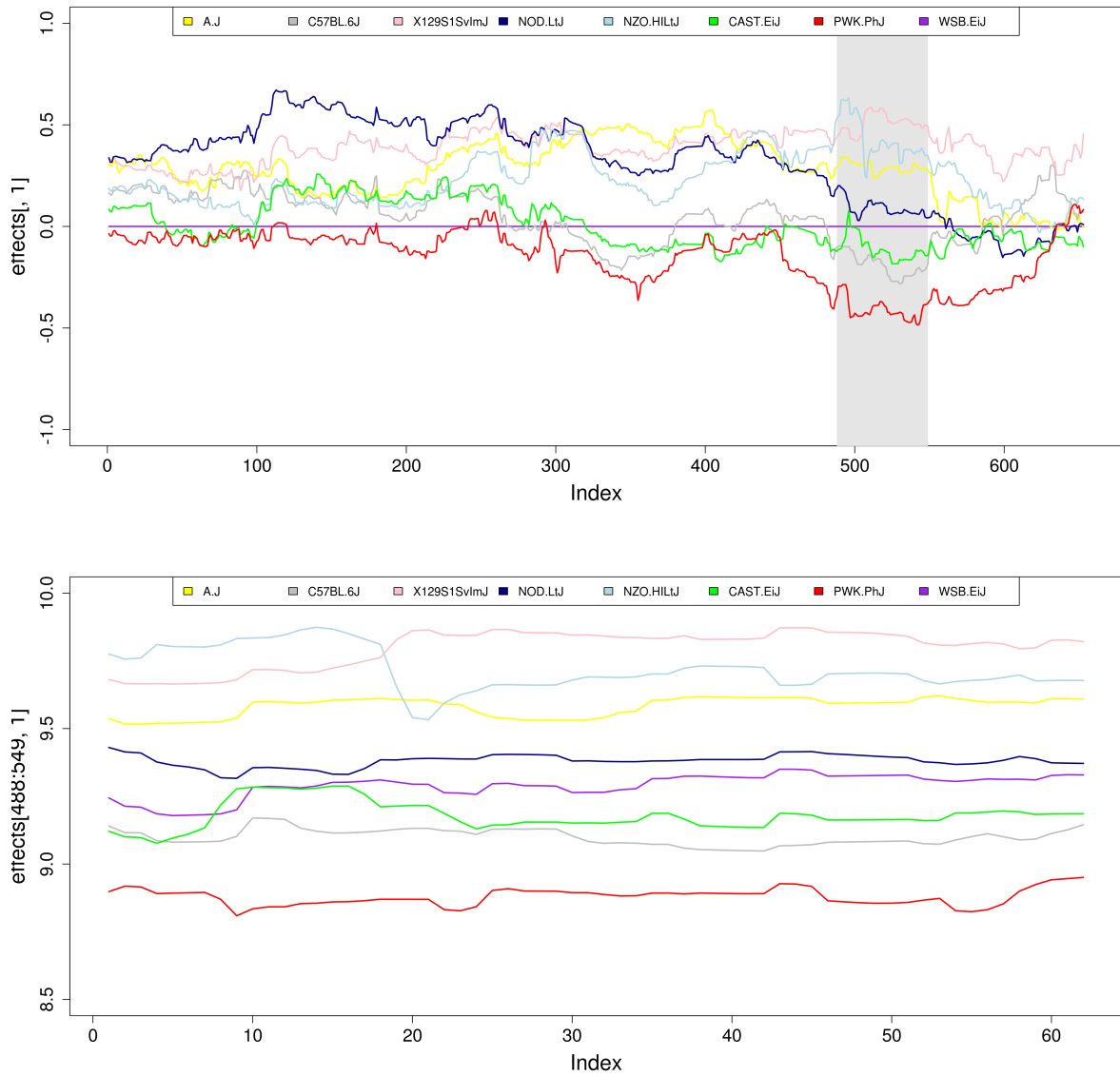


Figure 43: Allelic effects attributed to each founder across chromosome 5. The upper panel (a) shows the allelic effects across the chromosome relative to WSB/EiJ. PWK/PhJ founder haplotype is associated with smaller litter size across the QTL, whereas 129S1/SvlmJ is associated with large litter size. The lower panel (b) shows the absolute allelic effects across the QTL.

Chromosome 1 locus

A similar plot of the chromosome 1 locus is shown in Figure 44, with WSB/EiJ founder seemingly associated with smaller litters and 129S1/SvImJ associated with larger litters across the QTL. A distinct drop in litter size associated with PWK/PhJ founder is seen in proximal chromosome 1, at a region with some evidence for a second suggestive QTL.

Chromosome 17

The genetic mapping repeatedly identify chromosome 17 as showing a significant association with litter size. However, the mapping scans identify nonoverlapping significant QTLs, covering most of the chromosome. These mapping results suggest the involvement of multiple loci on chromosome 17. I show here the effect sizes across the whole chromosome for two QTL scans (Figure 45). These results suggest a complex effect on litter size from chromosome 17, which the mapping resolution of this study has been unable to disentangle. Chromosome 17 loci also show suggestive associations with time between litters indicating more-widespread involvement with reproductive traits.

Time between litters

Finally, I investigate the founder contributions and effect sizes for the significant chromosome 12 QTL for time between litters (Figure 46). This locus shows no notable variation in founder contribution, and some notable differences in effect between founders, with indication of clustering of founder effects. A/J haplotypes show a consistent association with the fastest time between litters, while CAST/EiJ, PWK/PhJ and NOD/LtJ haplotypes are associated with the slowest time between litters.

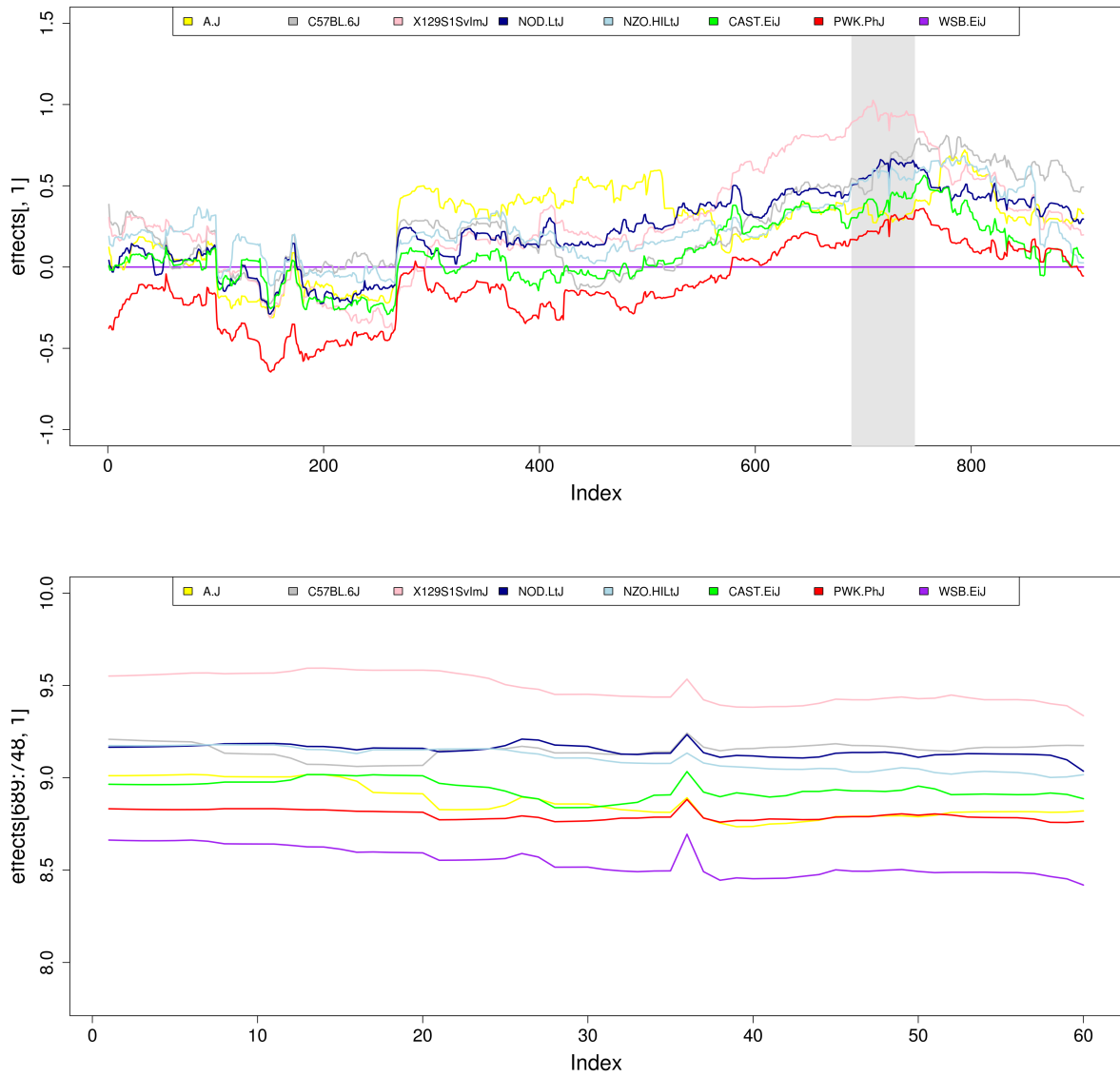


Figure 44: Allelic effects attributed to each founder across chromosome 1. The upper panel (a) shows the allelic effects across the chromosome relative to WSB/EiJ. WSB/EiJ founder haplotype is associated with smaller litter size across the QTL, whereas 129S1/SvlmJ is associated with large litter size. The lower panel (b) shows the absolute allelic effects across the QTL.

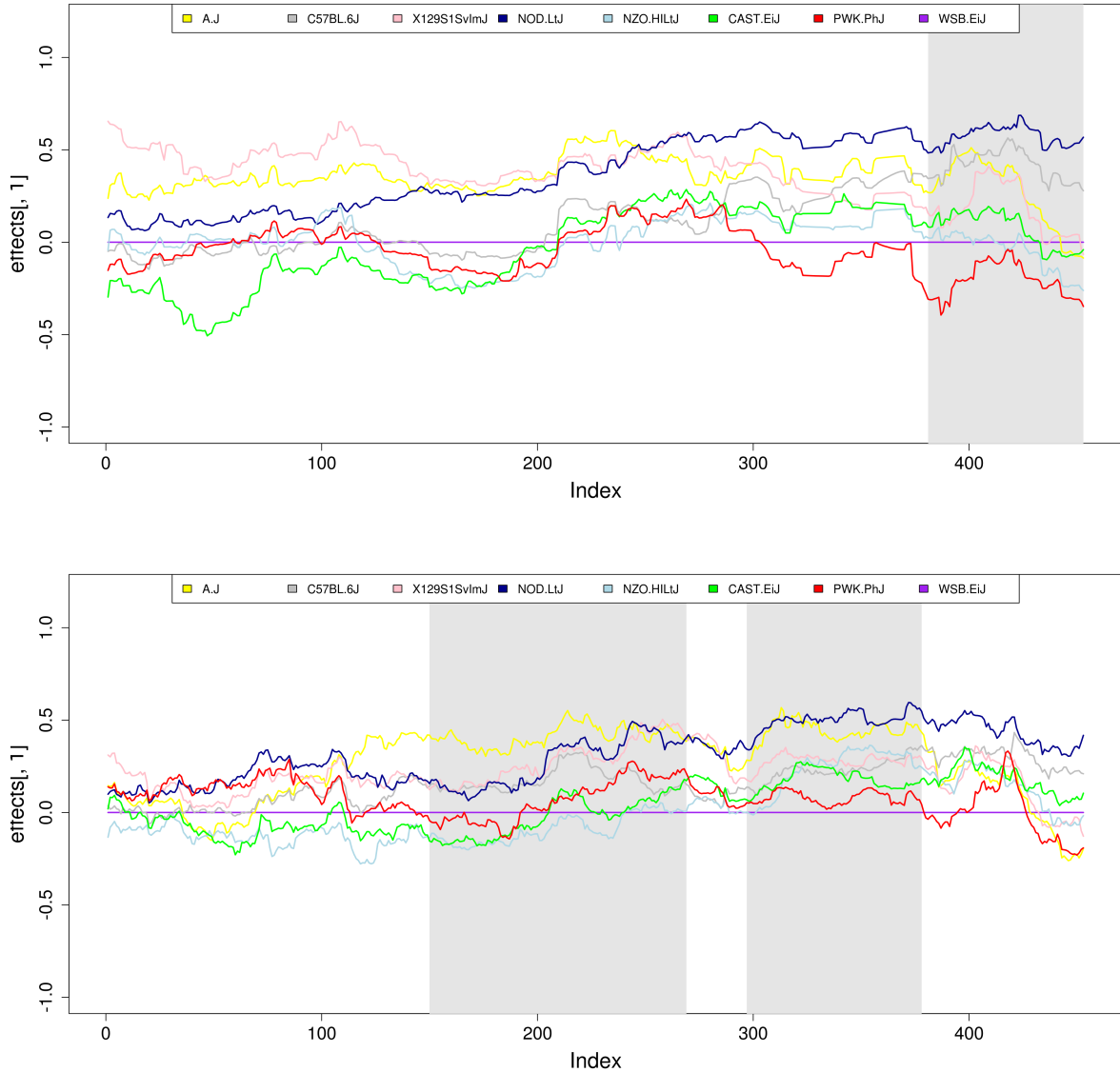


Figure 45: Allelic effects attributed to each founder across chromosome 17. The upper panel (a) shows the allelic effects across the chromosome relative to WSB/EiJ from QTL Scan 3. NOD/LtJ founder haplotype is associated with larger litter size across the QTL, whereas PWK/PhJ is associated with small litter size. The lower panel (b) shows the allelic effects across the chromosome relative to WSB/EiJ from QTL Scan 10. Two distinct QTLs have distinct patterns of effects.

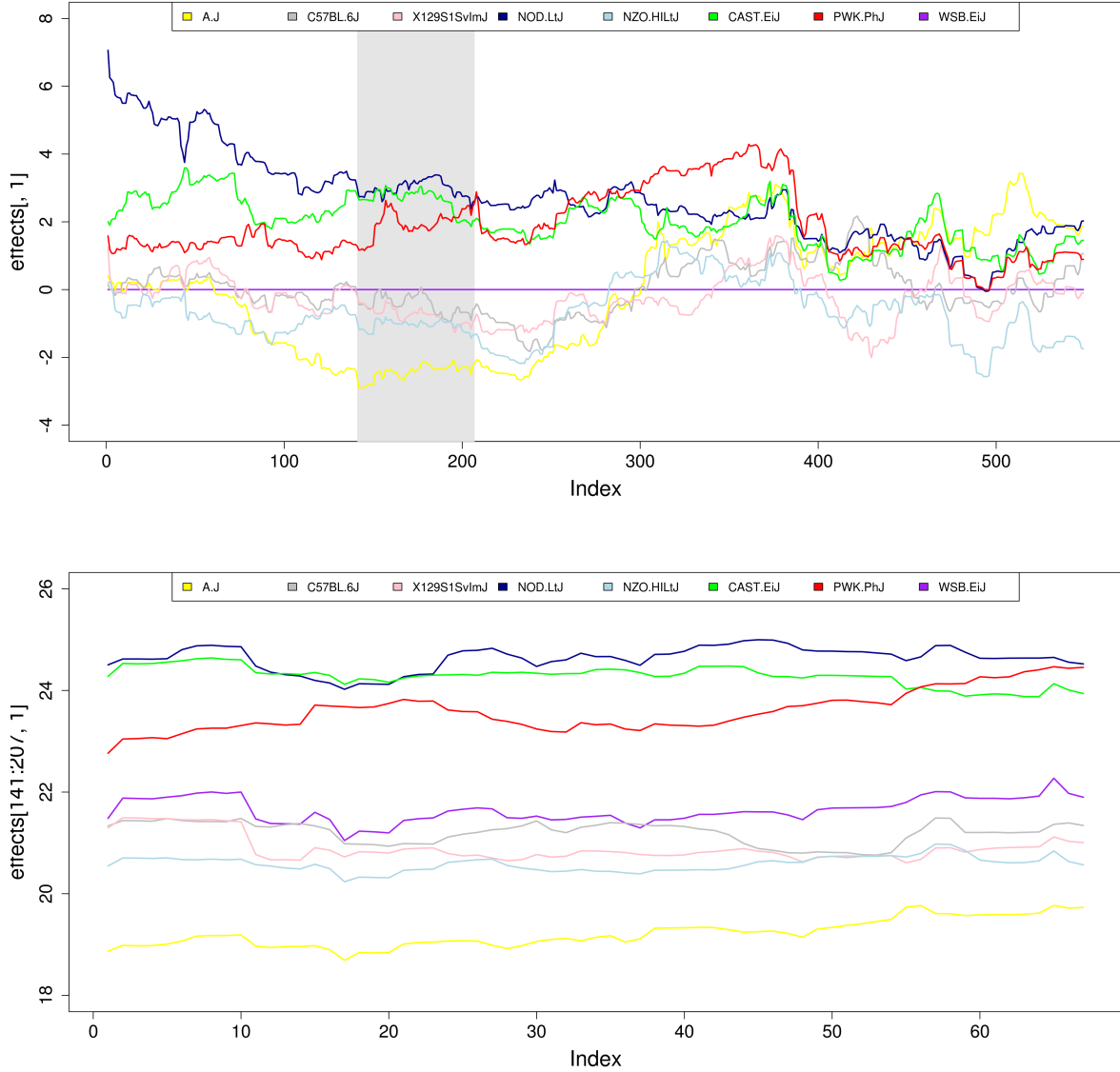


Figure 46: Allelic effects attributed to each founder across chromosome 12. The upper panel (a) shows the allelic effects on time between litters across the chromosome relative to WSB/EiJ. PWK/PhJ, CAST/EiJ and NOD/LtJ founder haplotypes are associated with longer time between litters across the QTL, whereas A/J is associated with shorter time between litters. The lower panel (b) shows the absolute allelic effects across the QTL. This QTL shows clear clustering of strain effects, which might be amenable to merge analysis.

Candidate genes

I further examine the identified litter size QTL on chromosome 5 to search for putative candidate genes and causal variants. The QTL spans ~11Mb and contains 215 genes (count includes predicted genes, RNA genes and pseudogenes; data from Archive Ensembl browser; <http://may2012.archive.ensembl.org/index.html>). Of these genes, 15 are reported to be associated with the term “reproductive system phenotype” in the Mouse Genome Informatics database (<http://www.informatics.jax.org/>) and seven are specifically associated with the term “abnormal litter size”: *Oas1d*, *Ptpn11*, *Aldh2*, *Atxn2*, *Atp2a2*, *Stx2*, *Tpst1*. These 15 genes are detailed in Table 13. Among these a number are expressed widely; only *Oas1d* and *Tpst1* are reported to be expressed exclusively in reproductive and related tissues (data from BioGPS [biogps.org] and MGI [<http://www.informatics.jax.org/>]). Among the remainder of the 15 genes *Scarb1*, *Adam1a* and *Piwil1* are also exclusively expressed in reproductive tissues, as are *Lhx5*, *B3gnt4*, and *Adam1b*, which are identified by Gene Ontology overrepresentation analysis (below), but not listed as associated with reproductive function in the Mouse Genome Informatics database. Other genes in the region, including *Abcb9*, *Brp* and *Ccdc92*, are also reported to be highly expressed almost exclusively in the testis.

Genes associated with reproduction and gamete generation are overrepresented threefold in the QTL, but are not significantly overrepresented after correction for multiple comparison (*Adam1b*, *Lhx5*, *B3gnt4*, *Adam1a*, *Gpr133*, *Piwil1*, *Rilpl1*, *Rilpl2*; analysis done using Gene Ontology Enrichment <http://geneontology.org/page/go-enrichment-analysis> and PANTHER www.pantherdb.org).¹⁸⁷ The genes in the QTL are strongly overrepresented for response to interferon- γ (Bonferroni-corrected P -value = 2.76×10^{-10}), as a result of the presence of ten members of the 2’–5’ oligoadenylate synthetase family (which includes *Oas1d*). Interferon- γ enhances blood flow to the site of implantation during pregnancy, and aids the development of the placenta suggesting a potential mechanism of action.

Table 13: **Chromosome 5 QTL genes associated with reproductive phenotype**

Symbol	Location (NCBIM37)	Reproductive phenotype	Affected sex	Litter size	Human disease
<i>Tbx3</i>	120120678-120134733	Abnormal female reproductive system biology	Female	No	Unhar-mammary sundrome
<i>Oas1d</i>	121364354-121371661	Decreased corpora lutea number and litter size	Female	Yes	Diabetes mellitus, insulin-dependent
<i>Ptprn11</i>	121580542-121641406	Impaired embryo impantation; decreased litter size	Female	Yes	Leopard Syndrome 1; Noonan Syndrome 1
<i>Adam1a</i>	121968613-121971704	Abnormal fertilization; asthenozoospermia	Male	No	NA
<i>Aldh2</i>	122016036-122043833	Decreased litter size	NA	Yes	Alcohol sensitivity
<i>Atxn2</i>	122161346-122266502	Enlarged seminal vesicle; decreased litter size	Male	Yes	Spinocerebellar ataxia 2; Parkinson disease
<i>Ppp1cc</i>	122608287-122625282	Abnormal sperm physiology	Male	No	NA
<i>Atp2a2</i>	122903531-122952183	Penis prolapse; decreased litter size	Male	Yes	Darier-White disease
<i>Kdm2b</i>	123320677-123465089	Oligozoospermia	Male	No	NA
<i>Clip1</i>	124027804-124134627	Teratozoospermia; reduced male fertility	Male	No	NA
<i>Scarb1</i>	125757457-125821464	Abnormal oocyte morphology; female infertility	Female	No	NA
<i>Piwi1</i>	129242121-129261349	Decreased testis weight; abnormal male germ cell	Male	No	NA
<i>Stx2</i>	129490432-129514449	Decreased seminal vesicle weight and litter size	Male	Yes	NA
<i>Gusb</i>	130464881-130478919	Male infertility	Male	No	Mucopolysaccharidosis
<i>Tpst1</i>	130549196-130611599	Reduced female fertility; decreased litter size	Female	Yes	NA

I examine the CC founder haplotypes using the CC viewer (<http://csbio.unc.edu/CCstatus/>), to identify any regions of common haplotype between founder strains. Of particular interest is the region 126.8–128.9Mb where 129S1/SvImJ and NZO/H1LtJ—the two founder haplotypes associated with largest litter size—largely share a haplotype (Figure 47). This region corresponds to the chromosome 5 QTL peak. It is also a region of very few genes, containing none of our previously identified candidates. It is however in close proximity to *Piwil1* and *Scarb1* on the forward and reverse strands respectively. This region also contains the transmembrane protein *Tmem132c*, which is expressed in uterus and umbilical cord.

Discussion

This work identifies putative loci associated with litter size in mice, and catalogs plausible candidate genes and functional variants at these loci. It also identifies loci associated with time between litters. Initial assessment show that the TAU cohort properties are consistent with those of the CC lines maintained at UNC,^{101,107} and that the heritability of reproductive traits is consistent with that seen in the UNC mice¹²⁴ and the wider literature. The strongest reported locus—the litter size QTL on chromosome 5—seems novel and not identified in other litter size mapping studies in mouse. This QTL explains around 3% of total variation in litter size, which is consistent with the effect size of other QTLs for complex traits.

Results

Investigation of the identified locus on chromosome 5 has identified a number of plausible candidate genes. These analyses suggest *Oas1d* and *Scarb1* as the most likely female factors in the target locus. Other genes in the locus are, however, reported to be associated with reproductive traits, and the possibility remains that male factors, multiple

genes or novel genes are involved. Investigation of sequence variants across this locus might help to further narrow the list of candidate genes. *Oas1d* is reported to be involved in ovarian follicle development and ovulation, and probably interacts with *Oas1a* (also found in this locus) to prevent cellular destruction during oogenesis.¹⁸⁸ The 2'–5' oligoadenylate synthetases also have a reported role in interferon-gamma sensing in the uterus, which modulates blood flow to the site of implantation during pregnancy, and aids the development of the placenta.

Scarb1 encodes a high density lipoprotein (HDL) receptor that mediates the selective uptake of plasma HDL cholesterol by liver and steroidogenic tissues. It is reported to affect fertility in humans and mice, through a range of potential mechanisms.^{189–193} Female mice with *Scarb1* knockouts are infertile, possibly as a result of premature activation of eggs caused by excess cholesterol.¹⁹⁴ During luteinisation, *Scarb1* expression increases dramatically to supply circulating HDL cholesterol to ovarian cells; *Scarb1* expression is essential for maintaining normal ovarian cholesterol homeostasis and luteal steroid synthesis.¹⁹⁵ *Scarb1* protein is also present in the yolk sac and placenta during development, and might therefore affect fetal growth by modulating maternal–fetal transport of cholesterol and other lipids.¹⁹⁶ A liver–HDL–ovarian axis is thus a plausible candidate for further investigation of fertility phenotypes in humans and model organisms. In support of this theory, a study that mapped HDL cholesterol in outbred mice identified significant QTLs on chromosomes 1 and 5 at locations that closely match our identified QTLs (Figure 48).¹⁹⁷ These authors identified *Scarb1* within the chromosome 5 locus and *Apoa2* within their chromosome 1 locus as candidate HDL cholesterol genes, raising the possibility that our measured litter size phenotype is conflated with circulating HDL cholesterol levels. *Apoa2* lies just outside our identified QTL, and contains a polymorphism associated with high HDL cholesterol levels that is only present in 129S1/SvImJ among the CC founders.¹⁹⁸ 129S1/SvImJ is also associated with larger litter size at the chromosome 1 locus. Future studies might explore HDL in CC lines in the context of *Scarb1*.

Our identified QTL on chromosome 1 contains nine genes previously associated with reproductive phenotype in mouse, and one (*Serpinc1*) associated with abnormal litter size phenotpye. Among these genes, *Adcy10* is reported to affect fertility with homozygous females exhibiting increased cholesterol and triglyceride levels. This QTL might also overlap with the reported modifier of obesity related sterility 1 (*Mors1*) QTL,¹⁹⁹ and with a large number of reported HDL cholesterol QTLs. Other QTLs identified in this study contain genes associated with embryonic cholesterol uptake, including the low density lipoprotein (LDL) receptor related proteins *Lrp2* on chromosome 2 and *Lrp6* on chromosome 6. The relationship of cholesterol with reproductive function, which might interact with obesity and metabolic syndrome, is discussed in several reviews.^{189,200–202}

Litter size has previously been investigated within the CC, as reproductive performance has been important to delivering the completed resource. An analysis of 102 traits in the UNC mice found one suggestive litter size QTL on distal chromosome 6 (93.9–128.3Mb; Build 36).¹²⁴ The authors report A/J haplotype to be associated with larger litter size, and NZO/HILtJ haplotype associated with smaller litter size. Our QTL mapping shows a suggestive peak in an overlapping but nonidentical region (113.9–128.4Mb; Build 37), and a nearby significant peak in the early generations of the inbreeding cross (143.6–149.3Mb; Build 37). A different pattern of strain effects is evident, with 129S1/SvImJ and C57BL/6J haplotypes associated with larger litter sizes (Figure 49), and 129S1/SvImJ is largely absent in the lines with the smallest mean litter sizes.

129S1/SvImJ and A/J share a haplotype in the distal region of chromosome 6, which is reflected in the HAPPY founder probabilities. This region contains the known reproductive genes *Capza3*, *Plekha5*, *Pde3a* and *Kras*, and corresponds to the low testis weight and sperm hammerhead QTLs *Ltsw2* and *Sh1*, respectively.²⁰³ One possible reason for the difference between studies is that the UNC study used data from G2:F5 generation and greater for mapping, whereas our strongest peaks in this region were associated with the early generations of inbreeding. Our analysis also uses substantially more data, by

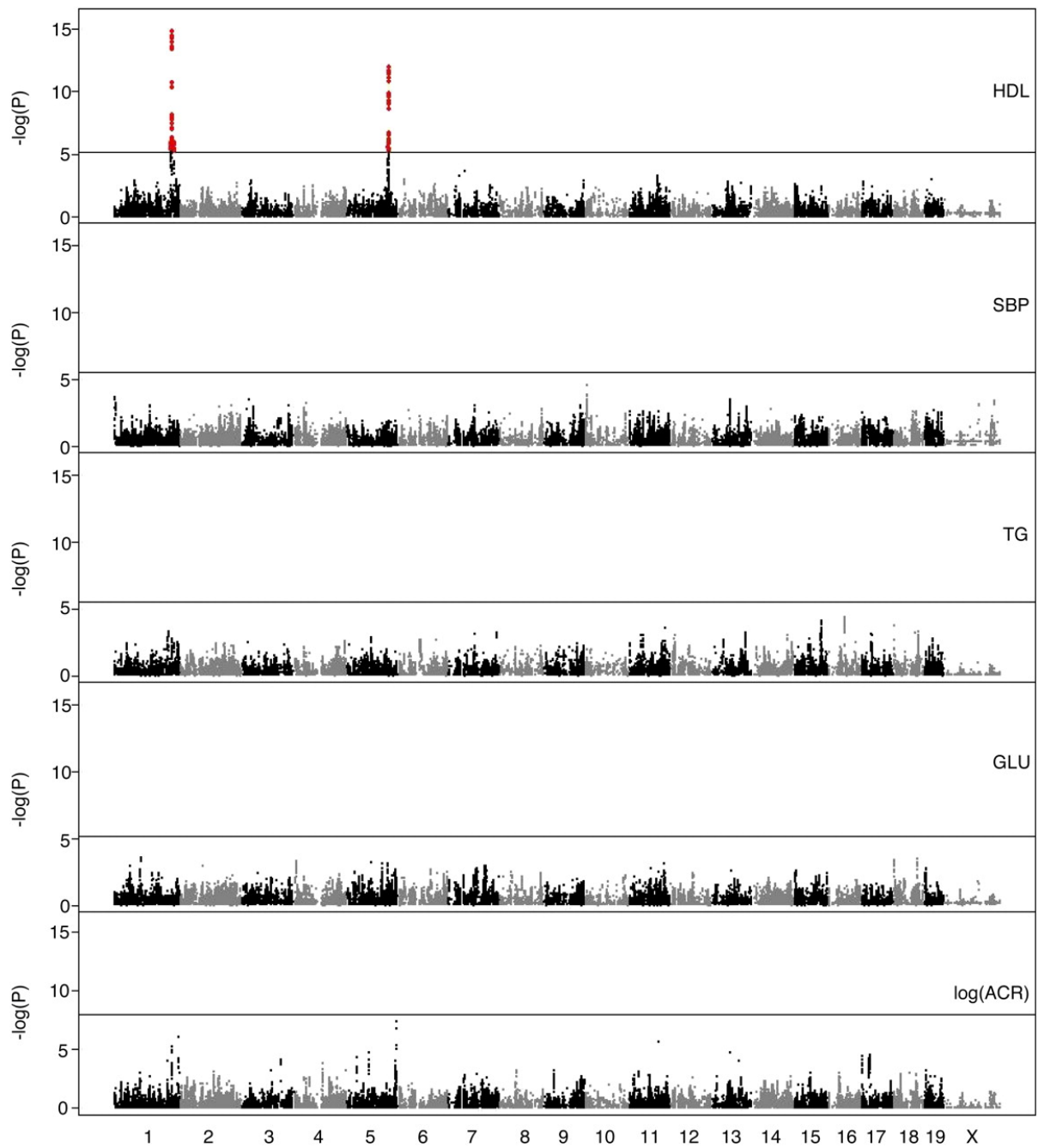


Figure 48: Genomewide association mapping with mixed model analysis (EMMA) over 44,428 SNPs for traits HDL, SBP, TG, GLU, and the log-transformed ACR. © 2012 Zhang *et al.*. Reproduced with permission under CC-BY.¹⁹⁷

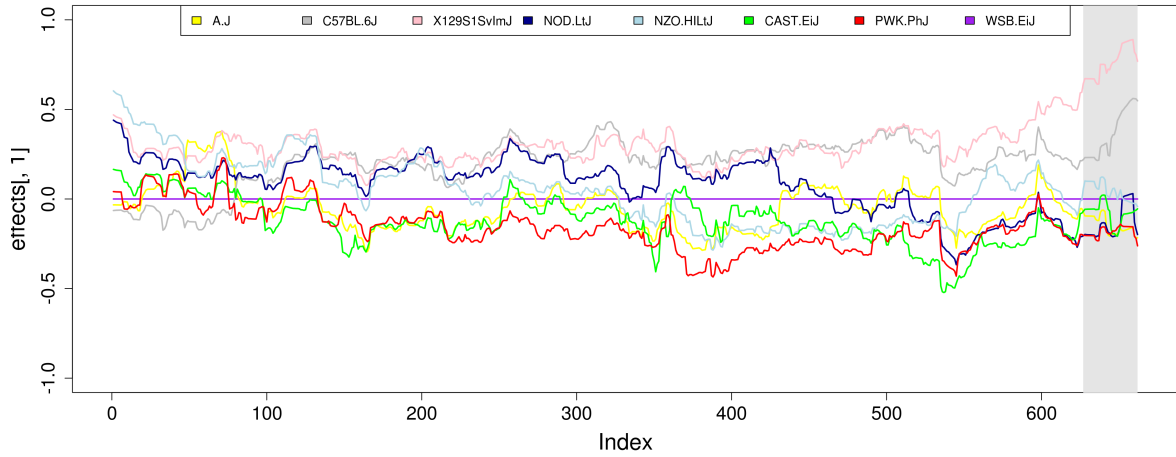


Figure 49: Allelic effects attributed to each founder across chromosome 6. C57BL/6J and 129S1/SvImJ founders are associated with larger litter size, particularly in the far distal portion of the chromosome.

using approximately the same number of CC lines, but including multiple generations and multiple litters from the same parents.

In our study, mapping in the early generations of the cross also identified a significant QTL on chromosome 7. This QTL contains 10 genes associated with reproductive phenotype, and one, *Ifitm5*, previously associated with bone defects and small litter size. Chromosome 7 was previously identified as *litsq* (litter size QTL) in chromosome substitution strain analysis of chromosome 7 derived from 129S1/SvImJ on a C57BL/6J genetic background, which also identified QTLs for body weight and composition.²⁰⁴ In this study, the substitution strains had significantly larger litter sizes. Interestingly, the TAU CC lines show an excess of 129S1/SvImJ haplotype across much of chromosome 7 (although not at the QTL), and 129S1/SvImJ is associated with larger litter size at the QTL, with PWK/PhJ associated with smaller litter size (Figure 50).

The chromosome 7 QTL also contains two imprinted gene clusters—the *Igf2–H19* and *Kcnq1* clusters—of four and fifteen genes, respectively, which are associated in mice with early embryonic lethality (paternal inheritance) and late fetal lethality (maternal



Figure 50: Allelic effects attributed to each founder across chromosome 7. The upper panel (a) shows the founder contributions across the chromosome. 129S1/SvImJ is overrepresented across much the chromosome, although not at the identified QTL. The lower panel (b) shows the allelic effects across the chromosome relative to WSB/EiJ. PWK/PhJ founder is associated with smaller litter size and 129S1/SvImJ with larger litter size.

inheritance). Genomic imprinting is thought to have a key role in fetal and placental growth, development and nutrition; it is discussed in more detail in Chapter 5. Some imprinted genes, including *Igf2*,²⁰⁵ have been reported to be associated with litter size in pigs.²⁰⁶ Parent-of-origin-specific effects, which might include imprinting effects, are also reported in female reproductive success in mice (defined as lifetime number of young surviving to sexual maturity), although the underlying mechanisms are unclear.²⁰⁷ Our litter size phenotype is likely to result from many complex processes including some influenced by imprinting effects. While our mapping approach might identify such regions, it lacks the capacity to discriminate between paternally and maternally inherited alleles. At later stages of inbreeding all individuals are likely to be homozygous at the locus so it would be impossible to identify imprinting effects. It might, however, be possible to further explore these effects using reciprocal crosses of well genotyped CC lines.

This analysis reveals evidence for involvement of chromosome 17 in determining litter size phenotype, with multiple loci likely to be involved. While our scans reveal significant peaks on this chromosome, the significant and suggestive regions span most of the chromosome. This extended region contains nine genes with reported associations with litter size including *Mep1a*, *Vegfa*, *Safb*, *Lbh* and *Fshr*, and overlaps with the previously reported mouse pregnancy rate QTL *Pregq4*. Also within the region is the luteinizing hormone/choriogonadotropin receptor *Lhcgr*, knockouts of which affect both male and female reproductive function, the major histocompatibility complex, which is suggested to be involved in mate choice in a range of taxa,^{208–210} and the linked *t* haplotype—a well-characterized selfish genetic element that contains recessive lethal alleles.²¹¹ Interestingly, our scans did not report an association with litter size on the proximal region of chromosome 17, a region that contains several genes with strong effect on reproductive function including *Prdm9*, *Igfr2* and *Sox8*. Two separate regions of chromosome 17 were also suggestive QTLs for time between litters indicating a widespread effect on reproductive function.

This study also identified significant QTLs for time between litters on chromosome 12 and 3. Gestation length—a component of time between litters—has been demonstrated to have a genetic basis in mice, primarily controlled by maternal genetics.¹⁶³ Notably, the founder effects seen at the chromosome 12 locus are inversely correlated with the reported gestation lengths for the CC founder strains in this prior study, with A/J haplotype associated with the shortest time between litters at this locus, despite exhibiting the longest gestation times.¹⁶³ One possible explanation for this contrast is that the QTL is associated with relative ease of initiating pregnancy, and gestation length is in fact a confounder. The chromosome 3 locus does not show the same distribution of effects. This prior study also reports a strong inverse correlation between gestation length and litter size, which was included as a covariate in our data, and a small seasonal effect on gestation length (most-notable in C57BL/6J mice).

Phenotype

Litter size is an extremely complex phenotype, in large part because it is a characteristic of both the parental and filial generations,⁸² as well as being affected by inbreeding depression and selection, and confounded with maternal and other indirect effects. Males are thought to contribute no more than 10% to observed variation in litter size,^{82,152} and I therefore focus discussion in this chapter mainly on female factor genes. However, a 2015 study reports that male reproductive phenotypes vary greatly among CC founder strains, with CAST/EiJ mice displaying a notable combination of deleterious male reproductive traits.²¹² The inclusion of the wild-derived founders in the CC population is intended to maximise diversity, and it is possible that prior studies that report a small contribution of males to observed variation investigated inbred lines lacking in relevant genetic variation. Nonetheless, the authors report that male infertility caused a large portion of the extinction of the inbred CC lines, and it seems likely that these male factors mainly affect the success or failure of mating, rather than quantitative variation in litter

size. A lack of information on failed matings, and on individuals and their sex, probably limits the ability of our approach to detect male factors that affect reproductive traits.

Litter size in mice has primarily been investigated in the context of informing animal husbandry, but the published research is fragmented and no coherent overview of the genetics of this phenotype is available. Litter size is reported to be responsible for most of the variation in reproductive performance in polytocous species (i.e. those ovulating multiple eggs),²¹³ which makes it a trait of interest for agricultural science. Nonetheless, polytocous and monotocous species are likely to have experienced differing selection pressures, meaning that the mechanisms governing litter size in laboratory mice might be different from those that are important in animal husbandry. Selection for both high and low litter sizes in mice is possible,^{82,152,155,214–217} although mouse litter size is a trait with low realized heritability,^{40,155,174,216,217} and consequently I might expect low power to detect genetic associations. In this regard, our calculated heritability and CV_G values indicate that relatively large-effect variants might be segregating in the population.

Litter size as a phenotype is determined by the interaction of many component traits including ovulation rate, fertilization rate, implantation rate, embryonic mortality and uterine capacity. Ovulation rate sets an upper limit to litter size, and increased ovulation rate is the primary cause of increasing litter size in mice selected for this trait. Litter size is a trait for which intermediate levels are likely to be evolutionarily optimal (although this pressure might be relaxed in laboratory populations). Adjustment to an intermediate level is probably most efficiently done by control of ovulation rate, as control of litter size mediated by implantation rates or embryonic mortality is wasteful of reproductive resources. Nonetheless, inbred mouse lines that produce large litters have been shown to do so using different adaptive strategies,²¹⁸ indicating that multiple male and female processes are involved in determining litter size phenotype; among female parameters, differences between lines were observed in progesterone concentrations, oocytes per follicle, ovulation rates and gestational loss.

Several genes are known to affect litter size in mice and agricultural animals. In particular, strong evidence exists that *Gdf9* (mouse chromosome 11) and *Bmp15* (mouse chromosome X) have roles in determining ovarian function and especially in regulating polyovulation.^{158,219–224} *BMP15* and *GDF9* are also associated with ovarian function and premature ovarian failure in humans.^{221,225,226} QTL mapping in our study did not identify these genes as being significantly associated with litter size in the TAU CC mice, maybe due to lack of functional variants in the founder strains. Mapping does, however, identify a suggestive QTL at a locus near *Bmp15* on chromosome X. *Gdf9* and *Bmp15* might act by inducing cholesterol synthesis in cumulus cells, suggesting a possible link with the postulated action of *Scarb1*, and other candidate genes²²⁷

Other QTLs have been associated with litter size in mice,^{40,159,160,228–231} often in crosses of lines with extreme phenotypes, although few underlying genes are identified; epistatic⁴⁰ and dominance²²⁹ effects are reported to have a notable role in determining this phenotype. QTLs have been identified for other reproductive phenotypes in mice including hormone-induced ovulation,^{156,157,229,232} and male factor infertility.²⁰³ The loci identified in our study do not greatly overlap with these prior loci, although this result is not necessarily surprising given the range of possible mechanisms that combine to produce the litter size phenotype. Many QTLs for litter size and reproductive traits are reported to overlap QTLs for growth traits.²²⁹ Selection for increased body weight results in increased litter size and a corresponding increase in ovulation rate, reinforcing a link between reproductive performance and body mass.²¹⁵ Selection for increased litter size also caused an increase in ovulation rate (60%), but showed an accompanying 87% decrease in luteinizing hormone receptor induction indicating that different mechanisms are involved. Many additional loci have been associated with reproductive traits in cattle, sheep, pigs and rabbits, and hundreds of genes are shown to be necessary for fertility and viability in mouse knockout studies.

Inbreeding depression

This study assesses genetic associations with litter size in a large mouse population in the context of severe inbreeding. In a progressively inbred population such as the CC, determination of litter size is confounded with the process of inbreeding depression, probably as a result of the accumulation of recessive deleterious alleles in both parents and offspring. Mice are consistently reported to suffer a decline of 0.5 offspring per 10% inbreeding and our data support this rate of decline. In addition, many matings during inbreeding are functionally infertile and unproductive, resulting in a reduction in the total number of litters produced. In the context of inbreeding, the effects on litter size are reported to be mediated largely through maternal factors (~40% of effect; especially implantation rates) and embryonic viability (~60% of effect),⁸² although inbreeding might also have different effects on components of litter size in different stocks of mice, and an interaction might exist between levels of inbreeding in the mother and offspring (in particular affecting preimplantation mortality).²³³ In this setting, it is likely that most oocytes get fertilized and become corpora lutea. These corpora lutea are, however, highly sensitive to the environment, including handling of the mother.¹⁷³ The effect of inbreeding on litter size is likely to not be mediated by an effect on body size, which tends to stay the same during inbreeding as the reduction in number of offspring caused by inbreeding leads to an improved maternal environment for growth.

In this study, I account for the effect of inbreeding depression by including generation of inbreeding in the mapping model as a linear effect, as a factor and as the estimated heterozygosity at each generation, all with largely similar results. Thus while I have mapped genetic associations with litter size in a progressively inbred population, the mapped trait is in effect the variation in litter size once the variation attributable to generation of inbreeding (as an approximation of the effect of inbreeding depression) is accounted for. Once the CC lines are fully inbred, parental and offspring genotypes will

be theoretically identical at all loci, negating the differences between the three different genotypes influencing a trait; inbreeding depression resulting from the accumulation of deleterious alleles and hybrid incompatibility will also not be issues in the fully inbred lines. Thus investigation of litter size in the completed CC lines might reveal more about the genetics of this trait.

CC population

The CC has been designed as a ‘genotype once, phenotype many times’ resource, whereby the final inbred lines will be genotyped and QTL mapping can be done on the assumption that the mice in which phenotype is measured will be at least as homozygous as the genotyped individual. This analysis, by contrast, uses genotypes from partly inbred mice and probably from later generations than many of the measured trait values. The precise identity and generation number of the genotyped individuals is not documented, although the mean seems to be 5 generations of inbreeding (see Chapter 3). Many of the recorded trait values, therefore, represent the phenotypes of individuals with more genetic variation than is recorded in the genotypes, which will have decreased the power to detect genetic associations. The full dataset even includes litter size data for the offspring of the F2 generation, who are not yet inbred. Including data from the earliest generation of inbreeding increased the number of litter size measurements that could be included in the analysis at the cost of the precision of the genotypes associated with them. As giving birth is a discrete event, I believe it appropriate to include multiple measurements from the same inbred lines and matings, analagous to taking measurements from multiple individuals in the final genetically identical lines. In this regard, it is reassuring that the inclusion of the Mating variable that denotes the individual matings within an inbred line has no material effect on the outcome of the mapping analysis.

One potential confounder of QTL mapping of litter size in mice is hybrid sterility—a well-known phenomenon in mice, which has chiefly been investigated in *M.m.domesticus*

$\times M.m.musculus$ crosses.^{85,234,235} These crosses produce sterile males, resulting from interactions between the *M.m.musculus* X chromosome and *Prdm9* on chromosome 17, as well as other loci.^{87,236} The CC design took this incompatibility into account, with PWK/PhJ mice selected instead of the better-studied PWD/PhJ (which are known to be infertile in crosses with C57BL/6J).²³⁷ Evidence now shows that *M.m.domesticus* $\times$ *M.m.castaneus* crosses also result in male hybrid sterility.⁹³ Other incompatibilities were discovered during the CC breeding process, including NZO/HlLtJ $\times$ CAST/EiJ (incompatible), NZO/HlLtJ $\times$ PWK/PhJ (incompatible), PWK/PhJ $\times$ 129S1/SvImJ (infertile male offspring), and 129S1/SvImJ $\times$ NZO/HlLtJ (low litter size with 129S1/SvImJ females).^{107,124} These incompatibilities have been investigated in more detail in the UNC lines of the CC, although no full published report is yet available (D Aylor, personal communication). Such incompatibilities represent another mechanism that might influence litter size in the incipient CC lines. Nonetheless, incompatibilities should typically result in infertile phenotype, which is not well measured by the TAU CC breeding records, and they should manifest early in the inbreeding process. Any inbred lines that suffered from hybrid sterility before the lines were genotyped is not included in the mapping analysis. Thus, primarily I map litter size and other traits in a population that has escaped the effects of hybrid sterility.

Previous work has shown that the pre-CC data is an appropriate resource for QTL mapping, with sufficient power to identify QTLs for a range of traits. In this context, our identified QTLs are wider than might be expected, and are of relatively low significance when assessed by permutation. Several factors might have confounded the mapping of litter size in these animals, and diminished our power to detect significant associations. First, as discussed above, traits such as litter size are extremely complex, representing the outcome of multiple processes. Second, many of our traits are measured in generations prior to the genotyped individual for each line, with the result that some variation in phenotype is not accounted for by the measured genotypes. Third, processes such

as inbreeding depression and hybrid sterility are highly likely to affect reproductive phenotypes in the CC lines, and our models might not adequately correct for these processes.

Data quality

Finally, the quality of the original litter size dataset might have affected the outcome of this analysis. Unfortunately, the quality and scope of available breeding records turned out to not meet the expectations I had when I initiated the project, which resulted in time-consuming data curation, and a lack of ability to investigate many aspects of fertility. I critique our project design further in Chapter 8. The TAU breeding records contain visible errors and duplications, which are likely to reduce the power of the QTL mapping. Efforts were made to correct these errors during data curation, but some may still remain. In addition, it is possible that errors in these files have led to inaccuracies in the mating of mice at TAU. This suspicion is reinforced by the fact that a number of the CC lines maintained at TAU have been found to contain contributions from fewer than eight founder strains.¹⁰¹ In the context of obvious errors in the TAU breeding records, it is reassuring that efforts to clean the data improve the strength of the identified genetic associations with reproductive traits. For example, careful curation and correction of the recorded dates of mating and birth resulted in the emergence of QTLs for time between litters that were not evident from the raw dataset. More insight might well be obtained from applying a better understanding of the biological processes involved to the QTL mapping. I discuss the importance of such biologically sensitive data management further in Chapter 8.

Similarly, despite much effort and time investment, I have been unable to satisfactorily unravel the natural history of the CC lines initiated in Kenya and subsequently maintained at TAU. I see at least two obvious problems: first, inconsistent generation labeling; and second, cousin lines that have been genotyped, but for which breeding records are absent.

The CC population initiated and maintained at TAU seems to not share these problems and I therefore base mapping results and analyses on this population—while not perfectly recorded, its history is at least more-consistently documented.

The manner of maintenance of the breeding records potentially discards important information relevant to reproductive phenotypes. For example, it might be that the most important points within our data are those at which a given line fails to breed, which typically involves a severe drop in litter size between generations from a normal number of births to none. In the context of mouse inbreeding, around half of vaginal plugs fail to be pregnancies (i.e. implantations) and could conceivably be regarded as a litter size of zero.¹⁷³ Owing to the nature of our dataset, it is difficult to identify failed matings, and indeed to differentiate these informative data points from those in which lines have been discontinued. I requested from TAU details of which lines are extant, in the hope that this would facilitate survival analysis of these data; to date this information has not been forthcoming. Nonetheless, some form of per-line survival analysis, investigating the maximum generation of inbreeding achieved by each inbred line (or a binary trait of survival to a given generation of inbreeding) might reveal additional loci associated with reproductive traits.

In this context it is important to be clear about definitions of fertility (the potential for reproduction) versus fecundity (the reproductive output of an individual). Our study has mapped a measure of fecundity (i.e. litter size), and owing to the structure of the TAU breeding records—which selectively omit data from unproductive matings—it can offer little insight into the genetics of fertility (i.e. proportion of successful matings). In fact, this study in effect maps QTLs associated with litter size as a quantitative but non-zero trait among a subpopulation of mice who bred relatively well in the context of severe inbreeding and which was presumably selected to some degree for large litter size (necessary for initiating matings to produce the subsequent inbred generation). Disentangling the various contributory factors in this setting is not straightforward, even

before the poor quality of the breeding records is considered.

Summary

In this chapter I report the mapping of QTLs for litter size and time between litters, and identify a number of significant loci. Unlike most previous studies, which have focused on identifying QTLs that affect litter sizes in high versus low litter size lines, this work maps litter size QTLs in the context of an inbreeding population. It identifies novel QTLs, which might reflect different underlying biology that determines litter size between the CC population and the mice used in prior studies. Importantly, the dataset only provides information on succesful littering events, so the trait mapped is litter size among only successful litters. Failed matings, which could theoretically be regarded as litter size of zero are not reliably included in the dataset. The most robust litter size QTL on chromosome 5 contains plausible candidate genes, including *Scarb1*. This QTL explained around 3% of total variation in litter size, and 24% of the variation attributable to genetic factors. Other QTLs with effects of similar magnitude are also identified for litter size and time between litters.

In the next chapters I move on to descibe a separate investigation of parent-of-origin effects in humans. I return to a further discussion of this mouse work in Chapter 8.

Chapter 5 | Parent-of-origin effects

Clever father, clever daughter; clever mother, clever son

—Russian proverb, unattributed

The epigenetic basis of variation

In prior chapters, I describe an investigation that links specific regions of the mouse genome with variation in reproductive traits. This study identifies loci associated with variation in litter size and time between litter traits, and suggests plausible candidate genes underlying these effects. One of the identified QTLs, for litter size on chromosome 7, contains two well characterized imprinted gene clusters associated with embryonic and fetal lethality.

At present, the extent of the contribution of epigenetic processes, including imprinting and other parent-of-origin effects, to complex trait variation remains to be fully determined. In the remainder of this chapter I discuss the definitions of parent-of-origin phenomena, and then give an overview of these effects and their underlying mechanisms. In subsequent chapters I report the results of an investigation into the prevalence of parent-of-origin-specific gene expression in humans, in a population of Dutch twins.

Genetic studies of mammalian complex traits, such as genomewide association studies, have not yet identified individual genes with effects sufficient to explain all observed heritable trait variation in most cases. A number of suggestions have been made as to the location of this missing heritability.¹³ First, heritable variation could result from rare and poorly tagged structural and copy number variants. Second, it could be caused by many common small-effect additive variants, rare large-effect variants or rare combinations of variants, which are not well analysed in genomewide association studies. Third, the causes of variation could be undetected owing to insufficient understanding of the genetic

architecture of investigated traits and the use of underpowered detection methods. Fourth, gene–gene and gene–environment interactions might have a significant role in determining variation. Fifth, epigenetic inheritance might contribute to observed heritable trait variation.

A number of epigenetic effects are suggested to contribute to the heritability of variation in complex traits.^{13,238,239} Epigenetics—“the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states”²⁴⁰—has a clear role in determining cellular and organismal phenotype, and epigenetic marks are (to varying degrees) stable and heritable through mitotic cell division. Several underlying mechanisms of epigenetic inheritance between generations are proposed, which form two broad groups: parental effects (i.e. fetal programming and behavioural transfer); and germline transmission (e.g. DNA methylation, persistent histone marks, and RNA transmission), which might include *bona fide* transgenerational germline transmission (i.e. transmission of acquired epigenetic modifications to the F3 generation and beyond).²⁴¹

The extent to which these mechanisms affect heritable traits is uncertain, and the proposed role of transgenerational germline transmission remains controversial.^{242–245} Importantly, only persistent epigenetic modifications will contribute to estimates of heritability; such modifications are equivalent to DNA mutations and are likely to be in linkage disequilibrium with marker SNPs, meaning they are not a source of missing heritability in genomewide association studies.²⁴⁶ A mutation study in arabidopsis demonstrates the action of such epigenetic quantitative trait loci,²⁴⁷ although such heritable epigenetic modifications are thought to have more of a role in plants than in mammals.

Parent-of-origin effects

One particular class of (primarily, although not exclusively) epigenetic effects that contribute to heritable trait variation are parent-of-origin effects, in which the phenotypic effect of a genetic factor is associated with the sex of the parent from which it is inherited. Such phenomena conflict with the expectations of mendelian inheritance, which assumes functional equivalence of maternally and paternally inherited genomes. To date, most genetic analyses of complex traits have similarly assumed equivalence of maternally and paternally inherited alleles, in part owing to the cost and difficulty of identifying parental origin. Nonetheless, differential effects of genetic variants depending on from which parent they are inherited is a plausible mechanism proposed to explain some of the observed missing heritability.¹³

Parent-of-origin phenomena are observed when the phenotypic effect of a genetic factor is associated with the sex of the parent from which it was inherited. This difference in phenotype manifests in heterozygotic offspring, such that an offspring that is A_1A_2 at a given locus shows a different phenotype to a sibling that is A_2A_1 (where, for example, the maternally inherited genotype is presented first; Figure 51). Parent-of-origin phenomena are reported in a wide range of traits and disorders across many taxa, with several possible underlying mechanisms. Such phenomena are typically loosely identified and discussed in published articles as ‘parent-of-origin effects’.

While reviews specifically addressing the topic exist,^{248,249} the term ‘parent-of-origin effect’ is not well defined in the published literature. It is often conflated with ‘genomic imprinting’ (see below), and is therefore associated with potentially misleading assumptions about the underlying mechanism of action. In addition, some confusion in definition results from the multiple meanings of the word ‘effect’, which is variously interpreted as the outcome of an event (i.e. Process A results in Effect B) and the mechanism or process of generating that outcome (i.e. Observation B results from Effect A). As an example,

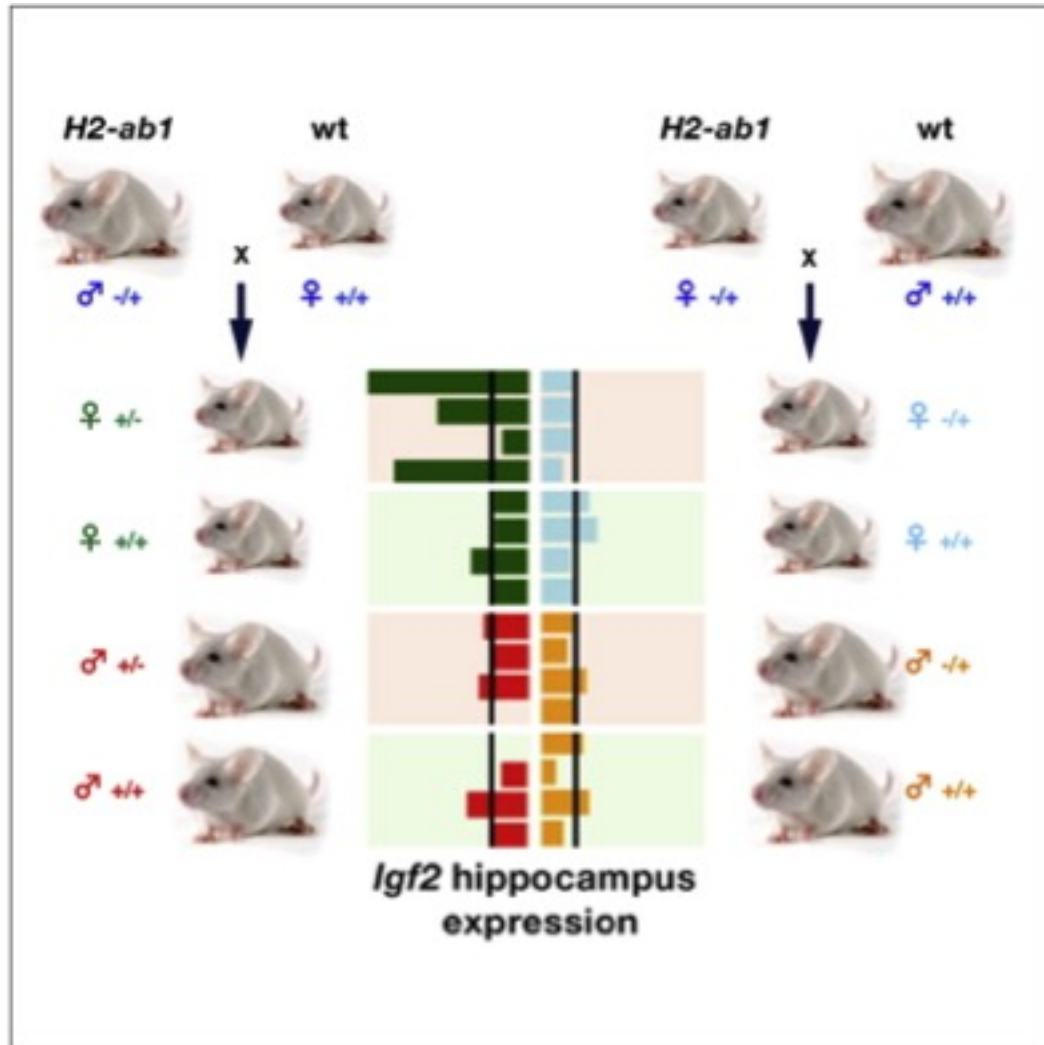


Figure 51: Example parent-of-origin effect in a heterozygote cross of *H2-ab1* knockout mice. When the *H2-ab1* knockout is inherited from the father (left side), significant changes in *Igf2* hippocampal gene expression are seen compared to homozygotes. When the *H2-ab1* knockout is inherited from the mother (right side) no such change is seen. In this case *H2-ab1* is differentially expressed by sex, so the parent-of-origin effect is only seen in female mice (green vs blue). Copyright Mott et al. 2014. Modified with permission under CC-BY.

the statement “the parent-of-origin effect of gene X produces a parent-of-origin effect on trait Y” uses two valid but distinct meanings of the term ‘parent-of-origin effect’. Such imprecision results in a lack of clarity as to what processes and phenomena are being discussed. Throughout the following chapters I thus endeavor to provide precise descriptions, such as parent-of-origin-specific gene expression when possible.

Examples

A classical example of a parent-of-origin phenomenon is in the breeding of hybrid species such as mules. A mule is the offspring of a male donkey and a female horse; it is phenotypically distinct from the offspring of the reciprocal cross—the hinny.²⁵⁰ Similar differences in phenotype are seen between the offspring of reciprocal matings of other closely related species,²⁵¹ and in reciprocal matings of phenotypically distinct breeds within a species.^{252,253} In such hybrids, the offspring phenotype is associated with parental origin of the inherited chromosomes. Note, that such phenotypic differences can be observed without invoking assumptions about underlying mechanisms.

Karyotype-associated parent-of-origin phenomena are also seen in clinical phenotypes; for example, parental origin of the extra X chromosome is associated with symptoms in Klinefelter Syndrome (47,XXY),²⁵⁴ and parental origin is possibly important in other chromosomal disorders. Parental origin of the deleted region of the long arm of chromosome 15 determines phenotype of the reciprocally inherited disorders Prader–Willi and Angelman Syndromes.²⁵⁵ Importantly, these syndromes result from the action of two complementary parent-of-origin mechanisms: parent-specific loss of the chromosomal region interacts with parent-of-origin-specific expression of a number of imprinted genes in the same region.²⁵⁶ Parent-specific chromosomal deletions are also associated with phenotype in other disorders (e.g. William Syndrome).^{257,258}

A range of other human traits and disorders are reported to exhibit parent-of-origin-specific

genetic associations, including type 2 diabetes,^{259–261} obesity and metabolism,^{262–265} cancer,^{259,266} age at menarche,²⁶⁷ autoimmune disease,^{268–270} and specific language impairment.²⁷¹

Genomic imprinting

The best-documented parent-of-origin mechanism is genomic imprinting (often simply referred to as ‘imprinting’), whereby offspring express only the maternally or paternally inherited copy of certain genes. Forms of imprinting are found across many taxa including flowering plants, fungi, insects and mammals; here I focus discussion on genomic imprinting in eutherian mammals—primarily mouse and human. Genomic imprinting results in silencing of an allele: a maternally imprinted gene results in only the allele from the father being expressed, and *vice versa*. Evidence suggests that genomic imprinting is a dynamic tissue-specific process, possibly associated with regulation of gene dosage. The mechanisms, origins and effects of genomic imprinting are comprehensively reviewed elsewhere.^{249,272–276}

The canonical demonstration of genomic imprinting effects in mammals was the creation of mouse uniparental embryos (and uniparental disomy) in the 1980s, whose gene function differed depending on parental origin.^{277,278} Prior work had identified paternal imprinting of the X chromosome in marsupials as a mechanism of sex-chromosome dosage compensation.²⁷⁹ These observations were followed by the identification of specific mammalian imprinted genes in the 1990s (e.g. *Igfr2*,²⁸⁰ *Igf2*^{281,282} and *H19*²⁸³), and diverse methods have since identified a range of additional imprinted genes in both mice and humans. The timeline of the emergence of genomic imprinting as an explanation for parent-of-origin effects is discussed in detail in the above reviews and elsewhere.²⁸⁴

The number of imprinted mammalian genes is not known with certainty, but is thought to be low, with no more than a few hundred imprinted genes predicted. Currently 100–150

imprinted genes are verified in mouse, with around half these genes also verified in humans. Reported imprinted genes generally concur in mouse and humans, but some discordance exists between reports; around 12 loci containing a core of imprinted genes seem conserved between the two species. While some studies showing much higher numbers of imprinted genes²⁸⁵ have been shown to be at risk for false positives,^{286,287} novel imprinted loci continue to be reported.^{288–290}

Catalogs of imprinted genes in humans and mouse are available online (e.g. *MouseBook Imprinting Resource*,²⁹¹ *geneimprint*,²⁹² *Web Atlas of Murine Genomic Imprinting and Differential Expression*,^{293,294} and *Catalogue of Parent of Origin Effects*^{295–298}), although whether all these resources are actively maintained is unclear.

Allele-specific expression from imprinted genes is potentially tissue specific, varies by individual and depends on developmental stage.^{289,299} Such variation might explain why surveys of imprinting typically identify overlapping but nonidentical sets of imprinted genes. Random monoallelic expression—single allele expression in which the random allelic choice occurs somatically—also contributes to this variation.^{300–302}

A number of epigenetic mechanisms are associated with differential gene expression in imprinted genes: primarily DNA methylation, but also allele-specific histone modification and the action of long noncoding RNAs.²⁸⁴ Imprints are transmitted intact (sometimes modified) through mitotic cell division but are erased during germline development, with imprinting patterns re-established in a sex-specific manner in spermatogenesis and oogenesis. This erasure of methylation and other imprinting marks limits the opportunity for transgenerational inheritance of acquired epigenetic modifications. Similarly, DNA methylation and chromatin modifications are almost entirely erased during epigenetic reprogramming at fertilization; the persistence of imprint marks into the zygote seems to involve an array of factors (e.g. *Dnmt1*, *PGC7*, *Zfp57*), which act to preserve a small number (i.e. 20–30) of gametic differentially methylated regions that subsequently control the establishment of more-widespread imprinting.^{303,304}

The majority of mammalian imprinted genes are located in clusters,³⁰⁵ indicating that silencing is in some way coordinated and regulatory elements are shared across genes. Regulatory sequences that affect the imprinting of multiple genes are termed imprinting control regions (ICRs). All verified ICRs are characterized by profound differences in DNA methylation between maternally and paternally inherited genomes, which are initiated in the germline.^{303,304} The properties of these differentially methylated regions can vary, with methylation variously associated with the silenced allele or the expressed allele.²⁷² Methylation at the majority of gametic differentially methylated regions is acquired in the female germline,³⁰⁶ perhaps consistent with the suggestion that the majority of silencing (for both maternally and paternally inherited alleles) is under matrilineal control.^{307,308} Notably, both the distribution of methylation and the process of epigenetic reprogramming differ markedly between male and female gametes, indicating possible mechanisms of postzygotic imprint specification.^{303,304}

The origin of genomic imprinting in mammals remains uncertain, as is the benefit of functional haploidy caused by gene silencing (which would be expected to negatively affect fitness). The most-cited hypothesis invokes conflict between the maternal and paternal genomes, which from an evolutionary stand point have conflicting interests during offspring development (also described as the kinship theory of imprinting, as the alleles in question are posited to affect interactions among kin).^{309–311} In polygamous species such as mice, in which paternity of future litters is uncertain, paternal genes within an embryo are most likely to be passed on if the mother invests more resources in the pregnancy (which theoretically results in larger, healthier litters at birth, at a cost to the mother's future potential for reproduction). Thus a parental conflict exists, whereby paternal genes expressed in the embryo and later are expected to increase nutritional demands on the mother, whereas maternal genes are expected to alleviate these excess demands.

In support of a parental conflict theory, many parent-of-origin effects are observed in

traits associated with growth or nutritional behaviour, and a large portion of imprinted genes are reported to have effects in the placenta,^{312,313} brain,³¹⁴ and early embryo.³⁰⁵ An excess of strongly expressed maternal–paternal gene pairs is seen among imprinted genes, also consistent with a hypothesis of genetic conflict.²⁹⁰ Nonetheless, a number of other explanations for the origin of genomic imprinting exist, including defence against ovarian trophoblastic disease,³¹⁵ sexual antagonism,³¹⁶ and maternal–offspring coadaptation.³¹⁷ The contrasting features of these models are reviewed in detail elsewhere.^{318–320} Notably, none of the available hypotheses perfectly explains observed data, and imprinting of different genes could plausibly have varied mechanisms of origin.

Importantly, some level of imprinting seems necessary for normal mammalian development, and aberrant imprinting is associated with several diseases and disorders. For example, the first recognized human imprinting disorders—Prader–Willi and Angelman Syndromes—are reciprocally inherited disorders caused by loss of a region of the long arm of human chromosome 15; maternal inheritance of the deletion is associated with Angelman Syndrome, and paternal inheritance with Prader–Willi Syndrome. This region contains both maternally and paternally imprinted genes, which are thought to be associated with the differential symptoms.²⁵⁶ Other imprinting disorders include Beckwith–Wiedemann Syndrome, Silver–Russell Syndrome, transient neonatal diabetes mellitus type 1, pseudohypoparathyroidism and uniparental disomy 14.^{321–324} Notably, an increased incidence of imprinting disorders has been reported in children conceived through *in vitro* fertilization and other assisted reproductive techniques.³²⁵

Mammalian imprinted genes are typically reported to have roles in embryonic and placental growth and development, and postnatal development, metabolism and nutritional behavior. More-recently, however, diverse roles for imprinted genes have been reported after birth and into adulthood.²⁷⁶ Genomic imprinting is associated with phenotypic differences in reciprocal hybrids and is implicated in hybrid sterility.³²⁰ Abberant expression of imprinted genes is suggested to contribute to many common diseases, in-

cluding obesity, diabetes mellitus, cancer, intrauterine growth restriction and psychiatric disorders. Genomic imprinting thus makes a contribution to complex trait variation in humans, and mapping such traits using additive models might fail to identify correct genetic architecture. One study in mice suggests that one-third of identifiable variance in BMI was lost if imprinting was not accounted for in the mapping model.³²⁶ Thus genomic imprinting—i.e. parent-of-origin-specific gene silencing—is a prominent mechanism resulting in observable parent-of-origin phenomena (i.e. parent-of-origin-specific gene expression, and potentially affecting complex traits).

Other parent-of-origin mechanisms

Besides genomic imprinting, a number of other possible explanations exist for the parent-of-origin phenomenon. First, the different phenotypes of reciprocal heterozygotes might result from strictly maternal effects—“the *causal* influence of the maternal genotype or phenotype on the offspring phenotype”³²⁷—for example, through direct gametic transmission of protein or RNA factors to the offspring,³²⁸ or the provision of the *in utero* environment or maternal care. Paternal effects are also plausible, but assumed to be less influential.

Maternal effects are reported in traits such as offspring viability and size at birth as well as postnatal traits such as offspring aggression.³²⁹ One possible explanation for the persistence of such effects is that maternal genetics, environment or diet affect offspring methylation and other epigenetic marks, which persist into later life.^{330–333} Such maternal effects can be confounded with genomic imprinting, and produce the same patterns of phenotypic variation, as homozygotic parents can produce only one of the possible reciprocal heterozygotes.³³⁴ They can be distinguished by using genetic data from both parents and offspring and restricting analysis to offspring of heterozygous mothers.

Several other mechanisms can also result in parent-of-origin effects,²⁴⁸ although their relative contribution to observed parent-of-origin effects is not well documented. First,

mothers and fathers might have different rates of mutational transmission, leading to a greater likelihood of the offspring inheriting disease-associated mutations from one parent. Examples include differential risk of trinucleotide repeat expansion in male versus female meiosis associated with onset and severity of triplet repeat disorders; differential risk of X chromosome rearrangement; and possible sex-linked differences in *de novo* germline mutation rates. Finally, certain genetic factors are transmitted exclusively from one parent; for example, mitochondrial DNA and the non-pseudoautosomal region of the Y chromosome. These factors are associated with diseases and traits, and might interact with other regions of the genome in a parent-of-origin-specific manner.

Detecting parent-of-origin effects

A number of methods can be used to detect parent-of-origin effects. First, a number of studies have investigated genomic imprinting by examining parent-of-origin-specific gene expression by sequencing the transcriptome, coupled with careful technical control.^{286,290,335,336} This approach can also be used to examine related phenomena including *cis*-regulated expression, X chromosome inactivation and random monoallelic inactivation. Such studies have revealed incomplete imprinting in many genes, pervasive *cis*-regulatory variation, and genomewide parent-of-origin allelic imbalance favoring the paternal allele in brain (at genes associated with CpG islands).^{314,337} Other studies have used DNA sequence analysis³³⁸ to identify potential imprinted genes, which can then be validated experimentally.

The other general approach to identifying parent-of-origin effects does not assume these effects to be synonymous with genomic imprinting. Instead of investigating differential expression, it is possible to do association mapping for binary disease or quantitative traits using genotypes phased by parent of origin, where reciprocal heterozygotes are classified as distinct genetic classes. Such approaches typically use trio,³³⁹ extended family,³⁴⁰ or long-range phasing²⁵⁹ information to assign parental origin to alleles. Increasingly

common is to use surrounding SNPs to incorporate haplotype estimation to improve the estimates of parental origin at a given test SNP.^{259,340,341} These approaches can be extended to identify and control for maternal (and paternal) effects.

Human gene expression

While parent-of-origin effects have been linked to a range of diseases and disorders, the extent to which they contribute to determining variation in complex traits remains to be determined. A 2014 *Cell* article³⁴² suggests that parent-of-origin effects contribute to the majority of 97 traits in mice, possibly mediated by interaction of nonimprinted genes with imprinted loci. This study uses a method that partitions heritability of the investigated traits into portions accounted for by inheritance of alleles from matching versus nonmatching parent of origin.

The method used relies on being able to define the kinship between pairs of individuals and subsequently partition this kinship into two parts representing matching and nonmatching parent of origin. The heritability attributable to each of these two partitions can then be calculated using standard methods. Traits with evidence of parent-of-origin effects will show a high proportion of heritability attributable to matching parent of origin. The methods are detailed in the supplement to the published paper.³⁴²

In the following chapters I describe the extension of this method to investigate the incidence of parent-of-origin effects on human gene expression in peripheral blood. Gene expression is known to be a heritable trait in humans, and is expected to be affected by parent-of-origin mechanisms, including genomic imprinting.

In order to apply this approach, a number of items are needed. First, a genotyped population of individuals (with differing degrees of relatedness) for whom gene expression trait data are available, and whose parents have also been genotyped. This role is fulfilled by the Netherlands Twin Register (NTR) population, which is discussed in detail in

Chapter 6. Second, a method for identifying parent-of-origin effects on gene expression. This proposed method is adapted from the above article³⁴² and its application is presented in Chapter 7.

Chapter 6 | Netherlands Twin Register data

“I don’t need to speak to him to know what he thinks. He is my twin. I don’t wish to see him. I don’t need to. Do you understand the nature of twins? Do you understand the meaning of the word *cleave*?”

—Magid Iqbal, in *White Teeth* (2000)³⁴³

NTR cohort

The Netherlands Twin Register (Nederlands Tweelingen Register; NTR) is a longitudinal research cohort initiated in 1986 and based at Vrije Universiteit, Amsterdam, The Netherlands (www.tweelingenregister.org). The NTR supports two main lines of research: following young twins from birth, primarily to investigate the development of emotions and behavior;³⁴⁴ and following adolescents, adults and their family members to investigate variability in personality, health and disease.³⁴⁵

The adult NTR cohort was initiated with recruitment of young adult twins during 1991–1993, and it remains open to additional eligible participants. Parents, siblings, spouses and offspring of twins are also invited to participate. To date, over 34,000 individuals have taken part in NTR research, and the cohort comprises approximately 25,000 twins and other multiples over 18 years of age. The cohort spans a wide age range, with the oldest male and female participants born in 1918 and 1909, respectively.

NTR participants are asked to complete a survey on health, lifestyle and personality every 2–3 years. The questions have varied from survey to survey. Initially these surveys were sent to families; individual respondents were subsequently targeted. A number of phenotyping studies have been done in parallel with the surveys, including measures of cardiovascular function, depression, cognition, brain function, diabetes and asthma.

During 2004–2008, the NTR project developed a biobank, which comprises phenotypic

measures, a DNA collection, biological measures and samples from NTR participants (twins and family members).³⁴⁶ The biobank recruited 9,530 participants from 3,477 nuclear families (some of whom form larger pedigrees), representing 3,554 male and 5,976 female participants (Table 14). The biobank collected blood (for DNA, RNA, cell line creation, hematology and biomarkers), urine (for metabolomics), and buccal tissue (for DNA). Additional phenotypic information was collected by interview. Gene expression data have subsequently been generated from the biobank RNA samples.³⁴⁷

Table 14: **NTR Biobank cohort—2010 update**

Role	Number of participants
Twin/multiple	4,293
Sibling	2,070
Father	1,209
Mother	1,587
Spouse	371
Complete twin pairs	1,663
Monozygotic male	273
Monozygotic female	622
Dizygotic male	144
Dizygotic female	309
Dizygotic opposite sex	299
Unknown zygosity	16
Triplet trios	7

By 2013, DNA samples have been collected for 15,437 (45%) of the NTR research participants.³⁴⁵ These samples have been collected through various efforts, including cardiovascular research projects, the NTR-biobank projects and research into anxious depression and nicotine addiction. DNA measures have been integrated into the assessment of zygosity, which interestingly is not completely concordant with questionnaire reports (i.e. self-reported monozygous versus dizygous status is not completely concordant with DNA).

Initial genotyping in the NTR was done using microsatellite markers. Subsequently, a number of rounds of genotyping were done using a range of SNP microarrays (Table 5). Importantly, a number of parents and other family members have been genotyped alongside the index twins. Genotype data obtained from the various SNP platforms are stored as a single merged dataset.

Table 15: **NTR genotyping platforms—2013 update**

Genotype platform	Number of participants
Microsatellite	2,694
Affymetrix-Perlegen 5.0	1,982
Illumina 370	166
Affymetrix 5.0	24
Illumina 660	1,340
Illumina Omni Express 1M	439
Affymetrix 6.0	5,056
Affymetrix 6.0	1,285 (not yet added to main data)
Unknown	1,000 (not yet added to main data)

NTR research

The NTR is an established research resource, and has formed the basis for over 100 peer-reviewed studies. The register was originally created to follow the development of young twins from birth.³⁴⁸ The adult NTR population was recruited from 1991 as a population-based longitudinal twin family study of lifestyle, personality and psychopathology.³⁴⁹ It has since developed into a resource that offers extended and complex family pedigrees for genetic epidemiology and heritability studies.³⁵⁰

Longitudinal questionnaire studies in the NTR have revealed the genetic contributions and causes of individual variation in many traits, including demographic characteristics, behaviour, personality, physical and mental health, and cognition.³⁴⁵ Experimental and laboratory studies of cognition and cardiovascular health have also been done in subsets

of NTR individuals.

The rich phenotype data generated by these projects have led to the use of the NTR for linkage and genome-wide association studies. To date, over 15,000 participants have been genotyped and the NTR has contributed to a number of international gene-finding consortia and the Genome of the Netherlands project.³⁵¹ The NTR data are also crosslinked to other databases within the Netherlands to facilitate research.³⁵⁰

Over time, the NTR has grown into an important resource that enables a broad range of research question to be addressed. Its longitudinally followed population has contributed to epidemiological studies that track behavioural changes within individuals and between generations. Similarly, the NTR has evolved into a rich source of genetic data among family trios. The NTR biobank population is a control population in the Genetic Association Information Network study of major depressive disorder.³⁵² Twinning is ubiquitous among all socioeconomic strata, and the NTR cohort is therefore representative of the broader Dutch population. The NTR exemplifies the value of long-term, maintained research resources.

A 2014 *Nature Genetics* paper used the adult NTR cohort to quantify the heritability of human gene expression.³⁴⁷ The authors used a classical twin design in which resemblance in MZ twin pairs is compared to that in DZ twin pairs (690 and 618 complete pairs, respectively, plus 136 individual twins). They use expression data from the NTR biobank to calculate the heritability of steady state transcription in peripheral blood for 43,628 probe sets from 18,392 genes.

The researchers conservatively highlight 777 genes with significant heritability of expression, and suggest that the frequency of expressed genes with heritability >0.3 is 7.9% (corresponding to 1,453 genes in their sample). This work confirms that gene expression is a heritable trait in humans and highlights the possibility of exploring this trait in the NTR population.

The NTR, therefore, provides a resource through which the heritability of human gene expression can be calculated, and thus the parent of origin effects on human gene expression can be estimated by partitioning of heritability. The following properties of the cohort are important:

1. Genotyped offspring with good covariate and pedigree data—for genetic trait analysis;
2. Genotyped parents (especially cases in which both parents are genotyped)—for phasing of offspring data by parent of origin;
3. Availability of a suitable phenotype for trait analysis, in this case genome-wide expression data;
4. A mixture of related and unrelated individuals (i.e. family trios not part of a larger pedigree) to allow comparison of true parent of origin effect versus maternal or shared environmental effects.

As per the project brief, I specifically requested the following information from the NTR:

1. Genotype data for those families in which both parents have been genotyped and for which gene expression data is available (i.e. genotypes from both parents and offspring, and/or phased imputed genotypes if these are available);
2. Gene expression data for these individuals;
3. Covariates and meta-information such as pedigree, sex, anonymised family ID, age, tissue used for gene expression etc., and any other covariates that NTR recommend for genetic analysis.

NTR genotype data

In total, I obtained pedigree and genotype information from the NTR for 2,762 individuals, comprising 638 family pedigrees. Among these data, 1,066 individuals from 448 families

have two parents who have been genotyped, and are therefore suitable for phasing by parent of origin. These individuals comprise 382 monozygotic twins and 684 other offspring (a mix of dizygotic twins and other siblings). The number of offspring per family ranges from one to eight, with the vast majority of families having two or three offspring (Figure 52). Monozygotic twins are labelled and matched to a co-twin ID in the pedigree, but dizygotic twins are not distinguished from non-twin siblings.

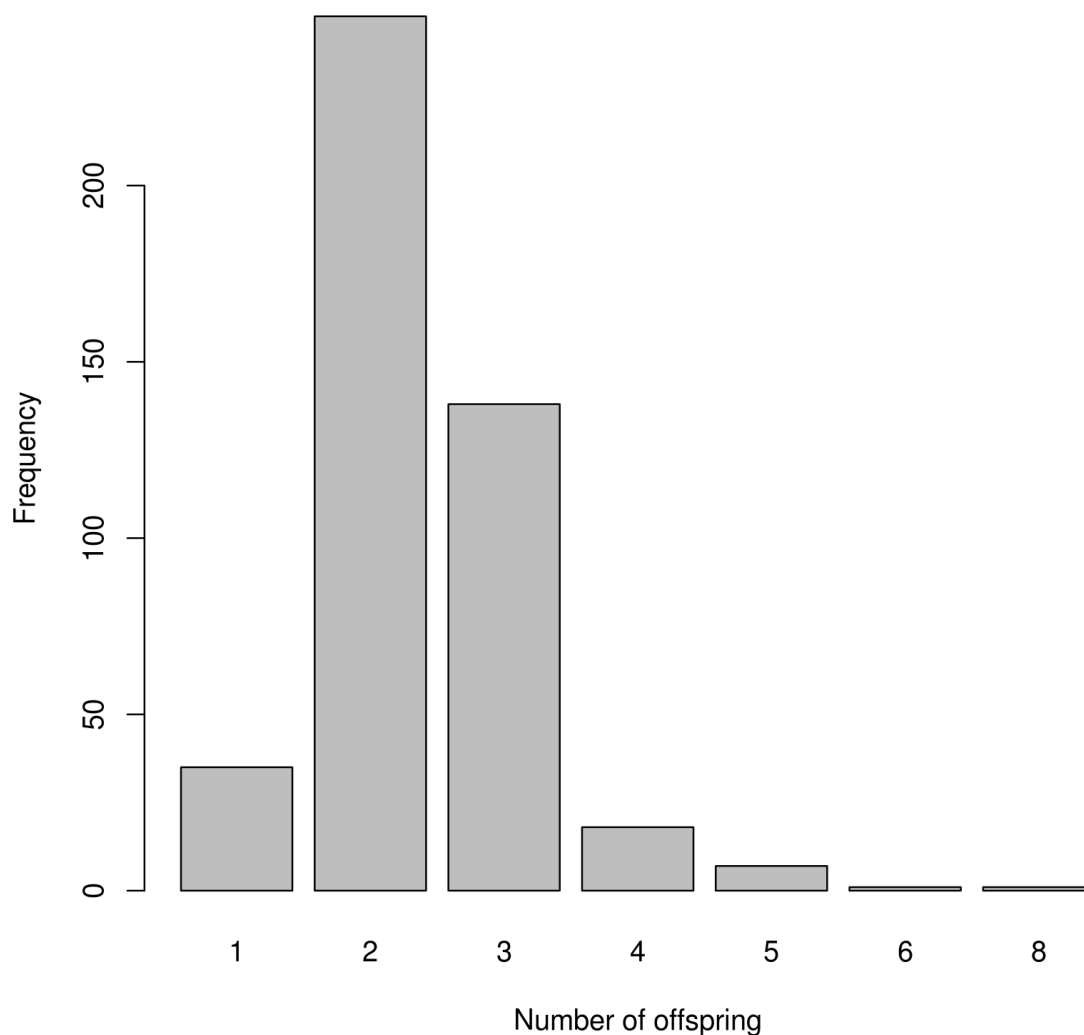


Figure 52: NTR family size. Most families in the dataset have two or three offspring. A few families are notably larger.

I use parental identities to check for non-nuclear families and monozygotic triplets. The population contains no monozygotic triplets; eight family IDs are assigned to non-nuclear families—two with grandparent relationships and six with two sets of parents and offspring, possibly representing sibling or cousin relationships at the parent level (Figure 53). In the context of this study, these families will contain individuals who share alleles, but who aren’t identified as siblings.

As anticipated, the NTR individuals had been genotyped using a range of platforms, and for multiple projects (Tables 16–18). The number of genotyped SNPs per individual ranges from 289,786 to 1,139,661, with a mean of 618,411 SNPs. For each platform the individual SNP markers have been lifted over to build 37 (hg19) of the human reference genome and aligned with the 1000 Genomes phase I integrated variant set. DNA has been extracted from whole blood or buccal samples. Further details of the quality control process are appended at the end of this chapter.

Table 16: **NTR genotyping projects and platforms**

Cohort	Array	<i>n</i>
NTR1	Perlegen-Affymetrix 5.0	614
NTR2	Illumina 660	360
NTR3	Illumina 1M	194
GO1	Affymetrix 6.0	1,186
GO2	Affymetrix 6.0	394
SDE1	Affymetrix 6.0	308
ADHD	Affymetrix 6.0	2
GEUT	Illumina 370	176
M2000	Affymetrix 6.0	11
SCHZ	Illumina 660	36

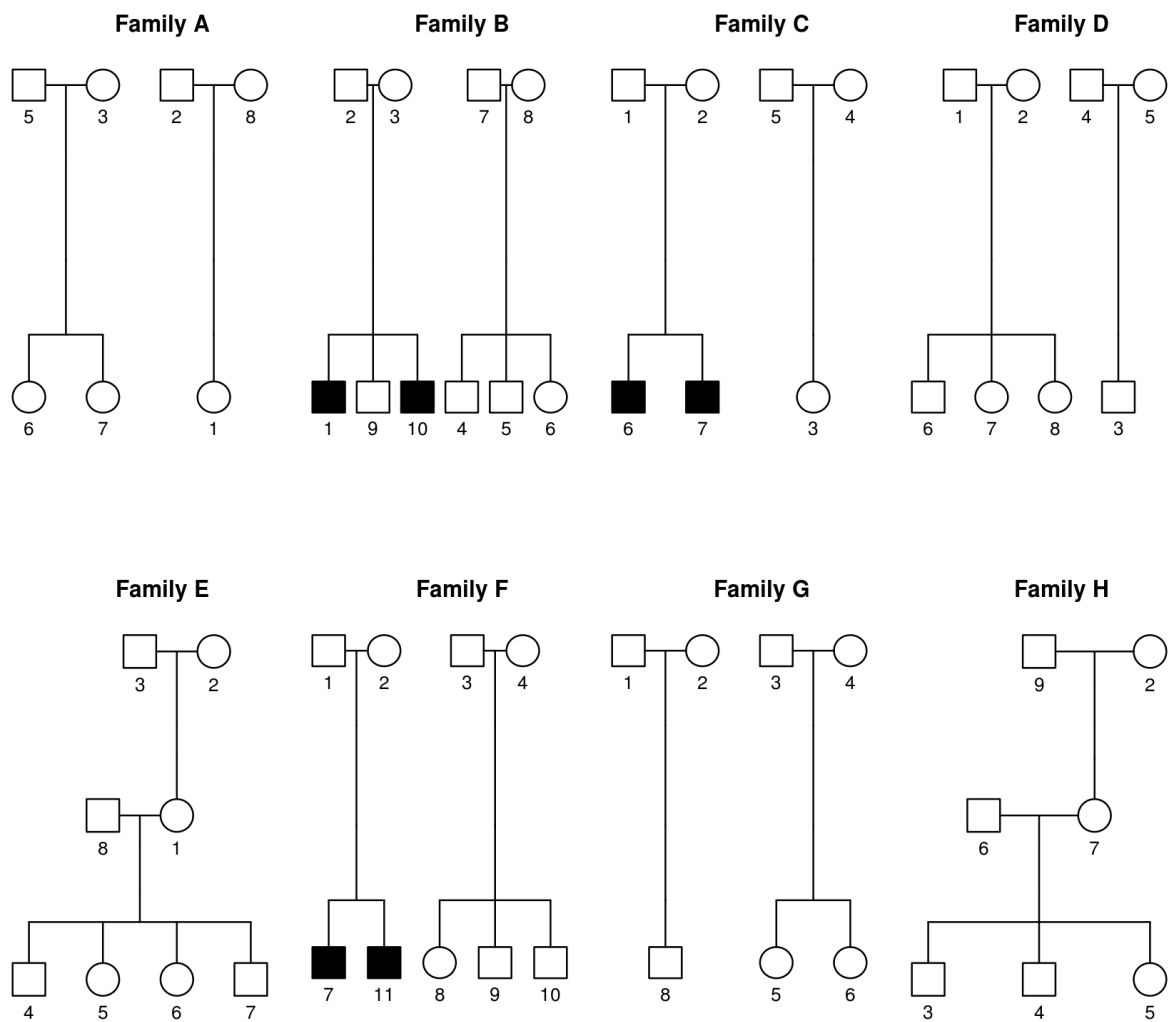


Figure 53: Unusual family pedigrees. Eight families demonstrate unexpected relationships, two with grandparent relationships and six with two sets of parents and offspring, possibly representing sibling or cousin relationships at the parent level. Monozygotic twins are shaded.

Array	<i>n</i>
Perlegen–Affymetrix Human SNP Array 5.0	614
Illumina Human660W-Quad	396
Illumina OmniExpress 1M	194
Affymetrix Human SNP Array 6.0	1,901
Illumina HumanCNV370-Duo	176

Table 18: **Individuals genotyped on multiple platforms**

Platforms used	<i>n</i>
1	2,323
2	363
3	72
4	4

I received genotype data from the NTR in binary PLINK (.bed, .bim, .fam) format. I used the `read.plink()` function from the `snpStats` Bioconductor R package to examine the binary data.³⁵³ The whole-genome dataset is too big for R’s memory capability³⁵⁴ so the data were split using PLINK,³⁵⁵ and each autosome has been examined and summarized individually as Class ‘`SnpMatrix`’ objects (Table 19).

Of the 1,416,992 SNPs covered by the five arrays, only 2,088 (0.00147%) were evaluated in all individuals. 31,788 (0.0224%) were assessed in >99% of individuals. In addition, individuals within a family are not always typed on the same platform. Thus there is substantial missingness in the data, which complicates subsequent analysis.

Table 19: Genotyping summary

Chromosome	SNPs	Median_per_indiv	Min_per_indiv	Max_per_indiv	Median_per_SNP	Min_per_SNP	Max_per_SNP
chr1	112,156	0.438	0.196	0.818	0.438	0.001	1
chr2	117,683	0.436	0.201	0.798	0.436	0.001	1
chr3	97,066	0.442	0.209	0.804	0.442	0.001	1
chr4	89,559	0.436	0.201	0.8	0.436	0.001	1
chr5	91,710	0.441	0.199	0.801	0.441	0.001	1
chr6	99,351	0.416	0.199	0.804	0.416	0.001	1
chr7	78,534	0.435	0.2	0.8	0.435	0.001	1
chr8	79,529	0.438	0.215	0.801	0.438	0.001	1
chr9	67,862	0.44	0.219	0.804	0.44	0.001	1
chr10	76,730	0.443	0.191	0.811	0.443	0.001	1
chr11	71,902	0.438	0.193	0.813	0.438	0.001	1
chr12	70,136	0.436	0.2	0.808	0.436	0.001	1
chr13	56,196	0.435	0.196	0.798	0.435	0.001	1
chr14	46,384	0.438	0.2	0.801	0.438	0.001	1
chr15	41,359	0.441	0.198	0.804	0.441	0.001	1
chr16	42,787	0.446	0.196	0.812	0.446	0.001	1
chr17	34,857	0.431	0.223	0.811	0.431	0.001	1
chr18	43,127	0.441	0.229	0.805	0.441	0.001	1
chr19	22,790	0.402	0.224	0.831	0.402	0.001	1
chr20	36,619	0.443	0.202	0.815	0.443	0.001	1
chr21	20,646	0.442	0.251	0.806	0.442	0.001	1
chr22	20,009	0.425	0.234	0.803	0.425	0.001	1

Expression data

I obtained peripheral blood expression data from 1,535 individuals in the NTR population, provided as a .csv file with two header rows. 889 of these individuals have been genotyped *and* have both parents with genotypes, suitable for phasing by parent of origin. Expression levels were measured using the Affymetrix u219 chip, and reported for 44,241 probes. The expression data were generated from samples collected as part of the NTR biobank project, and have been subject to heritability estimation.³⁴⁷ Details of the expression pipeline are appended to this chapter.

The Affymetrix u219 array contains 530,467 probes for 49,293 transcripts (~11 probes per transcript), plus control probes (labeled ‘AFFX’).³⁵⁶ Arrays were hybridised using the GeneTitan system. Probes were mapped to the human genome version 19 (build 37) using Bowtie³⁵⁷ and annotated against GENCODE version 14;³⁵⁸ those probes with uncertain location or within polymorphic SNPs were removed. The NTR mapping might, therefore, differ from the manufacturer’s mapping and I will consider this possibility in our subsequent analyses. The NTR mappings are saved for future reference.

The distributions and mean values of the expression data vary by probe. The distributions for the first 100 expression traits are shown in Figure 54, with whiskers depicting data extremes. In general, probes are consistently expressed across individuals, although between-individual variation in expression is also evident. Expression data are typically not normally distributed, and do not show consistent distributions (Figure 55).

Covariates

I obtained data on 17 covariates for the 1,535 individuals with expression data. Covariate data is likely to have been obtained via the NTR biobank project. The covariate file is illustrated in Tables 20–22. 1,351 individuals have complete covariate records, 762 of whom have available expression data and both parents genotyped.

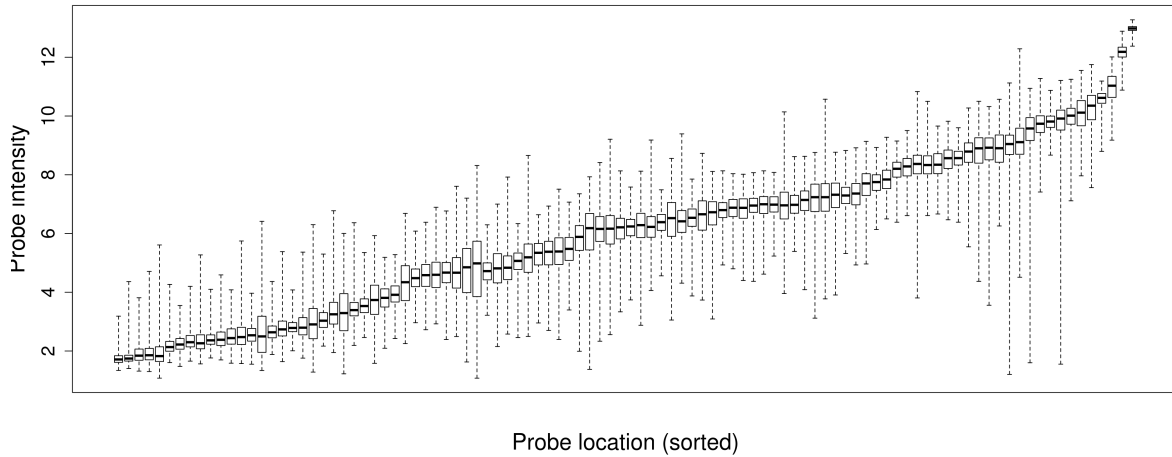


Figure 54: Variation in measured values of 100 expression traits. The traits are ordered by increasing mean probe intensity. There is substantial variation in the mean probe intensities, and in the range and skew of the intensities.

The NTR gene expression study³⁴⁷ used these data in its heritability analysis, finding “significant associations for plate, hybridization well position, age at blood sampling, sex, time interval between extraction and hybridization steps, total white and red blood cell counts, hematocrit and the top five expression principal components.” The covariate data that I received largely match these significant covariates, with minor differences.

The supplement to the paper³⁵⁶ lists the covariates that were included in the heritability and eQTL analyses, and their relative order of importance. Covariates were selected by type III sum of squares testing, which compares a full model containing all covariates with a partial model excluding the given covariate. This list does not agree exactly with that presented in the published paper methods, but is potentially more precise.

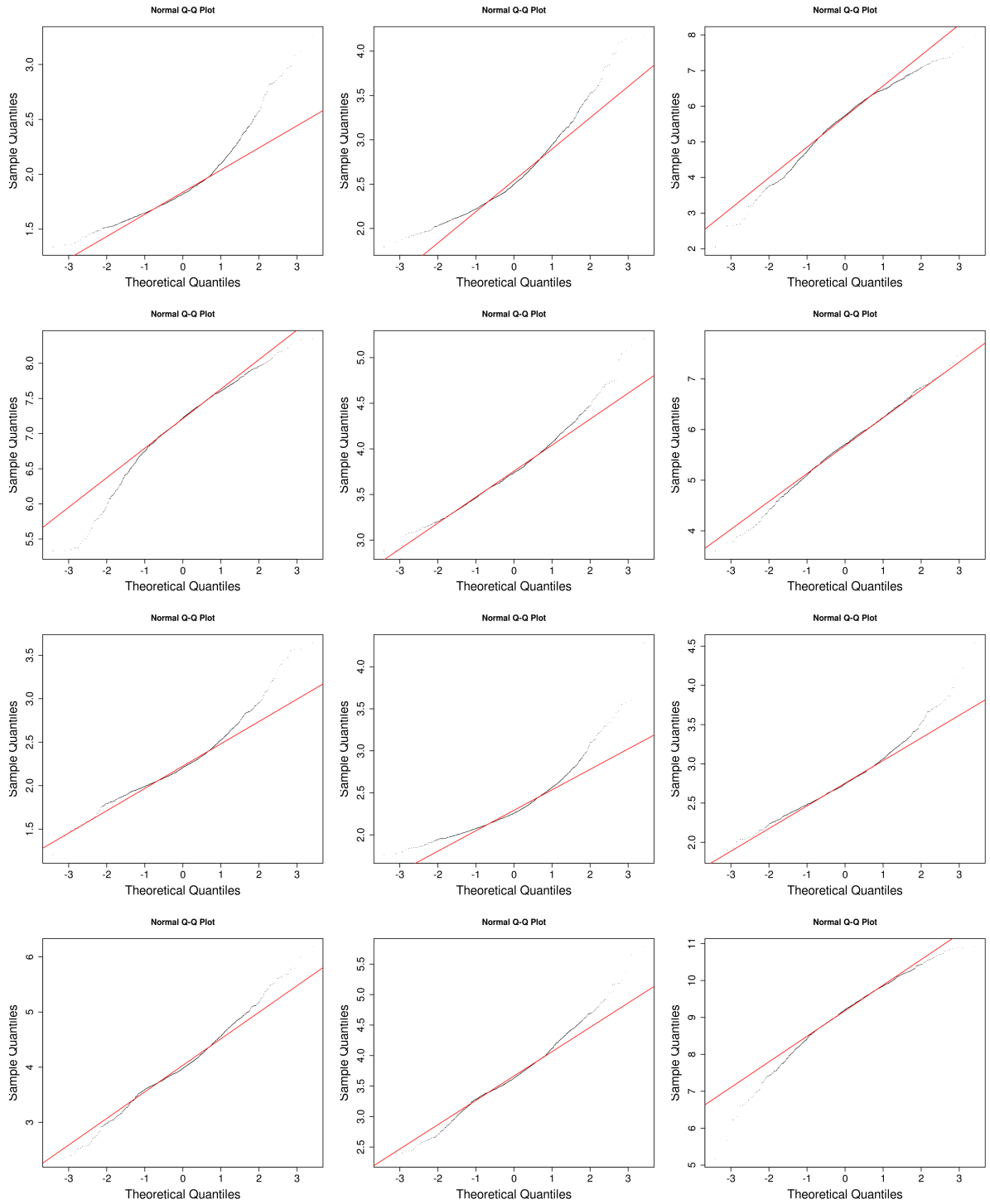


Figure 55: Normal Q–Q plots for selected expression traits. The data are not typically normally distributed and do not show consistent distributions.

Table 20: Covariate data

sex	age	neutn	lympn	monon	eosn	bason	rbc	plate	well	D	bmi	smoking	ndays	ndays_ext_amp	month	hour
FEMALE	17	NA	NA	NA	NA	NA	4.6	plate10	A10	1.316	19.5	1	1,900	140	3	38,040
FEMALE	43	3	3	0.6	0.3	0	4.22	plate10	B09	0.872	21.7	1	700	140	6	31,800
MALE	34	3.2	2.1	0.6	0.1	0	5.12	plate10	B10	1.288	30	1	1,368	140	8	33,180
MALE	34	3.5	3.1	0.6	0.2	0	5.14	plate10	B11	0.9	26.1	1	1,486	140	5	26,400
MALE	31	4.9	2.7	0.6	0.1	0	5.11	plate10	B12	0.537	22.5	2	1,451	140	6	35,400
FEMALE	36	3	1.5	0.6	0.2	0	4.19	plate10	C09	0.98	22.8	1	990	140	9	28,800

Colname	Key
sex	Sex
age	Age at sampling
neutn	Neutrophil count
lympn	Lymphocyte count
monon	Monocyte count
eosn	Eosinophil count
bason	Basophil count
rbc	Red blood cell count
plate	Array plate
well	Location on plate
D	Measure of sample correlation (see pipeline)
bmi	BMI
smoking	Current smoking status (1 = nonsmoking)
ndays	Days between blood sampling and RNA extraction
ndays_ext_amp	Days between RNA extraction and RNA amplification
month	Month of blood sampling
hour	Hour of blood sampling (seconds since midnight)

Table 22: NTR expression covariates

Covariate	Type	df	Received
Affymetrix GeneTitan plate	Factor	41	Yes
<i>D</i> statistic	Numeric	1	Yes
Well location on plate (as 24 sets of well quadruples)	Factor	23	Yes (but not as well quadruple)
Days from blood draw to RNA hybridisation	Numeric	1	No (draw to extraction instead)
Days from RNA amplification to fragmentation	Numeric	1	No
Calendar month of blood draw	Factor	11	Yes
Monocyte cell count	Numeric	1	Yes
Interaction of sex*monocyte count	Numeric	1	Yes
Interaction of sex*lymphocyte count	Numeric	1	Yes
Interaction of sex*eosinophil count	Numeric	1	Yes
Interaction of sex*neutrophil count	Numeric	1	Yes
Interaction of sex*basophil count	Numeric	1	Yes
Lymphocyte count	Numeric	1	Yes
Eosinophil count	Numeric	1	Yes
Basophil count	Numeric	1	Yes
Neutrophil count	Numeric	1	Yes
BMI	Numeric	1	Yes
Days from RNA extraction to amplification	Numeric	1	Yes
Hour of day of blood draw	Numeric	1	Yes
Age at blood draw	Numeric	1	Yes
Smoking status	Factor	1	Yes
Sex	Factor	1	Yes
Interaction of sex*RBC count	Numeric	1	Yes
RBC count	Numeric	1	Yes
Principal components 1–3 (estimated genotype data)	Numeric	3 x 1	Yes (but in pedigree files; eQTL analyses only)
109 SNPs associated with blood counts	Numeric	109 x 1	No
Principal components 1–7 (estimated on expression residuals)	Numeric	7 x 1	No

The supplementary note identifies three classes of covariate:

1. Covariates related to technical variation;
2. Clinical covariates that are subject-specific;
3. Covariates related to blood counts, which if not properly accounted for might produce spurious ‘eQTL’ relationships.

The technical covariates include the hybridization plate, well position and sample processing dates. In our data, hybridization plate has 50 unique levels, compared to 41 degrees of freedom listed in the NTR supplement; this number drops to 47 when only the individuals with complete covariate data are considered. Well position is provided with 96 unique levels, although this has been collapsed to 24 sets of 4 in the NTR supplement. Processing dates are provided for blood sampling to RNA extraction and extraction to amplification, but not for amplification to fragmentation or overall draw to hybridisation.

The subject-specific clinical covariates include sex, age at blood draw, BMI, blood counts, and month and time of day of blood draw. All these data are available, although some measurements are missing for white cell counts ($n = 127$) and red blood cell counts ($n = 14$; overlapping with missing white cell counts). The NTR paper report complete white blood cell counts for ~93% of the samples.³⁵⁶ By contrast, our rate of missing data is 11.9% for the full covariate set and 14.3% for those individuals with two parents and expression data, perhaps reflecting the larger sample in the previous NTR study.

The NTR project imputed missing total white and red blood cell counts by regressing on BMI and BMI², although the prediction accuracy was low (multiple R² ~0.03 for each; done separately by zygosity status). It does not report imputing individual white cell counts. As omitting the individuals with missing data would notably reduce the size of our dataset, I investigated the possibility of imputing the missing values. Plotting the distributions show that white blood cell counts are typically right skew, with a small proportion of relatively extreme values. Red blood cell counts seem roughly normally

distributed (Figure 56). Of note, blood cell subtypes contain many cell subpopulations and the measurements here represent relatively broad classes of cell subtypes.

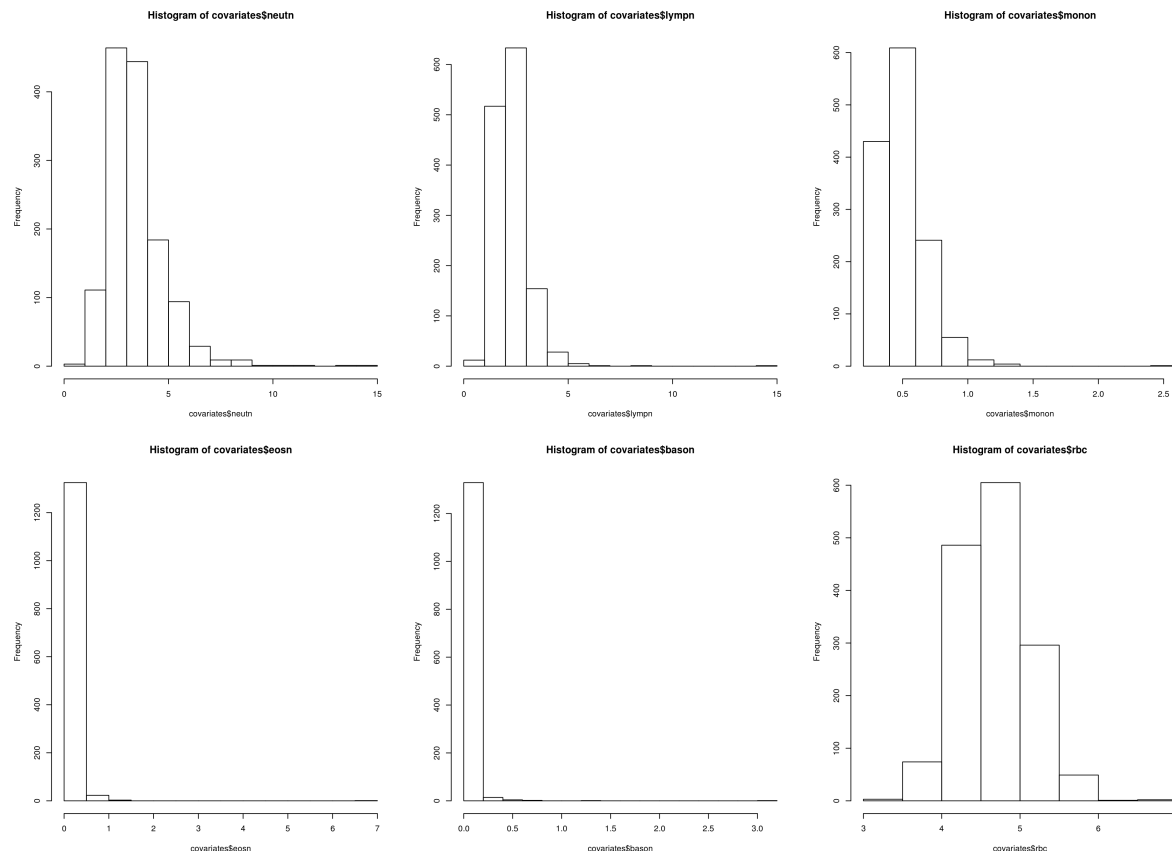


Figure 56: Blood count covariates. White blood cell counts are typically right skew, with a small proportion of relatively extreme values. Red blood cell counts seem roughly normally distributed

For imputation, a number of options are available.³⁵⁹ Most simply, the overall mean or the observed value from another similar case can be assigned to each missing value. More-complex methods include conditional mean imputation, which involves imputing missing values based on regression models, and multiple imputation (fully conditional specification), which involves specifying univariate models for each missing variable conditional on all other variables. I explored these more complex approaches.

Among the available clinical covariates, possible predictors of blood cell counts are sex, age, BMI and smoking status. I first regressed each blood cell type count on these

covariates and used the model predicted values to impute the missing blood cell count values. The predicted values match the observed values reasonably well, but as expected do not predict extreme outliers (Figure 57). I compared these predicted values to those generated by a multiple imputation method that uses a bootstrapping approach (Amelia II^{360,361}). Multiple imputation produces a broader range of values, more-consistent with the distributions of the observed measurements (Figures 58, 59). Masking the observed values individually, using the `overimpute()` function shows reasonable performance for the imputation (not shown). I use the multiple imputation approach for all subsequent analyses.

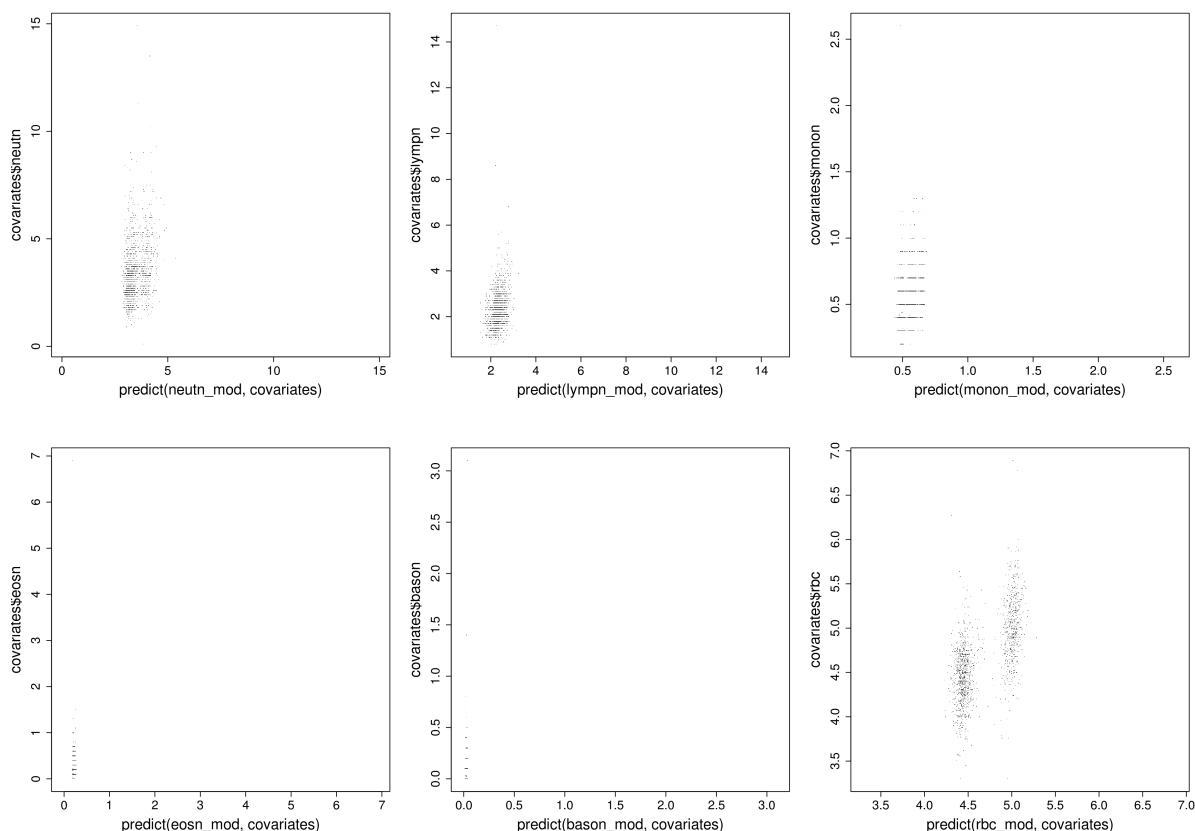


Figure 57: Predicted versus observed values for blood cell counts. The predicted values are based on regression of blood cell counts on sex, age, BMI and smoking status. The distributions of predicted results are typically narrower than the distribution of observed values.

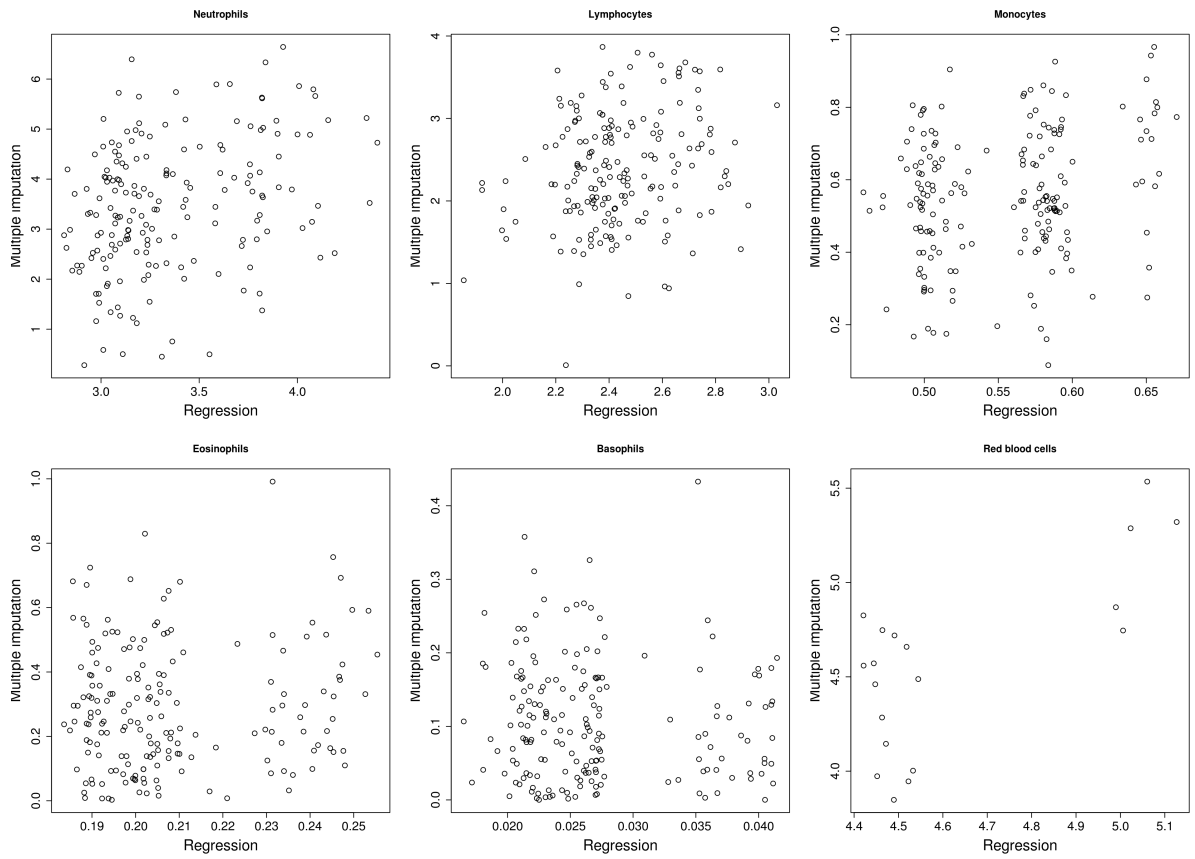


Figure 58: Regression versus multiple imputation methods. Multiple imputation methods produce similar estimates to regression, but with a wider distribution of predicted values.

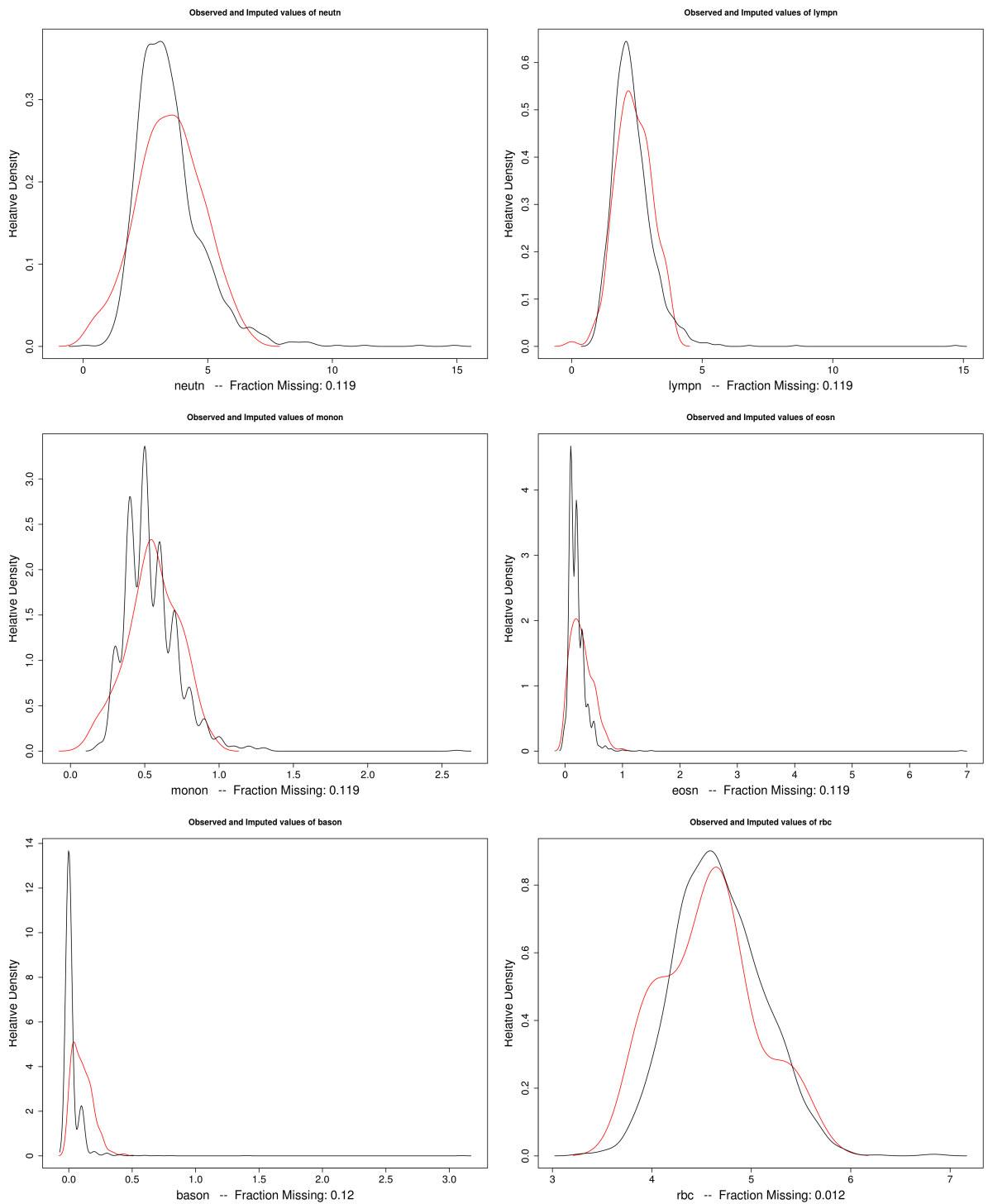


Figure 59: Observed and imputed (red) values on multiple imputation. The multiple imputation approach produces imputed values that match well the distribution of the observed results.

The NTR heritability study also identified 109 potentially confounding SNPs associated with blood cell count phenotypes (through GWAS and the NHGRI GWAS catalog). These SNPs are used as covariates for analysis, but initially none was found in our genotype data. On closer examination, the SNPs are mislabeled in the study’s supplementary information,³⁵⁶ with both chromosome and position wrongly given. After correcting the labeling, 70 of the SNPs were found in our genotype data (generally incomplete with much variation in missingness). As the genotype status of these 109 SNPs is largely missing from our dataset, I omit them from the analysis at this stage. If data on these SNPs existed they might also improve prediction ability for imputation of missing blood cell subtypes.

Finally, a calculated D -statistic was reported in the list of covariates. This statistic reflects the correlation of each array with all other arrays, and was used to identify and remove poor quality arrays. D is correlated with many transcripts, and is therefore included as a covariate. Once I had established the appropriate covariates, I reduced the expression data to residuals to remove the effect of the covariates on the expression phenotypes.

Following the approach of the NTR paper I calculated expression principle components on the residualised data and re-residualised the data to remove the effect of the first five of these components (described as analogous to surrogate variables³⁶²). None of these components explains a substantial amount of variance (all $<10\%$), although the components do seem to capture some variation in the data (cumulatively 14.9% of variance). Finally I again followed the approach of the NTR paper and inverse quantile normal transformed each expression trait to limit the possibility of identifying spurious eQTLs caused by expression outliers. Before transformation I first checked for possible bimodality of the variables (not shown). These normalised values are used for all subsequent analyses.

An example of the distribution of an expression variable during data transformation is shown in Figure 60. The final expression object is saved as a list comprising raw and transformed expression data and covariates (Table 23). This expression object is used for

the heritability analysis described in Chapter 7.

Table 23: **Final expression data**

Name	Description
exp	Raw expression data
exp_resids	Residualised expression data
exp_resids_nosex	Residualised expression data without sex
exp_resids_pca	PCA residualised expression data
exp_resids_pca_norm	Normalised PCA expression data
covariates	Raw covariate data
covariates_imp	Imputed covariate data with PCs
NTR_annotes	NTR probe annotations

Figure 14 | Expression data transformation example

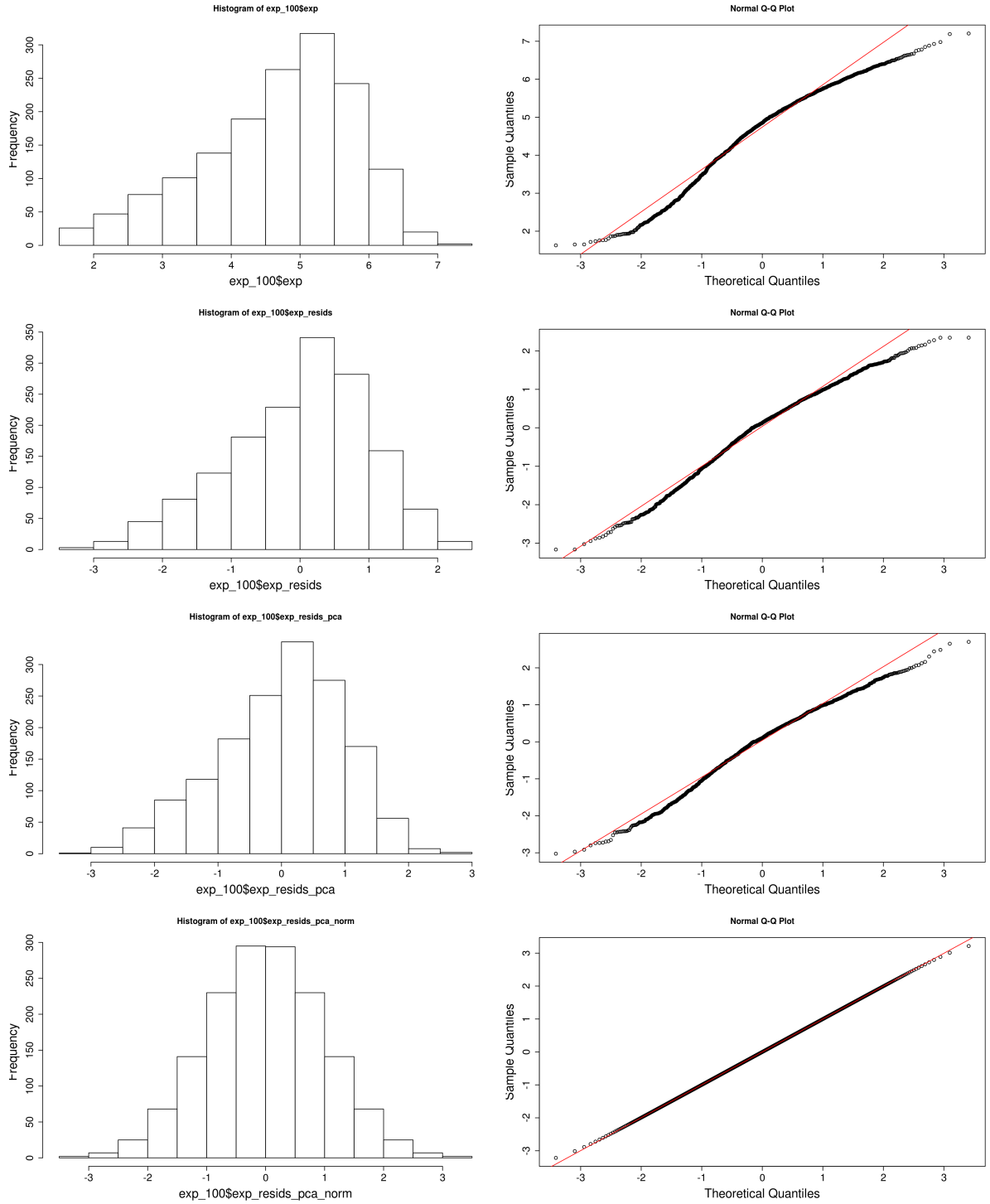


Figure 60: Sequential transformation steps, illustrated by one expression trait. The panels step through the distribution of trait values in (a) the raw data, (b) the residualized data, (c) the residualized data after accounting for principal components, and (d) the final normalized data. The final data is approximately normally distributed. Other expression traits show the same patterns across the transformation steps.

Chapter 7 | Parent of origin effects on gene expression

A little more than kin, and less than kind

—*Hamlet* (1.2.65), *aside*

Phasing NTR genotype data

In order to assess parent of origin effects on gene expression, I first needed to phase the genotypes with respect to parent of origin. The NTR twin data used in this analysis are discussed in detail in Chapter 6. To reiterate, the data used here comprises 889 genotyped individuals with expression data, each of whom have two parents also genotyped and are therefore suitable for phasing by parent of origin. This population contains 337 monozygotic twins (not all of whom have a corresponding twin pair), and 552 siblings (including dizygotic twins). 531 individuals are female and 358 male.

Phasing (also known as haplotype estimation) involves sorting genotype data to identify which alleles co-locate on a chromosome. Sequencing and array data are typically generated unphased; computational methods can determine which parental chromosome or haplotype a given allele is found on. Identification of haplotype phase has a number of uses in research, including imputation of untyped variation and modeling *cis*-regulated gene expression.³⁶³ For the current investigation, phasing by parental origin is essential to correct identification of parent of origin effects.

In general, computational haplotype phasing works by identifying common haplotypes that can explain observed patterns in genotype data. These common haplotypes can be informed by the genotypes of individuals in a sample, publicly available haplotype frequencies and reference sets (e.g. 1000 Genomes), and known pedigree information. A number of approaches to haplotype phase estimation have been used, with coalescent-based and hidden Markov model methods the state of the art at the time of the Browning and

Browning review (e.g. PHASE, BEAGLE, MACH and IMPUTE2).^{363–367} Other phasing programs have since been developed (e.g. SHAPEIT, HAPI-UR).^{149,368} All methods rely on the insight that a small number of distinct common haplotypes segregate in a population. In related individuals with known pedigrees, Mendelian inheritance constrains and informs haplotype phase estimation at many loci. For example, a given individual will have inherited a distinct chromosomal haplotype from each parent, allowing for a low number of recombination events. Such shared parent–offspring haplotypes also provide information as to which parent (i.e. mother or father) a given haplotype has been inherited from. Thus, when genotypes are known for both parents, parent of origin can be relatively assigned to each chromosome. More-generally, segments of long-range identity by descent (IBD) among unrelated individuals can also be used to inform phasing.^{259,369} Incorporating known relatedness and IBD considerations can improve the accuracy of phasing, by constraining the number of possible common haplotypes.

In order to investigate parent of origin effects on gene expression in the NTR population I first need to obtain accurate estimates of haplotype phase for each individual. As noted previously, I obtained genotype data for 2,762 individuals, 889 of whom had both parents available within the genotype dataset *and* available gene expression data. These individuals had been genotyped on a variety of SNP arrays, and as a result our genotype dataset contained substantial missing data. Members of a family were not necessarily genotyped on the same SNP platform. Thus, our phasing process has to handle this missing data.

I use SHAPEIT^{368,370} to phase our haplotypes. SHAPEIT is a hidden Markov model haplotype estimation tool developed in collaboration by groups at the Conservatoire National des Arts et Métiers, Paris, France and Department of Statistics, University of Oxford, UK. SHAPEIT uses a graph structure to represent all possible haplotypes consistent with an individual’s genotype data; a pair of paths through this graph represent a pair of possible haplotypes. It has proved effective for whole-chromosome phasing

projects,^{368,371} and for phasing individuals with all levels of relatedness.³⁷² Our initial pilot tests were done with version v2.r727 and the final haplotype sets were made using the last major updated version v2.r788 (which was current at 14 January 2014; now updated to v2.r837 as of 17 August 2015).

The primary function of SHAPEIT is to infer haplotypes from SNP genotypes; a genetic map specifies the genetic position of the sample SNPs in centimorgans and allows estimation of recombination rates between these SNPs. SHAPEIT only works on individual chromosomes so I split the NTR dataset accordingly using PLINK.³⁵⁵

Given the level of missingness in our data I had no guarantee that the phasing process would work as intended. However, several factors indicated that SHAPEIT could handle these data. First, SHAPEIT is recommended as a tool for pre-phasing before using IMPUTE2 to produce a fully imputed sequence.³⁷³ This suggests that it can accurately phase haplotypes in the presence of a relatively small number of informative SNPs. Second, the NTR population is mainly European in origin and the relevant haplotypes are likely to have good coverage on the 1000 Genomes and Genome of the Netherlands reference panels. Third, I obtained a larger sample of individuals than the 889 trios to be phased by parent of origin ($n = 2,762$). This large sample provides additional common haplotype coverage for haplotype estimation. Finally, the presence of good pedigree information increases the performance of phasing, as Mendelian constraints limit the possible common haplotypes. The process is analogous to creation of a phased scaffold, which is expected to be highly accurate in the presence of family trios and duos.³⁷⁴ I evaluate the program's performance below.

I phased the data using a reference panel generated from the 1000 Genomes Project, which matches that used in the NTR genotype data pipeline (phase 1, v3; current as of 14/01/14³⁷⁵; since updated to phase 3 in October 2014 by integration of sequence data^{374,376}). As per the suggested protocol, I excluded sample SNPs that were not found in the reference panel (around 20–100 per chromosome). I used the standard

SHAPEIT model and algorithm parameters, with the window size parameter set to 5Mb as recommended for populations with large amounts of IBD sharing. Phasing has the additional effect of imputing missing SNP calls in the dataset, based on the haplotype estimates. Full imputation can follow later, using a combined reference panel incorporating 1000 Genomes^{23,377–379} and Genome of The Netherlands^{380–382} data. Prephasing with SHAPEIT allows imputation to be done rapidly, and enables repeat imputation with updated reference panels.

I initially used an in-development version of the software DuoHMM to incorporate pedigree information and to phase each of the returned outputs with respect to its relevant family members (provided by J O’Connell). DuoHMM “infers the inheritance pattern in parent–child duos, detects genotyping errors and can correct switch errors.” This functionality was subsequently incorporated into SHAPEIT (since version v2.r788)^{372,383} and I have since re-done the phasing using the `–duohmm` flag within SHAPEIT.

The SHAPEIT output is saved in the haps/sample file format (example below). Each row of the .haps file represents one called SNP. The first five columns show chromosome, SNP ID, SNP location, minor allele, and major allele. Each subsequent pair of columns (e.g. columns 6 and 7) then show the two haplotypes for each of the 2,762 individuals in the full dataset. 0 denotes the minor allele and 1 denotes the major allele. The paternally inherited haplotype is shown first, and the maternally inherited haplotype second. The .haps files each have a matching .sample file that contains IDs for the individuals in the .haps file (not shown). I have also output haplotype graphs, which can be used to identify uncertainty in phasing for subsequent analysis.

##		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
##	22:16055171	22	22:16055171	16055171	A	G	1	1	1	1	1	1
##	22:16058758	22	22:16058758	16058758	A	C	1	1	1	1	0	1
##	22:16415579	22	22:16415579	16415579	T	G	1	1	1	0	1	0

```
## 22:16425814 22 22:16425814 16425814 T C 1 1 0 1 1 1
## 22:16489239 22 22:16489239 16489239 G A 1 1 0 1 1 1
## 22:16491206 22 22:16491206 16491206 T C 1 1 0 1 1 1
## 22:16504399 22 22:16504399 16504399 C T 1 1 1 1 1 1
## 22:16642256 22 22:16642256 16642256 T C 1 1 1 1 1 1
```

As expected SHAPEIT recognises substantial missingness in the input data. It outputs haplotype estimates with the missing SNP calls imputed (based on the most likely calls from the haplotype graphs). I use a standalone implementation of duoHMM (version 0.1.4) to check these haplotype estimates on one chromosome (chromosome 22), using the function for detecting subtle genotype errors (defined as those that would require unlikely recombination patterns). Among the 1,066 offspring with genotyped parents, such errors occur at a median of 1,526 of 19,965 (7.64%) chromosome 22 sites per individual (Figure 61).

Overall the subtle genotyping errors affect 19,926 of the 19,965 (99.8%) SNPs on chromosome 22. Most SNPs show errors in relatively few individuals (median 58), with a small number of SNP locations across the genome showing higher error rates (maximum 605 of 1,066 (56.8%; Figure 62).

I examined the chromosome 22 genotypes before and after phasing for one family trio. Of the 19,965 SNPs in the phased data, 15,479 have been genotyped in the child; all these phased SNP calls are consistent with the calls in the unphased data. In addition, 7,675 SNPs are called in the offspring and both parents. At 17 sites, SHAPEIT reported an offspring haplotype phase inconsistent with the unphased parental genotypes—in 4 cases the SNP had not been called in the child; in the remaining 13 the called SNP was phased the wrong way around compared to parent genotypes. These inconsistent calls might result from constraints of the SHAPEIT model parameters, particularly the conditioning states (-K) parameter, which specifies the number of conditioning states upon which

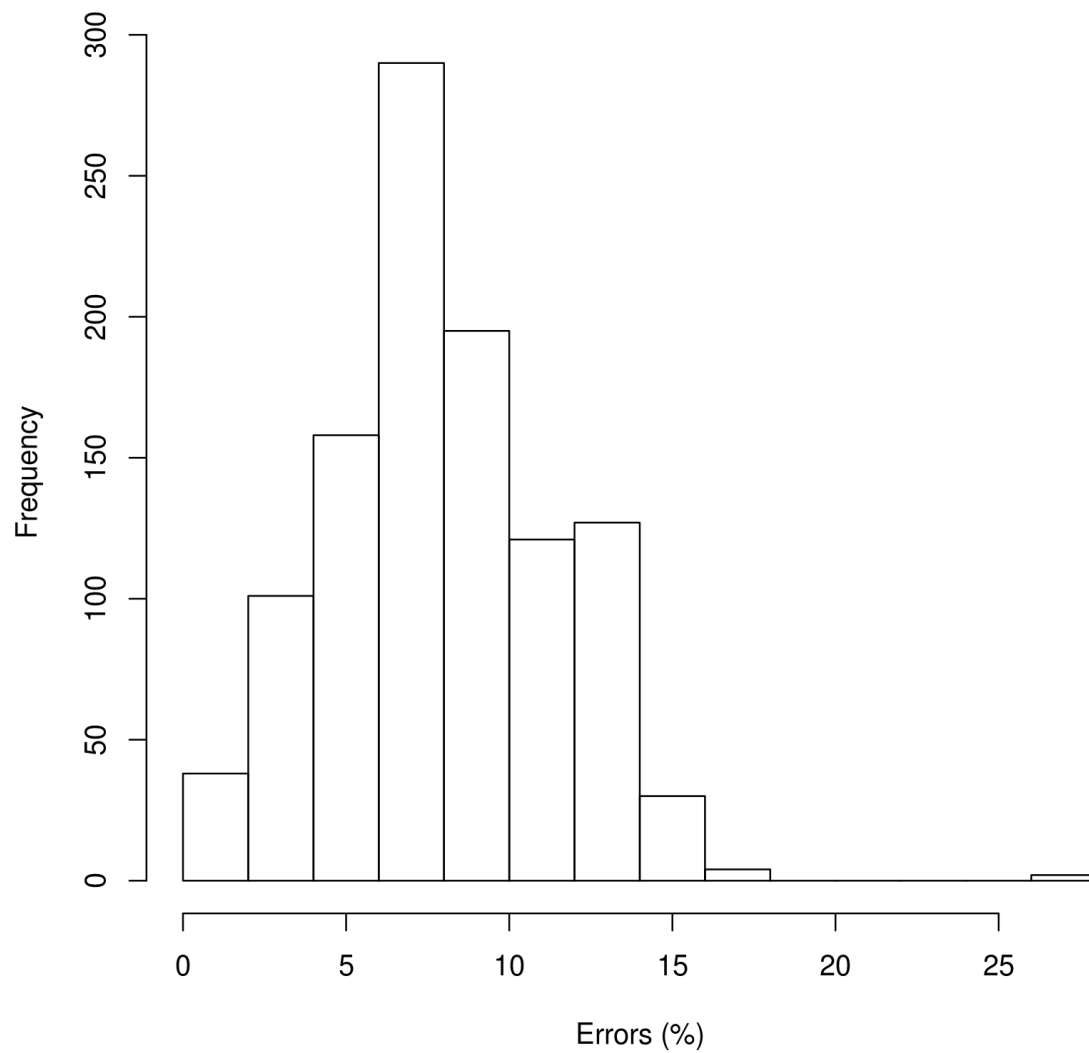


Figure 61: Proportion of genotyping errors on chromosome 22. A median of 7.64% of the SNP calls in each individual are reported as subtle genotype errors; i.e. those that would require unlikely recombination patterns.

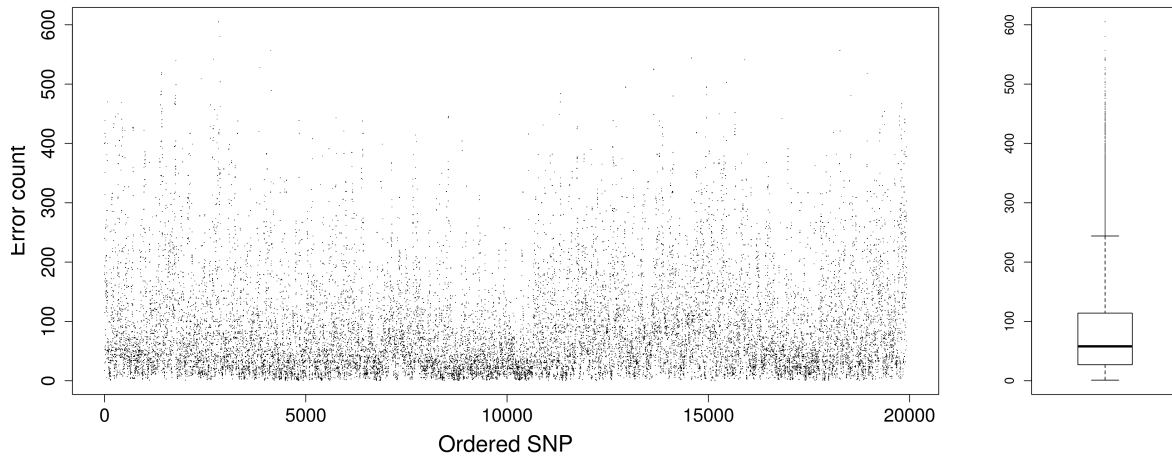


Figure 62: Per-SNP subtle genotyping errors across chromosome 22. The SNPs with the highest error rates are distributed relatively evenly across the chromosome. Most SNPs, however have relatively few subtle genotype errors (median 58), as shown in the accompanying boxplot.

haplotype estimation is based.

I investigated haplotype concordance among the monozygotic twins in the sample, as an indicator of phasing and imputation accuracy. 278 of 292 twin pairs show >95% concordance across the 19,965 chromosome 22 SNP locations after phasing and imputation (Figure 63). 14 twin pairs show poorer concordance (81–89%), probably as a result of a small number of poorly inferred recombination events. Examination of discordant loci generally shows matching genotypes with opposite phase, indicating discordance is the result of misphasing rather than poor imputation. Discordance generally results from a small number of haplotype blocks incorrectly phased in a given twin pair, but no region of the chromosome is particularly prone to discordance.

These results suggest that SHAPEIT is building sensible haplotypes despite the need to impute at many sites (a median of ~55% of the total SNPs were imputed for each individual; see Chapter 6 for details of the genotypes). Some of the errors will be a results of erroneous haplotype estimation among parents in the pedigree (who do not themselves have parental information available to help the phasing process). In addition,

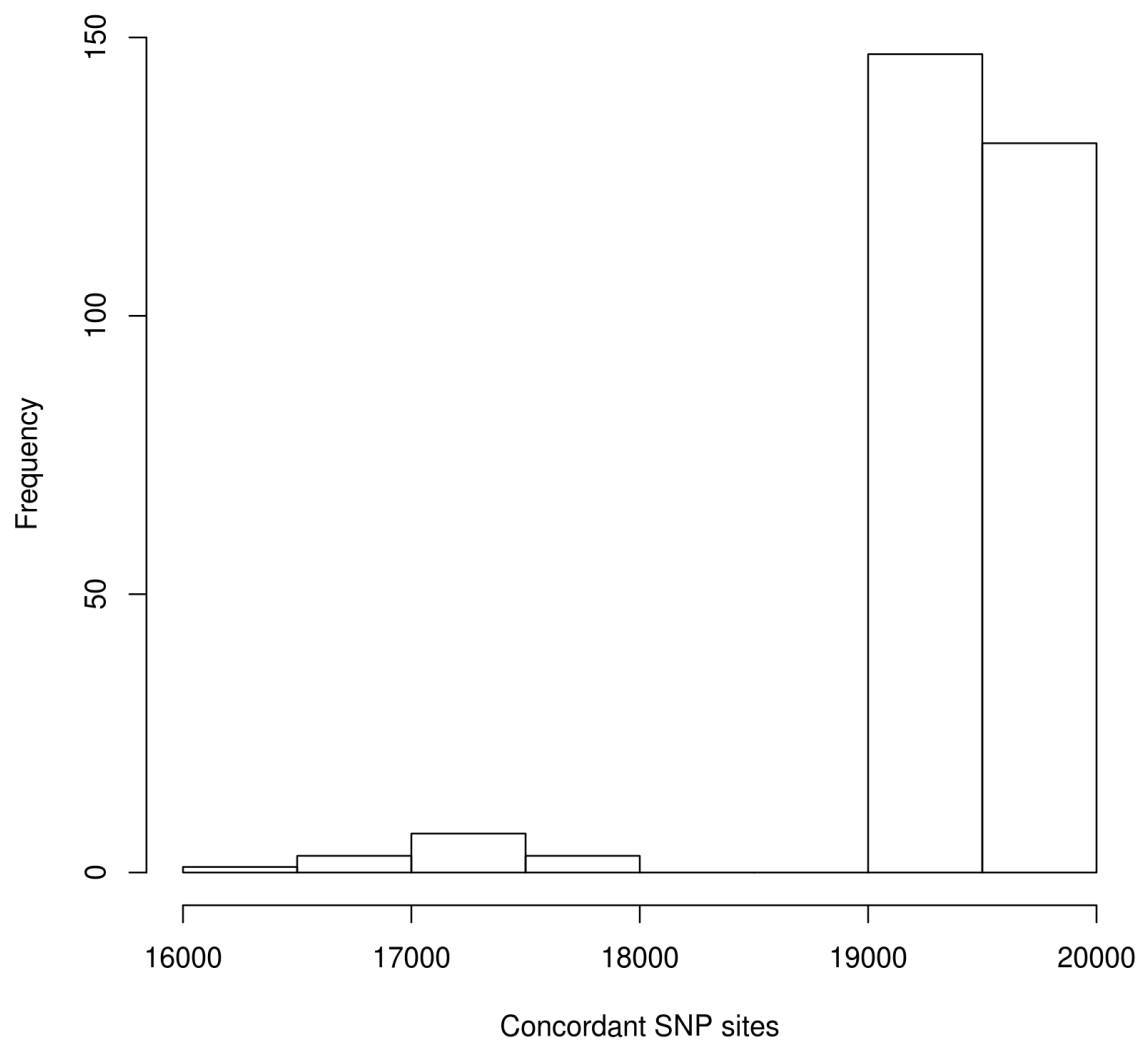


Figure 63: Haplotype concordance among monozygotic twins. Most monozygotic twin pairs show >95% concordance across the chromosome 22 SNP locations. 14 twin pairs show poor concordance, which probably results from a small number of poorly inferred recombination events.

the haplotypes tested are merely the most likely estimates based on the haplotype graphs generated by SHAPEIT. Subsequent kinship calculations can take into account the phase uncertainty, which alleviates the effect of some of the reported subtle errors in genotyping.

Kinship

In order to calculate heritability of the NTR expression traits by parent of origin, it is necessary to estimate the level of relatedness (kinship) among the members of the population. A number of tools are available to perform this task using either family pedigree information or genetic data. However, an analysis of parent of origin effects requires a more-complex approach. Specifically, this approach needs to adequately partition total kinship to determine how much is attributable to each parent of origin.

A number of methods exist for measuring relatedness between individuals. These approaches have classically been based on identifying expected sharing between individuals in pedigrees based on probabilities of being identical by descent. More-recently, methods have been developed to measure genome sharing directly from SNP data; such methods are reviewed in detail elsewhere.^{384–388} Nonetheless, no best practice is defined, and a number of problems exist in defining and measuring relatedness. One such problem that is important for the current study is that sequence identity at SNPs between any two haploid human genomes is typically much lower than the identity at non-SNP loci. Thus SNP-based measures for estimating relatedness are very dependent on the minor allele frequencies of the SNP set.

Many SNP-based methods exist for estimating relatedness.³⁸⁸ A common way to measure the similarity of individuals at a SNP is as the probability of a match between alleles drawn at random from them (K_{as}).

$$K_{as}(B, C) = \frac{1}{2} + \frac{1}{2m} \sum_{j=1}^m (S_{Bj} - 1)(S_{Cj} - 1)$$

$$= \frac{1}{2} + \frac{1}{2m} X_B X_C^T$$

Where $K_{as}(B, C)$ is the estimated relatedness between two individuals, B and C; m is the number of SNPs; and S_{Bj} and S_{Cj} are the respective genotypes of individuals B and C at the j^{th} diallelic SNP (coded 0, 1, 2). The diagonal elements of this matrix are not equal to unity and are instead defined by the level of homozygosity among individuals.

This scheme scores heterozygotes as having a relatedness of 0.5 (based on the probability of a match); in reality the identity by descent is actually 1 or 0 depending on phase, which is often known. Thus another possibility is to score matching heterozygotes as 1 (K'_{as}).

$$K'_{as}(B, C) = 1 - \frac{1}{2m} \sum_{j=1}^m |S_{Bj} - S_{Cj}|$$

These approaches give all SNPs equal weight, whereas greater weight can be given to rare shared alleles (e.g. by centering the genotype scores; K_{c0}). This genetic relationship matrix can contain negative values and be scaled so the diagonal is unity.

$$\begin{aligned} K_{c0}(B, C) &= \frac{1}{m} \sum_{j=1}^m (S_{Bj} - 2p_j)(S_{Cj} - 2p_j)^T \\ &= \frac{1}{m} X_B X_C^T \end{aligned}$$

Where p_j is the population minor allele frequency at the j^{th} SNP.

Finally, it's also then common to standardize (which assumes each SNP to be equally informative; K_{c-1}).

$$K_{c-1}(B, C) = \frac{1}{m} X_B X_C^T \text{ where } X_{Bj} = \frac{S_{Bj} - 2p_j}{\sqrt{2p_j(1-p_j)}}$$

The K_{c0} and K_{c-1} approaches are in fact part of the same one-parameter family ($K_{c\alpha}$), with $\alpha = -1$ and $\alpha = 0$; studies can consider alternative values for α when searching for

appropriate genetic relationship matrices. These methods correspond to equations (6–10) in Speed and Balding (2015).³⁸⁸

Importantly, no interpretable relatedness parameter exists with which to measure different approaches against; thus we lack good criteria with which to choose among approaches. This lack of a canonical method does, however, allow approaches to be adapted to best suit applications and the genomic architectures of the traits of interest.³⁸⁸ Other possible approaches to estimating relatedness include genome-wide distributions of time since most-recent common ancestor, and haplotype-based methods that account for the length of genomic regions shared by two individuals. This latter approach is particularly interesting in the context of our already-phased genotype data.

Method

I initially follow the methods for calculating and partitioning the kinship between individuals published by Mott and colleagues.³⁴² This method was developed for a heterogeneous stock population of mice descended from eight inbred founder strains and is described in detail in the published article and supplement. The method denotes kinship as the genomewide average number of shared founder haplotypes between individuals, broadly equivalent to K_{as} described above.³⁴² It also describes the partitioning of total kinship (the genetic relationship matrix K) into two components: “the genomewide average number of haplotypes inherited from parents of the same sex” (K_+), and “the genomewide average number of haplotypes inherited from parents of opposing sexes” (K_-). The genetic relationship matrices are additive, such that $K = K_+ + K_-$.

Central to the published method is the assumption that “if two individuals (not necessarily siblings) share an allele, then either it was inherited from parents of the same sex, or from parents of the opposite sex.” Actually, this is not strictly the case; for example, two matching homozygotes share alleles inherited from both parents of origin. This

description caused major issues in replicating the method. It would be more correct to say that the probability of a match between alleles drawn at random from two individuals can be further partitioned according to whether or not a match also exists in parent of origin.

For a human population, which is not described in terms of founder haplotypes, the same method can be adapted to score the similarity of two individuals at a SNP as the probability of a match between alleles drawn at random from each of them. This approach can be further extended so that the similarity is partitioned by parent of origin so that the scoring represents the probability of a match between alleles and parental origin on a random draw.

Following the published method I construct kinship matrices for each chromosome individually, and then combine them into a genomewide kinship matrix. Conservatively, SNPs with a Mendelian error rate greater than 0.05 (~20% of SNPs) are removed from the kinship calculation. The published method scales the kinship matrices so that the diagonal elements are unity. In fact, the diagonal elements in a K_{as} relationship matrix are by definition not unity³⁸⁸ and variation exists in levels of self relationship between individuals (Figure 64). I therefore construct kinship matrices both with and without this scaling step.

The above method was developed for a population of heterogeneous stock mice, comprising eight founder haplotypes in approximately equal proportions. In human populations such as the NTR the calculated allele sharing at any SNP will in fact be determined by the minor allele frequency at that SNP. As a result, kinship matrices for this project were also built using an alternative method, equivalent to K_{c-1} described by Speed and Balding (2015).

$$K_{c-1}(B, C) = \frac{1}{m} X_B X_C^T \text{ where } X_{Bj} = \frac{S_{Bj} - 2p_j}{\sqrt{2p_j(1-p_j)}}$$

X_{Bj} can be partitioned into maternal and paternal components

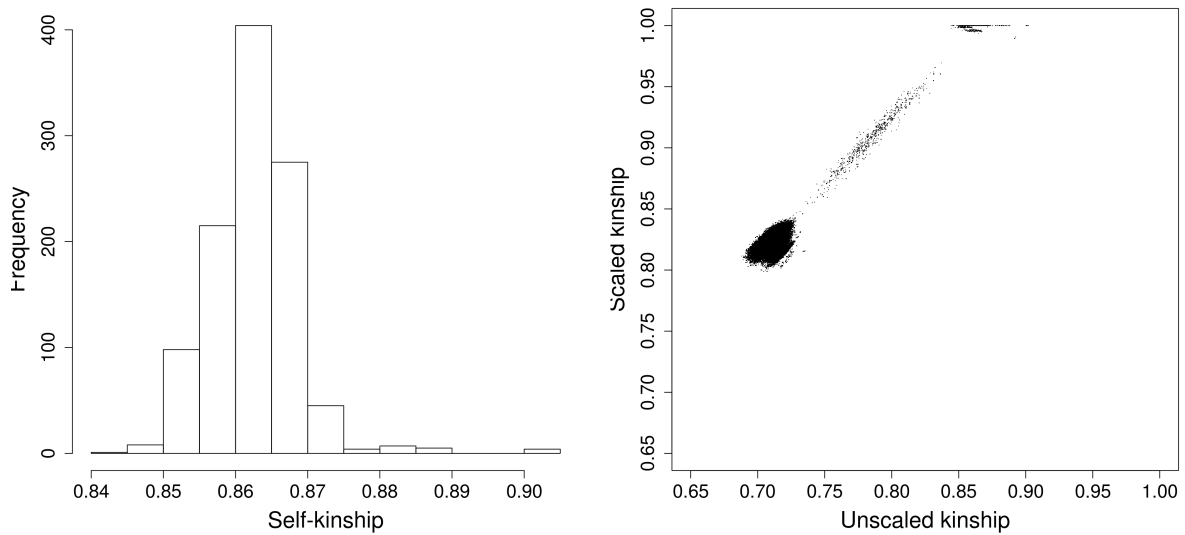


Figure 64: Evaluation of kinship estimated by allele sharing. The left panel (a) shows the distribution of self-kinship values (i.e. the kinship matrix diagonal) for the chromosome 1 kinship matrix. As expected some variation in kinship is seen, including some outliers with relatively high self-kinship. In fact, one individual (ID 999001996) from the parental generation has almost perfect self-kinship indicating that they are homozygous at all SNPs across the chromosome (not shown). The right panel (b) shows kinship calculated from raw allele sharing plotted against the rescaled kinship where the matrix diagonal is scaled to unity. Three distinct groupings are seen, corresponding to unrelated individuals, siblings, and monozygotic twins and self. The two approaches are relatively consistent, although the rescaling dramatically reduces the variability seen in the monozygotic twin/self grouping.

$$X_{Bj} = X_{Bj}^M + X_{Bj}^P$$

$$X_{Bj}^M = \frac{S_{Bj}^M - p_j}{\sqrt{2p_j(1-p_j)}} \text{ and } X_{Bj}^P = \frac{S_{Bj}^P - p_j}{\sqrt{2p_j(1-p_j)}}$$

Written in simple matrix form

$$K = GG' = (M + P)(M + P)' = (MM' + PP') + (MP' + PM')$$

$$K = K^+ + K^-$$

where K represents total kinship (genetic relatedness), which is partitioned into K^+ and K^- elements representing allele sharing from matching and nonmatching parent of origin respectively. G is the matrix of total number of reference alleles (0, 1, 2; subjects x SNPs), and M and P are matrices of the number of reference alleles on the maternally and paternally inherited chromosomes, respectively.

This method centers the genotype scores and standardizes the outputs, and in theory is more appropriate for human populations (it is equivalent to the method used for calculating genetic relationship matrices in GCTA). In practice, the results are very consistent with those of the K_{as} method outlined above (Figure 65). Similarities between the outcomes of different matrices have previously been reported.³⁸⁹

The final step in the creation of the kinship matrices was to transform them into a format suitable for passing to GCTA as a genetic relationship matrix (see below). These matrices are “always relative to an arbitrary base population in which the average relationship is zero. We use the individuals in the sample as the base so that the average relationship between all pairs of individuals is 0 and the average relationship of an individual with him- or herself is 1.”¹⁸ Thus most of the subsequent analysis is done with the K_{c-1} -format matrices, which match this description.

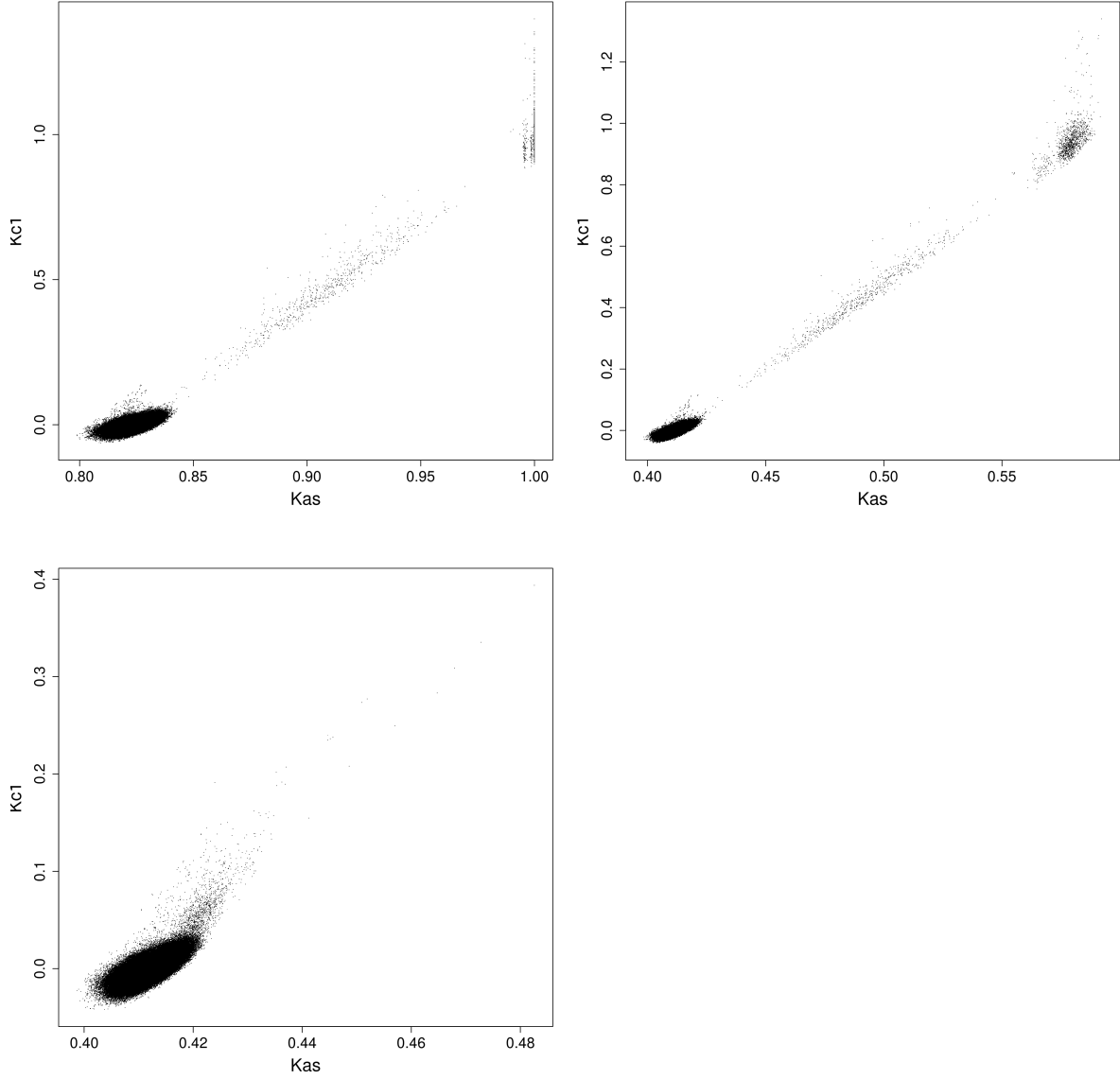


Figure 65: Comparison of the scaled K_{as} and K_{c-1} kinship methods. The panels show total kinship (K; top left), parent-of-origin kinship (K+; top right) and non-parent-of-origin kinship (K-; bottom left). Note the difference in x axis scale. In general, the results are consistent across kinship methods, with distinct groups representing unrelated, sibling and monozygotic/self kinship. Similar results are seen for the unscaled K_{as} kinship (not shown).

Family structure

Overall, as expected, nonsiblings do not differ in levels of allele sharing between parents of matching versus non-matching sex. By contrast, siblings show greater allele sharing between parents of matching sex than between parents of nonmatching sex (Figure 66). This excess of shared alleles between parents of matching sex is responsible for almost all the additional genetic similarity among siblings compared to nonsiblings. Interestingly, however, the allele sharing between parents of nonmatching sex does seem to be greater among siblings than among nonsiblings. In some cases this is likely to be attributable to errors introduced by the phasing process. The other option here is that siblings do in fact also share more alleles by nonmatching parent of origin. One possible cause is that the parents in the NTR cohort might be slightly more related than average, as a result of spatial structure in the Dutch population (D Boomsma, personal communication).

Heritability of gene expression

Heritability is a summary measure of the proportion of the variation in a trait that is attributable to variation in genetic factors.¹⁷⁹ In Chapter 4 I present an analysis of the heritability of reproductive traits in CC mice, using several methods to estimate broad and narrow sense heritability. In the remainder of this Chapter I build on the NTR study of heritability of gene expression in peripheral blood³⁴⁷ and present an analysis of parent of origin effects on the heritability of these traits.

Common approaches to heritability estimation model the similarity between individuals to separate the effects of genotype and environment—in humans a common design is the twin study in which the differing similarity of monozygotic and dizygotic twins is used to estimate heritability. The NTR study used three separate methods for heritability estimation: first, twin-based heritability via a REML mixed model with random additive genetic components of variation and shared and individual-specific environmental effects,

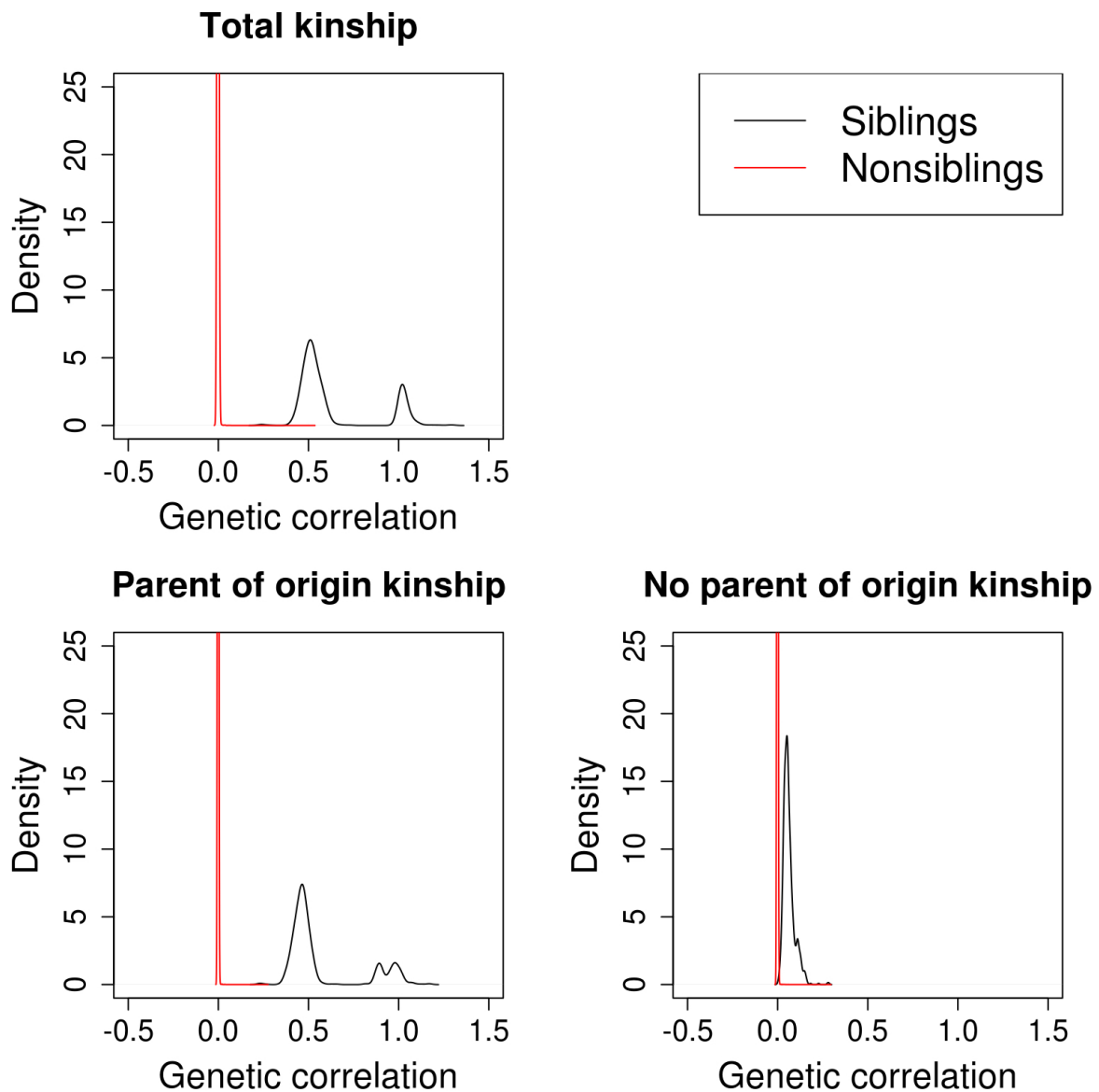


Figure 66: Allele sharing among siblings and nonsiblings. Overall (top left panel), nonsiblings do not differ in levels of allele sharing between parents of matching versus nonmatching sex. By contrast, siblings show greater allele sharing between parents of matching sex than between parents of nonmatching sex. This excess kinship is accounted for almost entirely by allele sharing from matching parent of origin (bottom left). Nonetheless, siblings also show consistently greater allele sharing by nonmatching parent of origin (bottom right). The double peak in the parent-of-origin kinship panel at genetic correlation of 1.0 results from a small number of misphasing events between monozygotic twin pairs in the population.

corresponding to a standard ACE model; second dizygotic only heritability following a constrained ACE mixed-model approach for full siblings; third, heritability estimated from a genetic relationship matrix, implemented in the software GCTA (discussed in more detail below). The NTR study used the first approach (standard ACE model) as their primary method.

In this Chapter I estimate the heritability of the 44,241 gene expression traits using the software GCTA.³⁹⁰ One of the functions of GCTA is to estimate the variance in complex traits explained by all SNPs on a chromosome or genomewide. The basic method fits the effects of all SNPs as random effects in a linear mixed model, and estimates V_G by restriction maximum likelihood based on a genetic relationship matrix estimated from the SNPs. The subsequent analysis was done with GCTA version 1.24.3 (current as of 5 June 2014; most recent is 1.25.2 as of 22 December 2015).

Total heritability

I first use GCTA with the total kinship matrix (K) to calculate heritability of each gene expression trait. The heritability of most transcripts is low (mean $h^2=0.0636$), indicating that many gene expression traits show no notable heritability in this population (Figure 67a). This finding is consistent with previous reports in peripheral blood,^{347,391} although higher mean expression heritabilities have been reported in tissues with more-homogeneous cell type.^{391–393}

I compare the heritability of the gene expression traits with the twin-based h^2 reported by the NTR expression study for 18,392 genes (selecting for each the transcript with the largest h^2 in the NTR study).³⁴⁷ In general, there is agreement between the results of this approach and the results provided by the NTR paper (Figure 67b). This agreement is improved slightly by selecting for each gene the transcript with the largest h^2 regardless of whether this matched the transcript used by the NTR. Nonetheless, the distributions

of reported heritability differ, with the NTR expression study generally reporting larger h^2 estimates, despite using an unbiased estimate of h^2 (i.e. not bounded by zero; Table 24). Other studies report even greater h^2 for peripheral blood gene expression (e.g. mean $h^2 = 0.150$).³⁹¹

The NTR study reports that heritability is strongly associated with expression mean and variance. In fact, our efforts to transform and normalize the data (see Chapter 6), make comparatively little difference to the heritability estimates, although they do bring them more in line with those of the NTR study.

Table 24: h^2 estimates for different study populations

NTR_h2	match_h2	max_h2
Min. :-0.430	Min. :0.003	Min. :0.003
1st Qu.: 0.010	1st Qu.:0.018	1st Qu.:0.030
Median : 0.100	Median :0.040	Median :0.065
Mean : 0.101	Mean :0.067	Mean :0.086
3rd Qu.: 0.180	3rd Qu.:0.085	3rd Qu.:0.112
Max. : 0.910	Max. :0.811	Max. :0.826

:The table lists the parameters from the NTR study³⁴⁷, and from our GCTA estimates with the same probes as the NTR and with the most-heritable probe for each gene.

The NTR study conservatively highlights 777 genes with significant expression h^2 ($q < 0.05$). 755 of the identified probes are present in our expression dataset, and in general these genes also show high heritability (Figure 67c). Not all of these probes have high heritability by GCTA, however, possibly as a result of inaccurate genetic imputation during the phasing process (Figure 67d). In fact, the NTR study notes “almost no correlation” of GCTA estimation with twin-based heritability.³⁴⁷ In this context, it is interesting that the present study does see a reasonable relationship between the two methods of estimating h^2 . The NTR study “reasoned that genomewide [identity by descent] might have reduced power for those genes influenced largely locally”, and therefore re-ran GCTA using kinship estimation performed in the local region within

1Mb of each transcript.

Parent-of-origin heritability

I next partition the heritability of gene expression traits into two portions, attributable to allele sharing from matching parent of origin (K+) and allele sharing from nonmatching parent of origin (K-). The variances associated with these components should be equal when no parent-of-origin effect exists.

$$V = K^+ \sigma_+^2 + K^- \sigma_-^2 + I \sigma_e^2$$

where the heritability attributable to K+ and K- is

$$h_+^2 = \frac{\sigma_+^2}{\sigma_+^2 + \sigma_-^2 + \sigma_e^2} \text{ and } h_-^2 = \frac{\sigma_-^2}{\sigma_+^2 + \sigma_-^2 + \sigma_e^2}$$

I again use GCTA for the analysis. GCTA has the functionality to estimate the heritability components from multiple kinship matrices. This functionality is most commonly used to partition the heritability of a trait between chromosomes,^{394,395} but also between subsets of SNPs and genomewide SNP sets³⁹⁶ and between *cis* versus *trans* components.³⁹¹ It also has the option to submit multiple genetic relationship matrices, which estimates the variance explained by the SNPs that were used to estimate the genetic relationship matrix. I use this option to partition heritability between matching and nonmatching parent of origin.

Among the full range of expression traits, little evidence is seen for widespread parent of origin effects (Figure 68). Among those traits that do show substantial differential heritability between the matching and nonmatching parent of origin components, the majority show a greater component attributable to allele sharing from nonmatching parent of origin (i.e. the K- genetic relationship matrix). The standard errors for these estimates are, however, large. One possibility is potential population substructure with parents

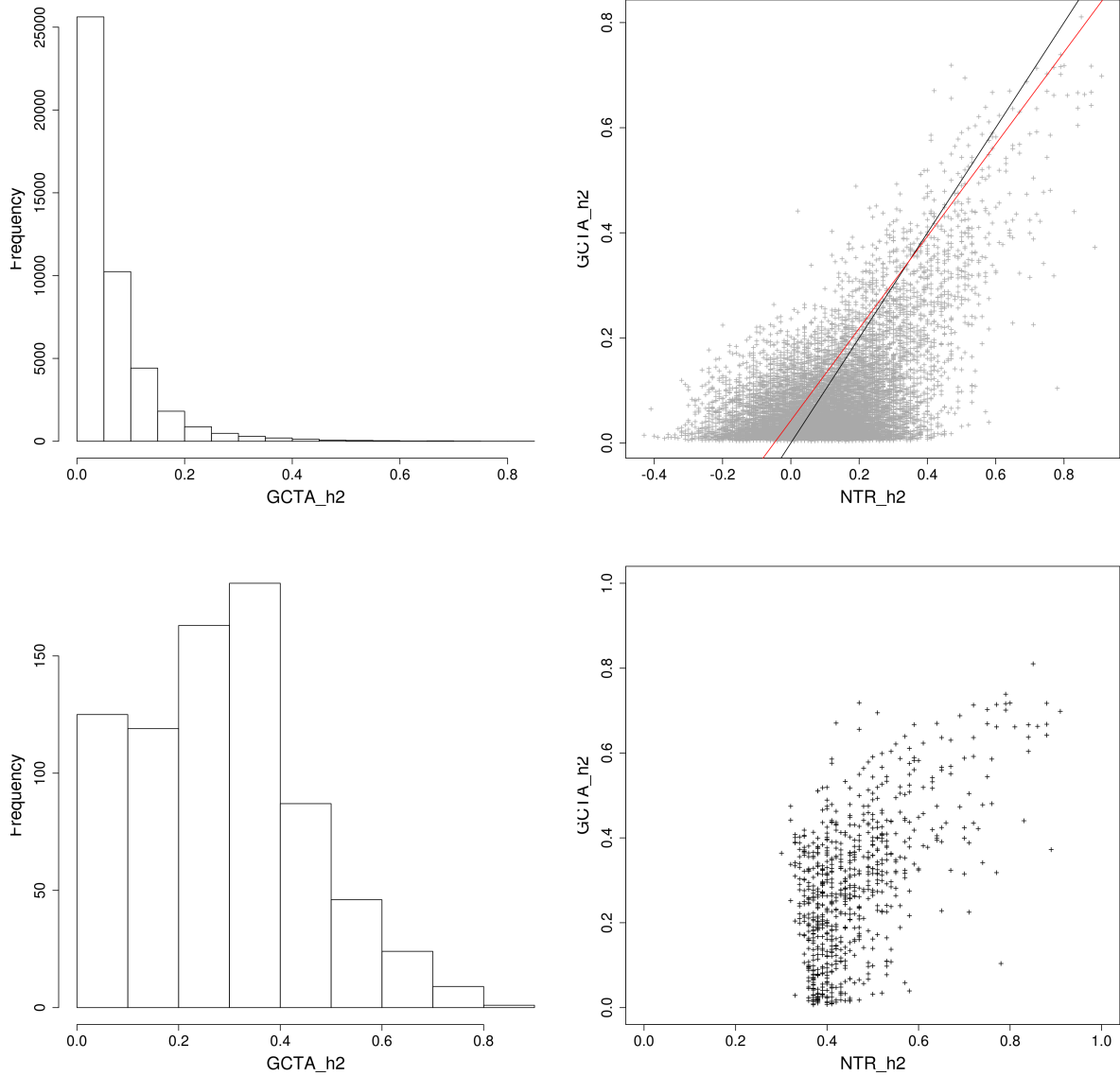


Figure 67: GCTA heritability estimation. The upper left panel (a) shows the distribution of heritability for the 44,241 gene expression traits. Most traits have relatively low h^2 . The upper right panel (b) shows the GCTA heritability estimates plotted against the NTR estimates for 18,074 genes. The black line is the line of $x=y$ and the red line is the regression fit. The lower left panel (c) shows the GCTA heritability estimates for 755 of the 777 genes identified by the NTR study as having significant heritability. Some of these genes have very little heritability estimated by GCTA. The lower right panel (d) shows these GCTA heritability estimates plotted against the NTR estimates. There is reasonable correlation ($r^2 = 0.57$) between the two estimates, but some GCTA estimates of heritability are very low.

who share a geographic location also sharing slightly more genome IBD than expected. Another more prosaic possibility is that the imputation done during phasing inserts an excess of shared versus non-shared alleles. The standard error for the heritability explained by matching parent of origin increases with increasing heritability, whereas the standard error for the heritability explained by nonmatching parent of origin decreases with increasing heritability (Figure 69). These large standard errors mean the results of this analysis should be treated with caution, as most h^2 estimates are not significantly different from zero.

As already established, the majority of traits have relatively low heritability, and these data points potentially confuse interpretation of these results. When only the 777 genes with significant expression h^2 are considered, however, the split becomes relatively even between those genes with a high matching parent of origin component and those with a high nonmatching parent of origin component (Figure 70). Using cutoff $K+$ of 0.3, similar to the cutoff for significant heritability in the NTR study, 270 probes in 200 genes show evidence of parent of origin effects. 142 of these genes are among the 777 NTR-identified genes. The sample of 200 genes is enriched for genes involved in a range of immune system processes (Bonferroni-corrected $P < 0.05$; Table 25). As it turns out, however, the full sample of 777 genes with significant heritability is also enriched for these processes. Immune system processes are among the traits for which substantial heritability makes sense in peripheral blood. I also see an indication of enrichment of some of these traits among the 142 traits with high parent of origin heritability within the 777 genes with highly heritable expression.

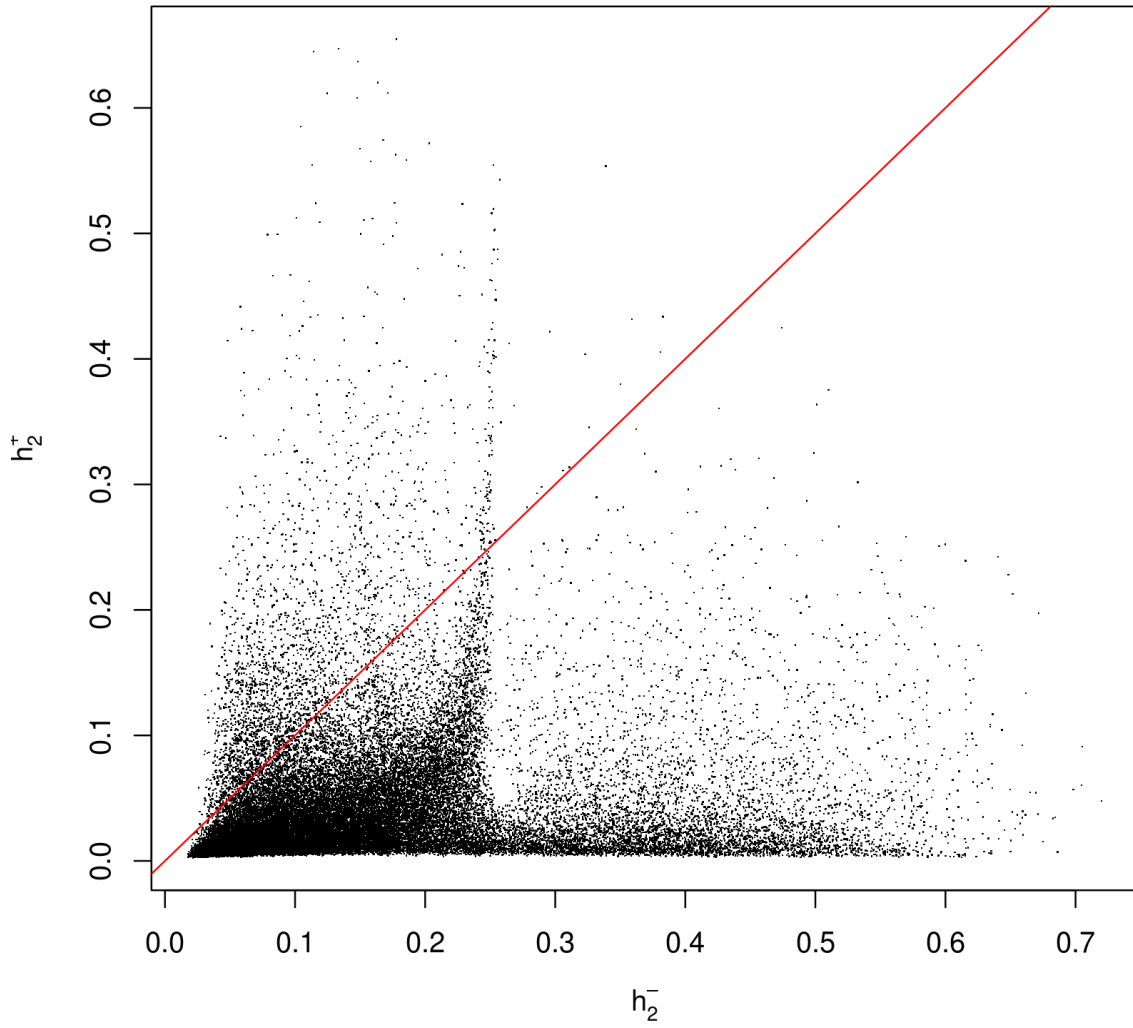


Figure 68: Heritability attributable to allele sharing by matching versus nonmatching parent of origin. Most gene expression traits tend to have a larger component of heritability attributable to nonmatching parent of origin. Some, however, show an indication of parent of origin effects.

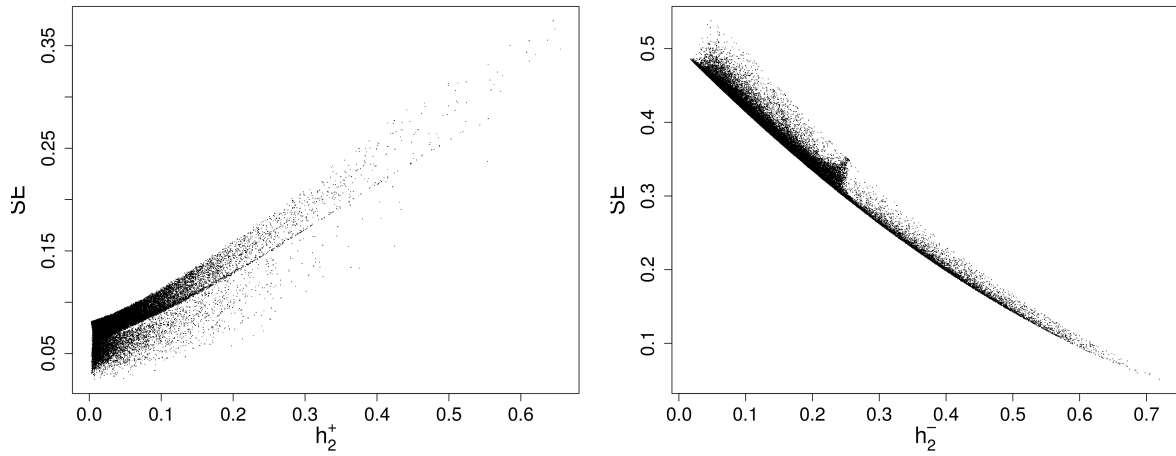


Figure 69: Standard errors for the H2+ and H2- components. Standard error tends to increase with increasing H2+ and decrease with decreasing H-

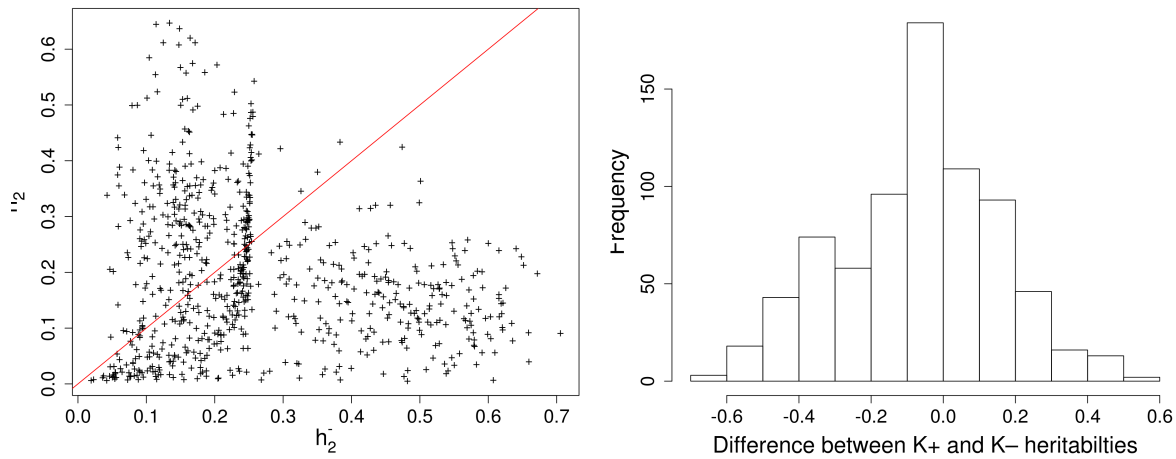


Figure 70: Heritability attributable to allele sharing by matching versus nonmatching parent of origin among 777 genes with significant heritability. The left panel (a) shows the 777 genes plotted in a similar manner to Figure 68. A much higher proportion of genes show evidence of parent of origin effects. The right panel (b) shows the distribution of the difference between the H2+ and H2- components. Roughly equal proportions of genes have a majority of heritability attributable to matching versus nonmatching parent of origin.

Table 25: Enrichment analysis for the 200 genes with evidence of parent-of-origin effects on gene expression

PANTHER GO-Slim Biological Process	Total#	Sample#	expected	Fold Enrichment	+/-	P value
natural killer cell activation (GO:0030101)	99	9	.92	> 5	+	1.06E-04
antigen processing and presentation (GO:0019882)	56	5	.52	> 5	+	4.34E-02
B cell mediated immunity (GO:0019724)	141	10	1.31	> 5	+	2.28E-04
macrophage activation (GO:0042116)	167	9	1.55	> 5	+	6.87E-03
cellular defense response (GO:0006968)	231	10	2.14	4.67	+	1.55E-02
immune response (GO:0006955)	518	18	4.80	3.75	+	4.46E-04
immune system process (GO:0002376)	1391	46	12.90	3.57	+	6.99E-12
response to stimulus (GO:0050896)	2170	49	20.12	2.44	+	6.73E-07
cellular process (GO:0009987)	6708	90	62.20	1.45	+	4.75E-03
Unclassified	8629	52	80.01	.65	-	0.00E00

:Analysis done by pantherdb.org. Reference: Homo sapiens; File: POE_genes_200.txt

To further check the method, I have similarly calculated and partitioned the heritability of the quantitative covariate traits according to parent of origin (Figure 71; Table 26). The results are interesting and in general they are consistent with the trends seen for the expression traits with no notable parent of origin effect and a tendency towards a greater component attributable to allele sharing by nonmatching parent of origin. Standard errors are reasonably large, especially for the nonmatching parent-of-origin portion of the heritability estimate. Importantly, not all these traits are expected to be heritable. For example the most extreme result is for participant age, a trait that seems to be completely heritable, with all heritability attributable to parent of origin. The ndays and D variables also make little sense in the context of heritable traits. The age results probably illustrate a problem with mixed model heritability estimation in humans—inferences vary according to levels of relatedness between individuals.³⁸⁸ In this case the age trait is shared between the most related individuals—i.e. monozygotic twins.

Table 26: **Heritability estimates of the sample covariates**

Trait	H2.POE	H2.NOPOE	SE.POE	SE.NOPOE	N
age	1	0	0.345	0.34	889
ndays	0.523	0.319	0.241	0.29	889
neutn	0.306	0.211	0.207	0.429	762
lympn	0.361	0.252	0.226	0.397	762
monon	0.229	0.463	0.119	0.214	762
eosn	0.038	0.704	0.03	0.067	762
bason	0.018	0.018	0.094	0.559	762
rbc	0.389	0.51	0.137	0.157	875
D	0.078	0.126	0.095	0.409	889
bmi	0.226	0.703	0.054	0.06	889

Imprinted genes

I compare our parent of origin results with a known list of imprinted genes. As discussed in Chapter 5 several options are available for the source of this list. Notably, none of these resources provides downloadable gene lists in a simple machine-readable format

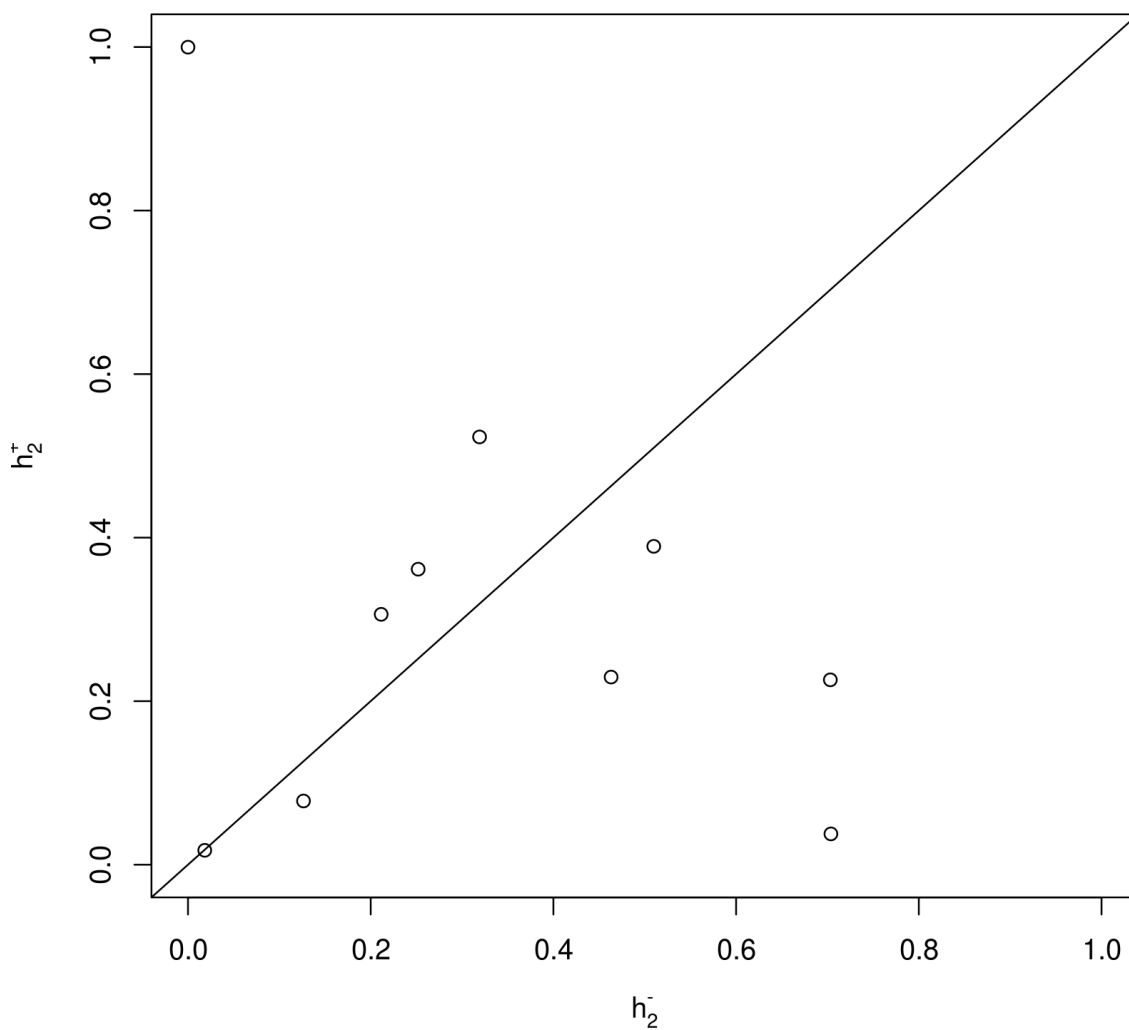


Figure 71: Heritability estimates partitioned by parent of origin for the 10 covariate traits. Some of these traits would be expected to be heritable, while others would not be. The traits generally show no evidence for an excess of heritability attributable to matching parent of origin. The exception is age, a trait that would not be expected to be heritable, but for which the results show almost 100% heritability completely attributable to matching parent of origin.

such as a text or .csv file. It is also not clear how well maintained the resources are, or what are the criteria for inclusion in the databases.

I obtained a list of imprinted human genes from *GeneImprint* (<http://www.geneimprint.com/>)²⁹² via the html table on the web page source. This resources lists 246 genes, categorized accoring to whether they are known to be imprinted, predicted or uncertain (Table 27). The other available resource—the Catalogue of Parent of Origin Effects^{295,296}—has a comparable list of 82 imprinted human genes among 340 entries for human parent-of-origin effects, but without simple categories.

Table 27: **Human imprinted genes**

Status	ExpressedAllele
ConflictingData: 8	Biallelic : 17
Imprinted : 97	Isoform Dependent: 4
NotImprinted : 15	Maternal :100
Predicted :107	Paternal :109
ProvisionalData: 6	Random : 2
Unknown : 13	Unknown : 14

:Data from GeneImprint (<http://www.geneimprint.com/>)²⁹²

Among the 246 human genes in the GeneImprint database, 186 are annotated in the NTR gene expression dataset (a further 10 are denoted by aliases). However, among the 97 genes categorized as ‘imprinted’ only 55 are annotated in the NTR dataset. 4 of the ‘imprinted’ genes and 12 of the total GeneImprint dataset are found in the 777 genes with significant heritability of expression, and in fact the heritability of gene expression in peripheral blood is low for such genes in the NTR and other studies. When overlaid on the prior plot of heritability attributable to matching versus nonmatching parent of origin it is clear that the majority of imprinted genes show no strong effect on heritability from allele sharing between matching parent of origin (Figure 72). The seven imprinted genes that do show these effects are noted in Table 28.

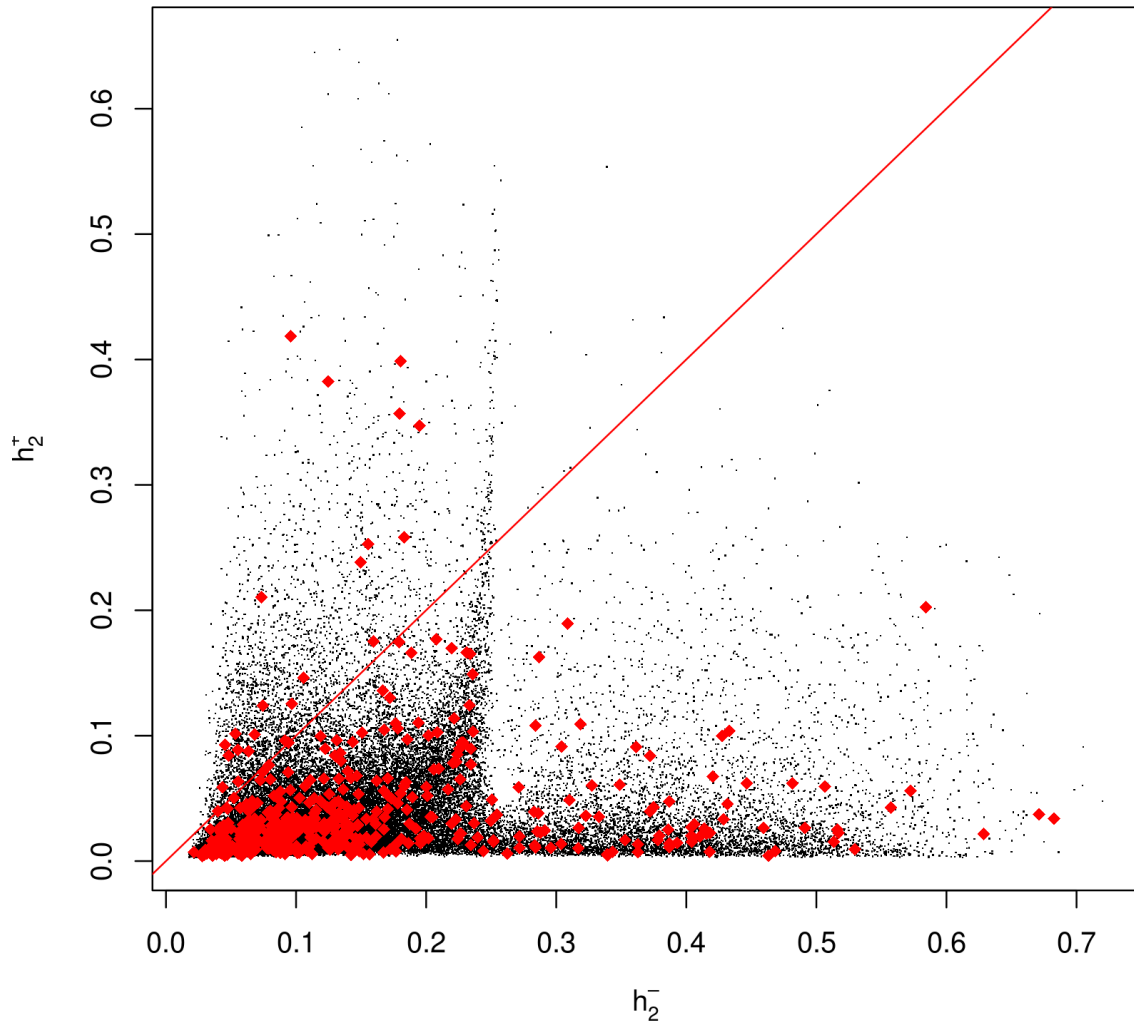


Figure 72: Imprinted genes show little evidence of heritability attributable to allele sharing from parents of matching sex. 246 human genes with confirmed or predicted imprinting are depicted in red.

Table 28: **Imprinted genes showing parent-of-origin heritability**

	H2.POE	H2.NOPOE	SE.POE	SE.NOPOE	N	gene.symbol
X11715394_s_at	0.238	0.15	0.161	0.419	889	CD81
X11715670_a_at	0.253	0.155	0.167	0.416	889	IFITM1
X11715671_x_at	0.258	0.183	0.164	0.393	889	IFITM1
X11718350_s_at	0.357	0.179	0.212	0.408	889	NLRP2
X11723050_a_at	0.419	0.096	0.265	0.491	889	HSPA6
X11723419_a_at	0.347	0.195	0.204	0.393	889	COL9A3
X11741832_a_at	0.382	0.125	0.238	0.46	889	SNRPN
X11754601_s_at	0.398	0.18	0.233	0.411	889	SPON2

Discussion

In this chapter I report the estimation of the heritability of gene expression traits in the NTR population of Dutch twins, and the subsequent partitioning of this heritability according to parent of origin. I have applied a method that partitions kinship between matching and nonmatching parent of origin, and used this to investigate differential contributions to the heritability of gene expression traits. I find genomewide h^2 estimates consistent with previous results, but little evidence of widespread parent of origin effects on gene expression. In fact the majority of the heritability of these traits seems to be attributable to allele sharing between nonmatching parent of origin. Nonetheless, approximately 270 probes in 200 genes enriched for immune system processes show evidence of parent-of-origin-specific differences in heritability. Known imprinted genes show little evidence of heritable expression attributable to allele sharing between matching parent of origin.

These findings contradict a previous report, in which an analogous method found evidence for parent-of-origin effects in the majority of 97 traits in mice (Figure 73a).³⁴² Two general possibilities exist for this discrepancy. First, the results might reflect problems with the method, specifically phasing, kinship calculation or heritability estimation. As

support for the correct application of the analytic approach, I confirm the heritability of gene expression traits in peripheral blood reported in the NTR gene expression study.³⁴⁷ Second, the results might reflect the gene expression trait, the choice of tissue, or errors in the prior reports. Below, I discuss each of these issues in turn.

Phasing

Correct partitioning of heritability and identification of parent-of-origin effects depends on accurate phasing by parent of origin. Errors in the phasing process will propagate into downstream analyses, where they affect the estimates of kinship and the construction of genetic relationship matrices in which allele sharing is partitioned between matching and nonmatching parent of origin. In this regard, it is important to note that our method uses SHAPEIT and DuoHMM—tools designed for accurate phasing of genetic data, but not validated in the context of large amounts of missing data resulting from multiple nonintersecting SNP platforms. Thus these tools are also imputing a large proportion of the data points during the phasing process. Encouragingly, evidence supports the approach of combining individuals genotyped on different platforms into a single genetic relationship matrix, by phasing together with a reference set.³⁹⁷

In general, the phasing method performs relatively well. The outputs resemble well-formed genotypes (however, how well they resemble the ‘true’ genotypes of the individuals is an open question). Subtle genotype errors (i.e. those requiring unlikely recombination patterns) are present at <10% of SNP sites per individual. Mendelian errors in which offspring genotypes are incompatible with parental genotypes are similarly common. It would be instructive to quantify the accuracy of the haplotype-based imputation done by SHAPEIT—one approach that would not require gold-standard data is masking the real data and assessing how many alleles are correctly imputed by the phasing algorithm.³⁶³

Importantly, while tests suggest the imputation works reasonably well, some monozygotic

twin pairs show poor haplotype concordance, with different parent of origin allocated to a given haplotype. These errors propagate downstream into kinship estimation, and presumably also affect kinship estimates between other pairs of individuals. Such errors do not affect the total kinship, but are important for the partitioned kinship.

Several steps could be taken to improve the quality of the phasing, which might ultimately improve the heritability estimates. First, the phasing could be redone with the latest SHAPEIT version, with the most recent 1000 Genome reference set. Second, the reference set could be enhanced by incorporating the Genome of the Netherlands data, which is a reference set specific to the Dutch population.^{380–382} Third, the SHAPEIT phasing process uses family pedigree information to inform possible haplotype states; being able to use monozygotic twin status would provide additional information to the algorithm. Finally, SHAPEIT also outputs phasing uncertainty for use in downstream analysis. At present this work uses the definitive haplotype calls, but adapting these methods to take into account the phasing uncertainty (i.e. using allele doses) might result in more-accurate kinship estimates.

Kinship

Estimation of genetic relatedness from genomewide SNPs is a well-established method for estimating heritability of traits. However, no optimum method exists, and how well genomewide kinship represents biological processes is not well defined.³⁸⁸ Our application of kinship estimation produces genetic relationship matrices that clearly differentiate unrelated, sibling and monozygotic pairs. Nonetheless, a number of potential improvements would be possible.

First, while our method accurately partitions kinship according to inheritance from matching or nonmatching parent of origin, it does produce a larger than expected K-element for some individuals. One reason for this is likely to be inaccurate phasing, which

mistakenly assigns haplotypes to nonmatching parent of origin in related individuals. A second possibility is that for some reason related individuals do in fact share more alleles by nonmatching parent of origin than do unrelated individuals. A solution to these issues might be to adapt the kinship calculation method to take into account the uncertainty in the phased haplotypes, as discussed above.

A potential problem with estimating heritability from genomewide SNPs is bias introduced by linkage disequilibrium. In short, contributions to h^2 are overestimated in regions of strong linkage disequilibrium and underestimated in regions of weak linkage disequilibrium. Another obvious extension to this kinship method would be to develop a linkage-disequilibrium-adjusted kinship method, analagous to tools such as LDAK.³⁸⁶

One further option for the extension of the kinship method would be to redefine genetic relationship as a measure of haplotype sharing.³⁸⁸ This approach is particularly appealing in the context of this study, as it would tie the kinship estimation to the process of phasing by parent of origin.

Heritability

The overall heritability results are consistent with previously published results and suggest that heritability estimation using GCTA is appropriate in this population. Nonetheless, a major problem with the GCTA approach to partitioning heritability seems to lie in the standard errors of the heritability estimates. While the standard errors are low for the overall heritability estimates they are high for the partitioned heritability estimates and seem to scale with the magnitude of the estimate (positively for h_2^+ versus negatively for h_2^-). In fact the GCTA-GREML power calculator and associated article suggest that standard errors should not vary with SNP heritability.^{398,399}

The standard error of the heritability estimate depends on the sample size and the genetic relatedness of the individuals (measured as the variance of the off-diagonal elements of

the genetic relationship matrix). This value is 0.000615 for the total kinship matrix (K), and the resulting standard error of 0.0641 agrees closely with the observed standard error of the total kinship estimates among the 889 individuals, with 95% power to detect significant h^2 at a true SNP heritability of 0.23 and above. By contrast, the empirical variances of the K+ and K- matrices are 0.000507 and 0.0000143, respectively, which translates to expected standard errors of 0.071 and 0.421 for the h^2+ and h^2- estimates. Much larger sample sizes would seem to be needed to reliably estimate heritability in these cases.

An additional point to note with this approach to estimating heritability is the inclusion of related individuals. In cases when individuals who share SNP genotypes more often than the average also tend to share a common environment then the heritability explained by the SNPs will be overestimated.³⁸⁷ In this case the partitioning of heritability between matching and nonmatching parent of origin requires that individuals in the study be differentially related. It is not clear whether the magnitude of this differential relationship needs to be so severe. Nonetheless, removing for example one of each monozygotic or sibling pair would noticeably reduce the sample size and thus the power to detect significant heritability. The NTR population are all adults, and siblings are therefore less likely to share a common environment than child participants would be.

The linear-mixed model approach for estimating heritability from genomewide SNPs, as used in GCTA, is a commonly used tool in statistical genetics. Nonetheless, at the time of finalizing this writing the validity of GCTA has been called into question.^{400–403} Public discussion of these criticisms will no doubt continue.

Strong *cis* effects

One possible explanation for the relative absence of parent of origin effects on heritability is that unlike complex traits gene expression could be governed largely by local effects, with

the result that they are not significantly heritable by genomewide estimation. A related point is that peripheral blood is a highly heterogeneous tissue, comprising many cell types, with expression subject to sex effects and other hormonal influences.⁴⁰⁴ Nonetheless, a portion of these traits is significantly heritable by genomewide assessment, and these traits largely overlap those previously reported. The genes with significant heritability of expression are enriched for immune system processes, which is consistent with the expectation of peripheral blood.

Importantly, while *trans* regulation is reported to make a notable contribution to the heritability of gene expression, *cis* regulation is suggested to be more important in tissues comprising heterogeneous cell types.³⁹¹ Thus an expectation exists that local effects might be relevant in this case. While the majority of genes have *cis* regulatory variation,³³⁷ genomewide effects continue to represent a major component of heritability of gene expression traits, possibly through action of *trans* factors on *cis*-regulatory variants.⁴⁰⁵ The role of local effects could be investigated further by further partitioning the heritability of expression into local and genomewide components. This approach would, however, likely require a larger sample size.

Context

The work in this chapter investigates the partitioning of heritability according to parent of origin for gene expression traits. This finding contradicts a previous report, in which a similar method found evidence for parent-of-origin effects in the majority of 97 traits in mice (Figure 73a).³⁴² These partitioned heritability estimates have surprisingly small standard errors relative to the size of the population (especially for the K- partition).^{342,398} For example, to obtain the observed standard error for the Biochem.HDL trait the variance of genetic relationship in the K- matrix must be 40 times greater than in the NTR data, at a level comparable to the variance of total kinship in the NTR data.

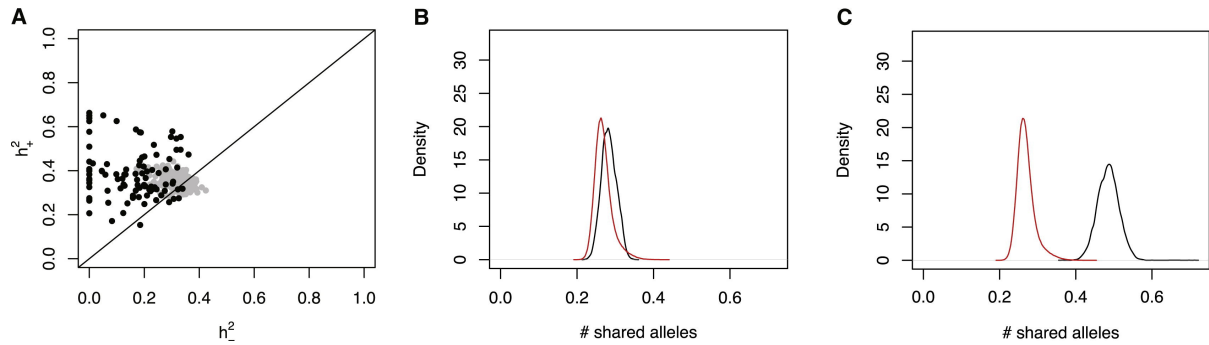


Figure 73: Heritability of parent-of-origin effects.(a) Heritability estimates for 97 traits measured in HS mice. Each black dot represents one trait. y axis, the heritability h^2_+ attributed to allele sharing from parents of the same sex; x axis, heritability h^2_- from parents of the opposite sex. Gray dots are corresponding estimates from simulations of nonimprinted complex traits. The diagonal is the line of equality between the heritabilities. In 91 of 97 (93%) of the traits, I found $h^2_+ > h^2_-$ ($P < 10^{-22}$ under the null hypothesis that $h^2_+ = h^2_-$, one-sided binomial test). Copyright Mott et al. 2014. Modified with permission under CC-BY.

An understanding of the incidence and mechanisms of parent-of-origin effects will enable phase-aware analysis and modeling of complex traits. The work presented in this chapter investigates the partitioning of heritability according to parent of origin, and would be enhanced by further investigation of *cis*-QTLs, parental effects and *bona fide* parent-of-origin effect. Such work is, however, outside the scope of this thesis.

Chapter 8 | Conclusions and perspectives

Science—and engineering, which here includes agriculture and medicine—
DOES NOT BEGIN WITH A TIDY QUESTION. Nor does it end with a
tidy answer.

—John W Tukey (1980)⁴⁰⁶

Outcomes

In this thesis I discuss two separate research projects unified by the broad general theme of measuring and explaining variation in complex traits. The first project identifies QTLs for reproductive traits in the Collaborative Cross population of mice through analysis of breeding records. I show that litter size and other traits are heritable in this population, and do genetic mapping using a mixed model approach. The most-robust QTL, on chromosome 5, explains around 3% of the total variation in litter size and approximately 24% of the variation attributable to genetics. This work also identifies QTLs for time between litters and identifies candidate genes suitable for follow up.

The second project uses a previously described method to investigate parent-of-origin effects on gene expression traits in a populations of 889 human twins. This work replicates prior genomewide heritability estimates for many of the gene expression traits in peripheral blood, but finds little evidence for widespread parent-of-origin effects on the heritability of these traits. This finding contradicts a prior report of widespread parent-of-origin effects on the heritability of 97 traits in mice. On close examination, the sample size severely limits the power to detect such effects. Nonetheless, I identify approximately 200 genes enriched for immune system processes that show evidence of parent-of-origin-specific effects on heritability.

These results provide incremental additions to the body of evidence in the domains of reproductive genetics and parent-of-origin effects.

Further work

A number of options are available to take this work forward. For further investigation of reproduction in mice, knockout studies of candidate genes might be fruitful, especially of *Scarb1* as a candidate for litter size. Another option would be to investigate crosses of high and low litter size lines with the correct genotypes at the QTL from the TAU CC population. An additional promising avenue of investigation might be a type of ‘survival’ analysis, where the CC lines are assessed for their survival through each successive generation, or to a given time point. Finally, bearing in mind the poor quality of the breeding records used in this analysis it might be productive to replicate the findings in another breeding population—potentially one of the other CC populations.

Some additional options for the project as it currently stands would be to calculate confidence intervals for the QTLs by simulation,¹¹³ and to do merge analysis of sequence variants over the QTLs.⁴⁰⁷ Merge analysis might work particularly well for the QTL for time between litter on chromosome 12, where the founder effects split into clear groups. Nonetheless, although merge analysis has been used in several CC mapping projects only acts to eliminate possible causal variants and has yet to provide any notable results in this population.

Further investigation of parent-of-origin effects in the NTR population would ideally entail further partitioning of the heritability of gene expression traits into local and genomewide components. Additional work includes investigation of *cis*-QTLs, parental effects and genuine parent-of-origin effects via association mapping. Some of this work has been done outside the remit of this thesis. Improving the quality of the phasing by parent of origin might also improve the heritability study; ultimately, however, the heritability estimation approach has been restricted by sample size and a population of 4,000–5,000 individuals would seem to be necessary to sufficiently reduce the standard error of the heritability estimates for the $K_{\perp}$ components.

The identification of appropriate follow-up work for these studies should, however, be guided by the wider context of knowledge and need. Research in genetics provides a means to improve health care, to increase the efficiency of agriculture, and to generate societal and economic benefit. It also offers a perspective on the origins of life and what makes us human. Too often, however, results are not well contextualized within these domains. In this regard, the question of what follow up work to do should really be reframed as whether follow-up work is beneficial. In this regard, considering the resources involved it is possible that these results are not sufficiently interesting to be high priority for follow up.

Critical perspectives

In terms of assessing the success of the two projects, it is instructive to evaluate them against the respective project proposals. The research done in the CC was initially intended to “dissect the genetic basis of variation in fertility” on the basis of preexisting “fertility statistics and DNA samples within each line from each generation”. Unfortunately the breeding records for the population were not as described, and should probably have prompted an immediate reevaluation of the project goals. The dataset is undocumented and poorly maintained, and is a definite limitation of this study.

The study of parent-of-origin effects on heritability of gene expression traits in humans has ultimately been hampered by small sample size resulting in low power to detect significant heritabilities. In fact this human population shows considerably less variation in kinship between unrelated individuals that does the mouse population used in which similar methods were previously applied. In this regard, it might have been instructive to evaluate the power of the analysis at the design stage to enable decisions about the value of the project to be taken.

When developing this thesis developing a coherent narrative proved a difficult task. The

conventional advice is to identify a clear problem definition that can be summarised in a single statement. Unfortunately, I can formulate two such statements, which are only loosely related. With the benefit of hindsight, this thesis would have benefitted from a very different approach to its design and planning, with the research projects focused around a clear question (or series of related questions).

With respect to John Tukey (quoted above), while science does not necessarily begin with a tidy question, some effort to clearly define a research question in the context of current knowledge does seem to be time well spent. This argument has been made elsewhere—although a research project is unlikely to follow a linear path from A to B,⁴⁰⁸ beginning with a well-defined question enables the researcher to make sensible decisions about the project. Beginning a project without well defined goals means there is no context with which to take decisions. Moreover, by not defining the parameters of a project one also does not define the criteria for success, or an objective measure of how to assess progress.

Data in biological research

The increasing availability of data, coupled with rapid development in technology, has had a substantial impact on biological research in the past decades. Within the domain of genetics, large data resources have become commonplace, and their maintenance and management has grown in importance.

A substantial portion of this thesis focuses on the technical details of data management and analysis. Although none of the observations on data maintenance in these sections is particularly novel,^{409,410} they do highlight issues that are rarely openly acknowledged and discussed in research articles—namely that the origins and maintenance of data matter. While universities and journals now typically offer repositories for data, there is very little by way of professional standards or institutional policy to ensure that data are collected,

stored and made available in appropriate forms.

For example, the CC TAU breeding records have been poorly maintained and on examination contain obvious errors and inconsistencies. I have invested substantial time in correcting as many of these errors as possible while ensuring all corrections are carefully documented—a somewhat thankless task. The cleaned dataset will be made available online once the dates have been further adjusted to meet the ISO-8601 standard, to facilitate replication of this research. The NTR data are much better documented; however, reconciling the differences in methodology between the published NTR gene expression paper,³⁴⁷ its methods section, and its supplementary information still required a considerable amount of work.

Reproducibility

This thesis also represents an experiment in reproducible research. All data analysis has been done with the open source R language, and the thesis itself is written in R Markdown, with almost all code used to create the figures and run the analysis contained within the source file. The source file and related documents will ultimately be available via the Oxford University Research Archive (<https://ora.ox.ac.uk/>) along with raw data where possible. This approach is, however somewhat over-reliant on a given computational environment and is vulnerable to package incompatibilities and system updates. One way around this would be to develop the R Markdown document within a Docker container with a known reproducible computational environment.

There are, however, a number of disincentives to working at this level of reproducibility. I have developed this system for reproducible research independently in response to a perceived need. Despite being good practice, the adoption of reproducible methods gets little recognition, and working in this manner is time consuming. Ultimately reproducible work should be better encouraged, and institutions should actively work to develop and

enforce best practices rather than relying on individual *ad hoc* practices.

Postscript

“You think you’ve got problems,” said Marvin as if he was addressing a newly occupied coffin, “what are you supposed to do if you are a manically depressed robot? No, don’t bother to answer that, I’m fifty thousand times more intelligent than you and even I don’t know the answer. It gives me a headache just trying to think down to your level.”

Trillian burst in through the door from her cabin.

“My white mice have escaped!” she said.

An expression of deep worry and concern failed to cross either of Zaphod’s faces.

“Nuts to your white mice,” he said.

—From *The Hitchhiker’s Guide to the Galaxy*, Douglas Adams (1979)

References

- 1.Chakravarti, A. Perspectives on human variation through the lens of diversity and race. *Cold Spring Harbor Perspectives in Biology* **7**, a023358 (2015).
- 2.Leslie, S. *et al.* Fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
- 3.Pääbo, S. The human condition—a molecular approach. *Cell* **157**, 216–226 (2014).
- 4.Mayr, E. The biology of race and the concept of equality. *Daedalus* **131**, 89–94 (2002).
- 5.Epstein, D. *The Sports Gene*. (Current, 2013). at <<http://thesportsgene.com/books/the-sports-gene/>>
- 6.Mendel, G. Experiments in plant hybridization (EN transl. W. Bateson). *Journal of the Royal Horticultural Society* **26**, 3–47 (1901).
- 7.Taliaferro, W. H. & Huck, J. G. The inheritance of sickle-cell anaemia in man. *Genetics* **8**, 594–598 (1923).
- 8.International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- 9.Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
- 10.Visscher, P. M., Brown, M. a., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American Journal of Human Genetics* **90**, 7–24 (2012).
- 11.Maher, B. The case of the missing heritability. *Nature* **456**, 18–21 (2008).
- 12.Manolio, T. a *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
- 13.Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**, 446–450 (2010).

14. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**, 1114–1120 (2015).
15. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics* **96**, 377–385 (2015).
16. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111**, E5272–E5281 (2014).
17. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455–E464 (2014).
18. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).
19. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
20. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).
21. Bateson, P. Why are individuals so different from each other? *Heredity* **115**, 285–292 (2014).
22. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
23. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
24. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
25. Cotton, R. G. H. *et al.* The Human Variome Project. *Science* **322**, 861–862 (2008).

26. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Annals of Human Biology* **40**, 463–471 (2013).
27. Kilpinen, H. & Dermitzakis, E. T. Genetic and epigenetic contribution to complex traits. *Human Molecular Genetics* **21**, 24–28 (2012).
28. Lu, Y., Goldstein, D. B., Angrist, M., Cavalleri, G. & Cooper, R. S. Personalized medicine and human genetic diversity. *Cold Spring Harbor Perspectives in Medicine* **4**, a008581 (2014).
29. Davies, K. The era of genomic medicine. *Clinical Medicine* **13**, 594–601 (2013).
30. McCarthy, J. J., McLeod, H. L. & Ginsburg, G. S. Genomic medicine: a decade of successes , challenges , and opportunities. *Science Translational Medicine* **5**, 189sr4 (2013).
31. Manolio, T. a *et al.* Implementing genomic medicine in the clinic: the future is here. *Genetics in Medicine* **15**, 258–267 (2013).
32. IBM Watson for oncology. at <<http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html>>
33. The 100,000 Genomes Project. at <<http://www.genomicsengland.co.uk/the-100000-genomes-project/>>
34. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Research* **43**, D662–D669 (2014).
35. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996–1006 (2002).
36. Strong, L. C. in *Origins of inbred mice* (ed. Morse, H. C. I.) 45–67 (Academic Press, 1978). doi:[10.1016/B978-0-12-507850-4.50011-1](https://doi.org/10.1016/B978-0-12-507850-4.50011-1)
37. Kosova, G., Abney, M. & Ober, C. Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences* **107 Suppl**, 1772–1778 (2010).

38. Gagliardi, C., Falkenstein, K. P., Franke, D. E. & Kubisch, H. M. Estimates of heritability for reproductive traits in captive rhesus macaque females. *American Journal of Primatology* **72**, 811–819 (2010).
39. Cammack, K. M., Thomas, M. G. & Enns, R. M. Reproductive traits and their heritabilities in beef cattle. *Professional Animal Scientist* **25**, 517–528 (2009).
40. Peripato, A. C. *et al.* Epistasis affecting litter size in mice. *Journal of Evolutionary Biology* **17**, 593–602 (2004).
41. Schultz, N., Hamra, F. K. & Garbers, D. L. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proceedings of the National Academy of Sciences* **100**, 12201–12206 (2003).
42. Matzuk, M. M. & Lamb, D. J. Genetic dissection of mammalian fertility pathways. *Nat Cell Biol Fertility*, s41–s49 (2002).
43. Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. *Nature Medicine* **14**, 1197–1213 (2008).
44. Matzuk, M. M. & Lamb, D. J. The biology of infertility: supplementary data. *Nature Medicine* **14**, (2008).
45. Matzuk, M. M. & Burns, K. H. Genetics of mammalian reproduction: modeling the end of the germline. *Annual Review of Physiology* **74**, 503–528 (2012).
46. Zhao, H. & Chen, Z.-J. Genetic association studies in female reproduction: from candidate-gene approaches to genome-wide mapping. *Molecular Human Reproduction* **19**, 644–654 (2013).
47. Montgomery, G. W., Zondervan, K. T. & Nyholt, D. R. The future for genetic studies in reproduction. *Molecular Human Reproduction* **20**, 1–14 (2014).
48. Kosova, G., Scott, N. M., Niederberger, C., Prins, G. S. & Ober, C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *The*

American Journal of Human Genetics **90**, 950–961 (2012).

49.Sato, Y. *et al.* An association study of four candidate loci for human male fertility traits with male infertility. *Human Reproduction* **30**, 1510–1514 (2015).

50.Hu, Z. *et al.* A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia. *Nature Genetics* **44**, 183–186 (2011).

51.Yu, J. *et al.* Identification of seven genes essential for male fertility through a genome-wide association study of non-obstructive azoospermia and RNA interference-mediated large-scale functional screening in *Drosophila*. *Human Molecular Genetics* **24**, 1493–1503 (2014).

52.Massart, A., Lissens, W., Tournaye, H. & Stouffs, K. Genetic causes of spermatogenic failure. *Asian Journal of Andrology* **14**, 40–48 (2012).

53.Collins, J. *et al.* Genetic aspects of female reproduction. *Human Reproduction Update* **14**, 293–307 (2008).

54.Fauser, B. C. J. M. *et al.* Contemporary genetic technologies and female reproduction. *Human Reproduction Update* **17**, 829–847 (2011).

55.Fortuño, C. & Labarta, E. Genetics of primary ovarian insufficiency: a review. *Journal of Assisted Reproduction and Genetics* **31**, 1573–1585 (2014).

56.Griffin, D. K. & Finch, K. A. The genetic and cytogenetic basis of male infertility. *Human Fertility* **8**, 19–26 (2005).

57.O’Flynn O’Brien, K. L., Varghese, A. C. & Agarwal, A. The genetic causes of male factor infertility: a review. *Fertility and Sterility* **93**, 1–12 (2010).

58.Zheng, K., Yang, F. & Wang, P. J. Regulation of male fertility by X-linked genes. *Journal of Andrology* **31**, 79–85 (2010).

59.Ioannou, D. & Griffin, D. Male fertility, chromosome abnormalities, and nuclear organization. *Cytogenetic and Genome Research* **133**, 269–279 (2011).

- 60.Hwang, K. *et al.* Mendelian genetics of male infertility. *Annals of the New York Academy of Sciences* **1214**, E1–E17 (2010).
- 61.Jamsai, D. & O'Bryan, M. K. Mouse models in male fertility research. *Asian Journal of Andrology* **13**, 139–151 (2011).
- 62.Rajender, S., Avery, K. & Agarwal, A. Epigenetics, spermatogenesis and male infertility. *Mutation Research* **727**, 62–71 (2011).
- 63.Aston, K. I. & Carrell, D. T. Emerging evidence for the role of genomic instability in male factor infertility. *Systems Biology in Reproductive Medicine* **58**, 71–80 (2012).
- 64.Ogorevc, J., Dovc, P. & Kunej, T. Comparative genomics approach to identify candidate genetic loci for male fertility. *Reproduction in Domestic Animals* **46**, 229–239 (2011).
- 65.Dada, R. *et al.* Epigenetics and its role in male infertility. *Journal of Assisted Reproduction and Genetics* **29**, 213–223 (2012).
- 66.Davis, G. H. Major genes affecting ovulation rate in sheep. *Genetics, Selection, Evolution* **37 Suppl 1**, S11–S23 (2005).
- 67.Guénét, J., Benavides, F., Panthier, J. & Montagutelly, X. in *Genetics of the mouse* 1–17 (Springer-Verlag, 2015).
- 68.Lilue, J., Müller, U. B., Steinfeldt, T. & Howard, J. C. Reciprocal virulence and resistance polymorphism in the relationship between *Toxoplasma gondii* and the house mouse. *eLife* **2**, e01298 (2013).
- 69.Haldane, J. B. S., Sprunt, A. D. & Haldane, N. M. Reduplication in mice (Preliminary Communication). *Journal of Genetics* **5**, 133–135 (1915).
- 70.MGI-Guidelines for Nomenclature of Mouse and Rat Strains. at <http://www.informatics.jax.org/mgihome/nomen/strains.shtml/#inbred/_strains>
- 71.Guénét, J. L., Benavides, F., Panthier, J.-J. & Montagutelli, X. in *Genetics of the mouse* 319–359 (Springer Berlin Heidelberg, 2015). doi:[10.1007/978-3-662-44287-6](https://doi.org/10.1007/978-3-662-44287-6)

- 72.Silver, L. in *Mouse genetics: Concepts and applications* (Mouse Genome Informatics, 1995). at <<http://www.informatics.jax.org/silver/>>
- 73.Rader, K. *Making Mice: Standardizing Animals for American Biomedical Research, 1900-1955*. (Princeton University Press, 2004). at <<http://press.princeton.edu/titles/7787.html>>
- 74.Artzt, K. Mammalian developmental genetics in the twentieth century. *Genetics* **192**, 1151–1163 (2012).
- 75.JAX Mice Database - Inbred Strain. at <<http://jaxmice.jax.org/list/cat481365.html>>
- 76.Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nature Genetics* **24**, 23–25 (2000).
- 77.Beck, J., Lloyd, S., Hafezparast, M. &... Genealogies of mouse inbred strains: Poster. *Nature genetics* (2000). at <<ftp://ftp.informatics.jax.org/pub/datasets/misc/genealogy/genealogy.pdf>>
- 78.Chesler, E. J. Out of the bottleneck: the Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. *Mammalian Genome* **25**, 3–11 (2014).
- 79.Bronson, F. H., Dagg, C. P. & Snell, G. D. in *Biology of the laboratory mouse* (ed. Green, E. L.) (Dover, 1966). at <<http://www.informatics.jax.org/greenbook/index.shtml>>
- 80.in *Mouse genetics: Concepts and applications* (Mouse Genome Informatics, 1995). at <<http://www.informatics.jax.org/silver/contents.shtml>>
- 81.Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature Reviews Genetics* **10**, 783–796 (2009).
- 82.Falconer, D. S. The genetics of litter size in mice. *Journal of cellular and comparative physiology* **56 Suppl 1**, 153–167 (1960).
- 83.Dobzhansky, T. Experimenta on sexual isolation in *Drosophila*. X. Reproductive isolation between *Drosophila pseudoobscura* and *Drosophila persimilis* under natural

and under laboratory conditions. *Proceedings of the National Academy of Sciences* **37**, 792–796 (1951).

84.Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics* **12**, 101–109 (1922).

85.Turner, L. M., Schwahn, D. J. & Harr, B. Reduced male fertility is common but highly variable in form and severity in a natural mouse hybrid zone. *Evolution* **66**, 443–458 (2011).

86.Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C. & Forejt, J. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* **323**, 373–375 (2009).

87.Dzur-gejdosova, M., Simecek, P., Gregorova, S., Bhattacharyya, T. & Forejt, J. Dissecting the genetic architecture of F1 hybrid sterility in house mice. *Evolution* **66**, 3321–3335 (2012).

88.White, M. A., Steffy, B., Wiltshire, T. & Payseur, B. A. Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics* **189**, 289–304 (2011).

89.Bhattacharyya, T. *et al.* Mechanistic basis of infertility of mouse intersubspecific hybrids. *Proceedings of the National Academy of Sciences* **110**, E468–E477 (2013).

90.Bhattacharyya, T. *et al.* X chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids. *PLoS Genetics* **10**, e1004088 (2014).

91.Turner, L. M. & Harr, B. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *eLife* **3**, e02504 (2014).

92.Turner, L. M., White, M. a, Tautz, D. & Payseur, B. a. Genomic networks of hybrid sterility. *PLoS Genetics* **10**, e1004162 (2014).

93.Stubbings, M., Dumont, B. L., White, M. A. & Payseur, B. A. Genetics and evolution of hybrid male sterility in house mice. *Genetics* **191**, 917–934 (2012).

- 94.Elliott, R. W., Poslinski, D., Tabaczynski, D., Hohman, C. & Pazik, J. Loci affecting male fertility in hybrids between *Mus macedonicus* and C57BL/6. *Mammalian Genome* **15**, 704–710 (2004).
- 95.Oka, A., Mita, A., Takada, Y., Koseki, H. & Shiroishi, T. Reproductive isolation in hybrid mice due to spermatogenesis defects at three meiotic stages. *Genetics* **186**, 339–351 (2010).
- 96.Suzuki, T. a. & Nachman, M. W. Speciation and reduced hybrid female fertility in house mice. *Evolution* **69**, 2468–2481 (2015).
- 97.Latour, Y. *et al.* Sexual selection against natural hybrids may contribute to reinforcement in a house mouse hybrid zone. *Proceedings. Biological sciences / The Royal Society* **281**, 20132733 (2014).
- 98.Williams, R. W. *et al.* *A Collaborative Cross for high-precision complex trait analysis.* (Complex Trait Consortium, 2002). at <<http://www.complextait.org/workshop1.pdf>>
- 99.Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* **36**, 1133–1137 (2004).
- 100.Valdar, W., Flint, J. & Mott, R. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* **172**, 1783–1797 (2006).
- 101.Hall, M. *et al.* The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190**, 389–401 (2012).
- 102.Threadgill, D. W. & Churchill, G. A. Ten years of the collaborative cross. *G3 Genes, Genomes, Genetics* **2**, 153–156 (2012).
- 103.Yang, H., Bell, T. A., Churchill, G. A. & Villena, F. P.-m. D. On the subspecific origin of the laboratory mouse. *Nature Genetics* **39**, 1100–1107 (2007).
- 104.Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse.

Nature Genetics **43**, 648–655 (2011).

105.Roberts, A., Pardo-Manuel de Villena, F., Wang, W., McMillan, L. & Threadgill, D. W. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mammalian Genome* **18**, 473–481 (2007).

106.Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

107.Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mammalian Genome* **19**, 382–389 (2008).

108.Iraqi, F. A., Churchill, G. & Mott, R. The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. *Mammalian Genome* **19**, 379–381 (2008).

109.Morahan, G., Balmer, L. & Monley, D. Establishment of ‘The Gene Mine’: a resource for rapid identification of complex trait genes. *Mammalian Genome* **19**, 390–393 (2008).

110.Welsh, C. E. *et al.* Status and access to the Collaborative Cross population. *Mammalian Genome* **23**, 706–712 (2012).

111.Ferris, M. T. *et al.* Modeling host genetic regulation of influenza pathogenesis in the Collaborative Cross. *PLoS Pathogens* **9**, e1003196 (2013).

112.Bottomly, D. *et al.* Expression quantitative trait Loci for extreme host response to influenza a in pre-Collaborative Cross mice. *G3 Genes, Genomes, Genetics* **2**, 213–221 (2012).

113.Durrant, C. *et al.* Collaborative Cross mice and their power to map host susceptibility to *Aspergillus fumigatus* infection. *Genome Research* **21**, 1239–1248 (2011).

114.Vered, K., Durrant, C., Mott, R. & Iraqi, F. a. Susceptibility to *klebsiella pneumoniae*

infection in collaborative cross mice is a complex trait controlled by at least three loci acting at different time points. *BMC Genomics* **15**, 865 (2014).

115.Aylor, D. L. *et al.* Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research* **21**, 1213–1222 (2011).

116.Ram, R., Mehta, M., Balmer, L., Gatti, D. M. & Morahan, G. Rapid identification of major-effect genes using the Collaborative Cross. *Genetics* **198**, 75–86 (2014).

117.Mathes, W. F. *et al.* Architecture of energy balance traits in emerging lines of the Collaborative Cross. *American Journal of Physiology. Endocrinology and Metabolism* **300**, E1124–1134 (2011).

118.Kelada, D. S. N. P. *et al.* Integrative genetics of allergic inflammation in the murine lung. **51**, 436–445 (2014).

119.Weiser, M., Mukherjee, S. & Furey, T. S. Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations. *Genetics* **198**, 879–893 (2014).

120.Kelada, S. N. P. *et al.* Genetic analysis of hematological parameters in incipient lines of the Collaborative Cross. *G3 Genes, Genomes, Genetics* **2**, 157–165 (2012).

121.Thaisz, J. *et al.* Genetic analysis of albuminuria in collaborative cross and multiple mouse intercross populations. *American Journal of Physiology. Renal Physiology* **303**, F972–F981 (2012).

122.Phillippi, J. *et al.* Using the emerging Collaborative Cross to probe the immune system. *Genes and Immunity* **15**, 38–46 (2014).

123.Rutledge, H. *et al.* Genetic regulation of Zfp30, CXCL1, and neutrophilic inflammation in mouse lung. *Genetics* **198**, 735–745 (2014).

124.Philip, V. M. *et al.* Genetic analysis in the Collaborative Cross breeding population. *Genome Research* **21**, 1223–1238 (2011).

- 125.Rogala, A. R. *et al.* The Collaborative Cross as a resource for modeling human disease: CC011/Unc, a new mouse model for spontaneous colitis. *Mammalian Genome* **25**, 95–108 (2014).
- 126.Rasmussen, A. L. *et al.* Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. *Science* **346**, 987–991 (2014).
- 127.Graham, J. B. *et al.* Genetic diversity in the Collaborative Cross model recapitulates human West Nile Virus disease outcomes. **6**, e00493–15 (2015).
- 128.Levy, R., Mott, R. F., Iraqi, F. A. & Gabet, Y. Collaborative Cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. *BMC Genomics* **16**, 1013 (2015).
- 129.Mao, J.-H. *et al.* Identification of genetic factors that modify motor performance and body weight using Collaborative Cross mice. *Scientific Reports* **5**, 16247 (2015).
- 130.Gralinski, L. E. *et al.* Genome wide identification of SARS-CoV susceptibility loci using the Collaborative Cross. *PLOS Genetics* **11**, e1005504 (2015).
- 131.Didion, J. P. *et al.* A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLOS Genetics* **11**, e1004850 (2015).
- 132.Crowley, J. J. *et al.* Genetics of adverse reactions to haloperidol in a mouse diallel: a drug-placebo experiment and Bayesian causal analysis. *Genetics* **196**, 321–347 (2014).
- 133.Josset, L. *et al.* Annotation of long non-coding RNAs expressed in Collaborative Cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts. *RNA Biology* **11**, 37–41 (2014).
- 134.Xiong, H. *et al.* Genomic profiling of Collaborative Cross founder mice infected with respiratory viruses reveals novel transcripts and infection-related strain-specific gene and isoform expression. *G3 Genes, Genomes, Genetics* **4**, 1429–1444 (2014).
- 135.Chow, C. Y., Wang, X., Riccardi, D., Wolfner, M. F. & Clark, A. G. The genetic

- architecture of the genome-wide transcriptional response to ER stress in the mouse. *PLOS Genetics* **11**, e1004924 (2015).
136. Zheng, C. L. *et al.* Splicing landscape of the eight collaborative cross founder strains. *BMC Genomics* **16**, 1–11 (2015).
137. Xiao, H. *et al.* Genetically determined severity of anti-myeloperoxidase glomerulonephritis. *American Journal of Pathology* **182**, 1219–1226 (2013).
138. Patel, S. J., Molinolo, A. A., Gutkind, S. & Crawford, N. P. S. Germline genetic variation modulates tumor progression and metastasis in a mouse model of neuroendocrine prostate carcinoma. *PLoS ONE* **8**, (2013).
139. Zombeck, J. A., Deyoung, E. K., Brzezinska, W. J. & Rhodes, J. S. Selective breeding for increased home cage physical activity in collaborative cross and Hsd:ICR mice. *Behavior genetics* **41**, 571–582 (2011).
140. Gelinas, R. *et al.* A genetic approach to the prediction of drug side effects: bleomycin induces concordant phenotypes in mice of the collaborative cross. *Pharmacogenomics and Personalized Medicine* **4**, 35–45 (2011).
141. Kovacs, A. *et al.* Genotype is a stronger determinant than sex of the mouse gut microbiota. *Microbial Ecology* **61**, 423–428 (2011).
142. Campbell, J. H. *et al.* Host genetic and environmental effects on mouse intestinal microbiota. *ISME Journal* **6**, 2033–2044 (2012).
143. O'Connor, A., Quizon, P. M., Albright, J. E., Lin, F. T. & Bennett, B. J. Responsiveness of cardiometabolic-related microbiota to diet is influenced by host genetics. *Mammalian Genome* **25**, 583–599 (2014).
144. Ferguson, B. *et al.* Melanoma susceptibility as a complex trait: genetic variation controls all stages of tumor progression. *Oncogene* **34**, 2879–2886 (2015).
145. Iraqi, F. A. *et al.* Heritability and coefficient of genetic variation analyses of phenotypic

traits provide strong basis for high-resolution QTL mapping in the Collaborative Cross mouse genetic reference population. *Mammalian Genome* **25**, 109–119 (2014).

146.Shusterman, A. *et al.* Genotype is an important determinant factor of host susceptibility to periodontitis in the Collaborative Cross and inbred mouse populations. *BMC Genetics* **14**, 68 (2013).

147.Lorè, N. I., Iraqi, F. A. & Bragonzi, A. Host genetic diversity influences the severity of *Pseudomonas aeruginosa* pneumonia in the Collaborative Cross mice. *BMC Genetics* **16**, 106 (2015).

148.Scoville, D. K. *et al.* Susceptibility to quantum dot induced lung inflammation differs widely among the Collaborative Cross founder mouse strains. *Toxicology and Applied Pharmacology* **289**, 240–250 (2015).

149.Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**, 238–251 (2012).

150.Morgan, A. P. & Welsh, C. E. Informatics resources for the Collaborative Cross and related mouse populations. *Mammalian Genome* **26**, 521–539 (2015).

151.Bogue, M. a., Churchill, G. a. & Chesler, E. J. Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. *Mammalian Genome* **26**, 511–520 (2015).

152.Eisen, E. J. Single-trait and antagonistic index selection for litter size and body weight in mice. *Genetics* **88**, 781–811 (1978).

153.Canario, L., Roy, N., Gruand, J. & Bidanel, J. P. Genetic variation of farrowing kinetics traits and their relationships with litter size and perinatal mortality in French Large White sows. *Journal of Animal Science* **84**, 1053–1058 (2006).

154.Hare, E. & Leighton, E. A. Estimation of heritability of litter size in labrador retrievers

- and German shepherd dogs. *Journal of Veterinary Behavior: Clinical Applications and Research* **1**, 62–66 (2006).
- 155.Buis, R. Investigation of a selection limit for litter size in mice. *Livestock Production Science* **13**, 159–172 (1988).
- 156.Spearow, J. L. & Barkley, M. Genetic control of hormone-induced ovulation rate in mice. *Biology of Reproduction* **61**, 851–856 (1999).
- 157.Spearow, J. L., Nutson, P. a, Mailliard, W. S., Porter, M. & Barkley, M. Mapping genes that control hormone-induced ovulation rate in mice. *Biology of reproduction* **61**, 857–72 (1999).
- 158.Moore, R. K., Erickson, G. F. & Shimasaki, S. Are BMP-15 and GDF-9 primary determinants of ovulation quota in mammals? *Trends in Endocrinology and Metabolism: TEM* **15**, 356–361 (2004).
- 159.Liljander, M. *et al.* Identification of genetic regions of importance for reproductive performance in female mice. *Genetics* **173**, 901–909 (2006).
- 160.Liljander, M., Andersson, A., Holmdahl, R. & Mattsson, R. Increased litter size and super-ovulation rate in congenic C57BL mice carrying a polymorphic fragment of NFR/N origin at the Fecq4 locus of chromosome 9. *Genetics Research* **91**, 259–265 (2009).
- 161.Grubb, S. C., Bult, C. J. & Bogue, M. A. Mouse Phenome Database. *Nucleic Acids Research* **42**, 825–834 (2014).
- 162.Chesler EJ, Philip VM, V. B. Multi-system assessment of 8 inbred founder strains and 380+ emerging lines (pre-CC) of the Collaborative Cross. MPD:Chesler3. at <http://phenome.jax.org/db/q?rtn=projects/details/&sym=Chesler3/#docsumm>
- 163.Murray, S. A. *et al.* Mouse gestation length is genetically determined. *PLoS ONE* **5**, e12418 (2010).
- 164.Tukey, J. W. *Exploratory Data Analysis: Past, Present and Future.* (1993). at

<<http://oai.dtic.mil/oai/oai?verb=getRecord/&metadataPrefix=html/&identifier=ADA266775>>

165.R: The R Project for Statistical Computing. at <<https://www.r-project.org/>>

166.Tukey, J. W. *Exploratory Data Analysis*. 688 (Addison-Wesley, 1977). at <https://books.google.co.uk/books/about/Exploratory/_Data/_Analysis.html?id=UT9dAAAAIAAJ/&pgis=1>

167.Velleman, P. F. & Hoaglin, D. C. *Applications, basics, and computing of exploratory data analysis*. (The Internet-First University Press, 1981). at <http://dspace-local.library.cornell.edu/DSpace/IFUP/_Catalog.pdf>

168.Behrens, J. T. Principles and procedures of exploratory data analysis. *Psychological Methods* **2**, 131–160 (1997).

169.Pearson, R. *Exploring Data in Engineering, the Sciences, and Medicine*. (Oxford University Press, 2011). at <<https://books.google.com/books?id=nH4pAQAAMAAJ/&pgis=1>>

170.NIST/SEMATECH. in *E-handbook of statistical methods* (NIST, 2013). at <<http://www.itl.nist.gov/div898/handbook/>>

171.Seltman, H. J. in *Experimental design and analysis* 61–100 (2015). doi:[0201076160](https://doi.org/10.1007/978-1-4939-9832-7_4)

172.Peng, R. D. *Exploratory Data Analysis with R*. (2015).

173.Runner, M. *et al.* The genetics of litter size in mice [discussion on D Falconer 1960]. (1960).

174.Holt, M., Meuwissen, T. & Vangen, O. Long-term responses, changes in genetic variances and inbreeding depression from 122 generations of selection on increased litter size in mice. *Journal of Animal Breeding and Genetics* **122**, 199–209 (2005).

175.Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nature Methods* **6**, 663–666 (2009).

176. CompGen Tool Suite-CC genotyping. at <<http://csbio.unc.edu/CCstatus/index.py?run=Genotype>>
177. 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* **97**, 12649–12654 (2000).
178. Mott, R. Package ‘happy.hbrem’. (2013).
179. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–266 (2008).
180. Houle, D. Comparing evolvability and variability of quantitative traits. *Genetics* **130**, 195–204 (1992).
181. Garcia-Gonzalez, F., Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. & Evans, J. P. Comparing evolvabilities: common errors surrounding the calculation and use of coefficients of additive genetic variation. *Evolution* **66**, 2341–2349 (2012).
182. Mott, R. & Flint, J. Dissecting quantitative traits in mice. *Annual Review of Genomics and Human Genetics* **14**, 421–439 (2013).
183. Bates, D. lme4. (2015). at <<http://lme4.r-forge.r-project.org/>>
184. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, (2015).
185. Bates, D. M. *lme4: Mixed-Effects Modeling with R*. (Springer, 2010).
186. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
187. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols* **8**, 1551–1566 (2013).
188. Yan, W. *et al.* Mice deficient in oocyte-specific oligoadenylate synthetase-like protein OAS1D display reduced fertility. *Molecular and Cellular Biology* **25**, 4615–4624 (2005).

- 189.DeAngelis, A. M., Roy-O'Reilly, M. & Rodriguez, A. Genetic alterations affecting cholesterol metabolism and human fertility. *Biology of Reproduction* **91**, 117 (2014).
- 190.Yates, M., Kolmakova, A., Zhao, Y. & Rodriguez, A. Clinical impact of scavenger receptor class B type I gene polymorphisms on human female fertility. *Human Reproduction (Oxford)* **26**, 1910–1916 (2011).
- 191.Trigatti, B. *et al.* Influence of the high density lipoprotein receptor SR-BI on reproductive and cardiovascular pathophysiology. *Proceedings of the National Academy of Sciences* **96**, 9322–9327 (1999).
- 192.Miettinen, H. E., Rayburn, H. & Krieger, M. Abnormal lipoprotein metabolism and reversible female infertility in HDL receptor (SR-BI)-deficient mice. **108**, 1717–1722 (2001).
- 193.Yesilaltay, A. *et al.* Effects of hepatic expression of the high-density lipoprotein receptor SR-BI on lipoprotein metabolism and female fertility. *Endocrinology* **147**, 1577–1588 (2006).
- 194.Yesilaltay, A. *et al.* Excess cholesterol induces mouse egg activation and may cause female infertility. *Proceedings of the National Academy of Sciences* **111**, E4972–E4980 (2014).
- 195.Jiménez, L. M., Binelli, M., Bertolin, K., Pelletier, R. M. & Murphy, B. D. Scavenger receptor-B1 and luteal function in mice. *Journal of Lipid Research* **51**, 2362–2371 (2010).
- 196.Santander, N. G. *et al.* Developmental abnormalities in mouse embryos lacking the HDL receptor SR-BI. *Human Molecular Genetics* **22**, 1086–1096 (2013).
- 197.Zhang, W. *et al.* Genome-wide association mapping of quantitative traits in outbred mice. *G3 Genes, Genomes, Genetics* **2**, 167–174 (2012).
- 198.Svenson, K. L. *et al.* High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* **190**, 437–447 (2012).

- 199.Ewart-Toland, A., Mounzih, K., Qiu, J. & Chehab, F. F. Effect of the genetic background on the reproduction of leptin-deficient obese mice. *Endocrinology* **140**, 732–738 (1999).
- 200.Lobaccaro, J. M., Gallot, D., Lumbroso, S. & Mouzat, K. Liver X receptors and female reproduction: when cholesterol meets fertility! *Journal of Endocrinological Investigation* **36**, 55–60 (2012).
- 201.Baardman, M. E., Kerstjens-Frederikse, W. S., Berger, R. M. F. & Bakker, M. K. The role of maternal-fetal cholesterol transport in early fetal life: current insights. **88**, 24 (2013).
- 202.Montfoort, A. P. A. van, Plösch, T., Hoek, A. & Tietge, U. J. F. Impact of maternal cholesterol metabolism on ovarian follicle development and fertility. *Journal of Reproductive Immunology* **104–105**, 32–36 (2014).
- 203.L'Hôte, D. *et al.* Centimorgan-range one-step mapping of fertility traits using interspecific recombinant congenic mice. *Genetics* **176**, 1907–1921 (2007).
- 204.Reed, D. R., McDaniel, A. H., Avigdor, M. & Bachmanov, A. A. QTL for body composition on chromosome 7 detected using a chromosome substitution mouse strain. *Obesity (Silver Spring, Md.)* **16**, 483–487 (2008).
- 205.Muñoz, M. *et al.* Non-additive effects of RBP4, ESR1 and IGF2 polymorphisms on litter size at different parities in a Chinese-European porcine line. *Genetics, Selection, Evolution* **42**, 23 (2010).
- 206.Coster, A. *et al.* The imprinted gene DIO3 is a candidate gene for litter size in pigs. *PLoS ONE* **7**, e31825 (2012).
- 207.Hager, R. & Johnstone, R. A. Early experience and parent-of-origin-specific effects influence female reproductive success in mice. *Biology Letters* **2**, 253–256 (2006).
- 208.Milinski, M. The major histocompatibility complex, sexual selection and mate choice.

- Annual Review of Ecology Evolution and Systematics* **37**, 159–186 (2006).
209. Chaix, R., Cao, C. & Donnelly, P. Is mate choice in humans MHC-dependent? *PLoS Genetics* **4**, e1000184 (2008).
210. Wedekind, C. & Furi, S. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of The Royal Society London B* **264**, 1471–1479 (1997).
211. Manser, a, König, B. & Lindholm, a K. Female house mice avoid fertilization by t haplotype incompatible males in a mate choice experiment. *Journal of Evolutionary Biology* **28**, 54–64 (2014).
212. Odet, F. *et al.* The founder strains of the Collaborative Cross express a complex combination of advantageous and deleterious traits for male reproduction. *G3 Genes, Genomes, Genetics* **5**, 2671 (2015).
213. Pérez-Enciso, M. & Bidanel, J. Selection for litter size components: a critical review. *Genetics, Selection, Evolution* **29**, 483–496 (1997).
214. Eisen, E. J. & Durrant, B. S. Genetic and maternal environmental factors influencing litter size and reproductive efficiency in mice. *Journal of Animal Science* **50**, 428–441 (1980).
215. Pomp, D., Eisen, E. J. & Ziecik, A. J. LH receptor induction and ovulation rate in mice selected for litter size and body weight. *Journal of Reproduction and Fertility* **84**, 601–609 (1988).
216. Vangen, O. Results from 40 generations of divergent selection for litter size in mice. *Livestock Production Science* **37**, 197–211 (1993).
217. Holt, M., Vangen, O. & Farstad, W. Components of litter size in mice after 110 generations of selection. *Reproduction (Cambridge, England)* **127**, 587–92 (2004).
218. Langhammer, M. *et al.* High-fertility phenotypes: two outbred mouse models exhibit

substantially different molecular and physiological strategies warranting improved fertility. *Reproduction* **147**, 427–433 (2014).

219.Galloway, S. M. *et al.* Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nature Genetics* **25**, 279–283 (2000).

220.Yan, C. *et al.* Synergistic roles of bone morphogenetic protein 15 and growth differentiation factor 9 in ovarian function. *Molecular Endocrinology (Baltimore, Md.)* **15**, 854–866 (2001).

221.Otsuka, F., McTavish, K. J. & Shimasaki, S. Integral role of GDF-9 and BMP-15 in ovarian function. *Molecular Reproduction and Development* **78**, 9–21 (2011).

222.McIntosh, C. J., Lawrence, S., Smith, P., Juengel, J. L. & McNatty, K. P. Active immunization against the proregions of GDF9 or BMP15 alters ovulation rate and litter size in mice. *Reproduction (Cambridge, England)* **143**, 195–201 (2012).

223.Peng, J. *et al.* Growth differentiation factor 9:bone morphogenetic protein 15 heterodimers are potent regulators of ovarian functions. *Proceedings of the National Academy of Sciences* **110**, E776–E785 (2013).

224.Mottershead, D. G. *et al.* Cumulin, an oocyte-secreted heterodimer of the transforming growth factor-beta family, is a potent activator of granulosa cells and improves oocyte quality. *Journal of Biological Chemistry* **290**, 24007–24020 (2015).

225.Pu, D., Xing, Y., Gao, Y., Gu, L. & Wu, J. Gene variation and premature ovarian failure: a meta-analysis. *European Journal of Obstetrics Gynecology and Reproductive Biology* **182**, 226–237 (2014).

226.Fonseca, D. J. *et al.* BMP15 c.-9C>G promoter sequence variant may contribute to the cause of non-syndromic premature ovarian failure. *Reproductive BioMedicine Online* **29**, 627–633 (2014).

- 227.Su, Y.-Q. *et al.* Oocyte regulation of metabolic cooperativity between mouse cumulus cells and oocytes: BMP15 and GDF9 control cholesterol biosynthesis in cumulus cells. *Development (Cambridge, England)* **135**, 111–121 (2008).
- 228.Kirkpatrick, B. W., Mengelt, A., Schulman, N. & Martin, I. C. A. Identification of quantitative trait loci for prolificacy and growth in mice. *Mammalian Genome* **102**, 97–102 (1998).
- 229.Rocha, J. L., Eisen, E. J., Siewerdt, F., Van Vleck, L. D. & Pomp, D. A large-sample QTL study in mice: III. Reproduction. *Mammalian Genome* **15**, 878–886 (2004).
- 230.Wei, J. *et al.* An integrative genomic analysis of the superior fecundity phenotype in QSi5 mice. *Molecular Biotechnology* **53**, 217–226 (2013).
- 231.Suto, J.-I. Genetic analysis of litter size in mice. *Journal of Veterinary Medical Science* **77**, 353–358 (2015).
- 232.Spearow, J. L. Characterization of genetic differences in hormone-induced ovulation rate in mice. *Journal of Reproduction and Fertility* **82**, 799–806 (1988).
- 233.McCarthy, J. C. The effects of inbreeding on the components of litter size in mice. *Genetic Research, Cambridge* **10**, 73–80 (1967).
- 234.Good, J. M., Dean, M. D. & Nachman, M. W. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* **179**, 2213–2228 (2008).
- 235.Good, J. M., Handel, M. A. & Nachman, M. W. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution; International Journal of Organic Evolution* **62**, 50–65 (2008).
- 236.Flachs, P. *et al.* Interallelic and intergenic incompatibilities of the Prdm9 (Hst1) gene in mouse hybrid sterility. *PLoS Genetics* **8**, e1003044 (2012).
- 237.Williams, R. W. & Churchill, G. A. *The Collaborative Cross: rationale, implementation, and costs.* (2004).

- 238.Bohacek, J. & Mansuy, I. M. Epigenetic inheritance of disease and disease risk. *Neuropsychopharmacology* **38**, 220–236 (2013).
- 239.Trerotola, M., Relli, V., Simeone, P. & Alberti, S. Epigenetic inheritance and the missing heritability. *Human Genomics* **9**, 17 (2015).
- 240.Bird, A. Perceptions of epigenetics. *Nature* **447**, 396–398 (2007).
- 241.Grossniklaus, U. *et al.* Transgenerational epigenetic inheritance: how important is it? *Nature Reviews Genetics* **14**, 228–235 (2013).
- 242.Iqbal, K. *et al.* Deleterious effects of endocrine disruptors are corrected in the mammalian germline by epigenome reprogramming. *Genome Biology* **16**, 59 (2015).
- 243.Whitelaw, E. Disputing Lamarckian epigenetic inheritance in mammals. *Genome Biology* **16**, 60 (2015).
- 244.Nadeau, J. H. The nature of evidence for and against epigenetic inheritance. *Genome Biology* **16**, 137 (2015).
- 245.Szabó, P. E. Response to: the nature of evidence for and against epigenetic inheritance. *Genome Biology* **16**, 138 (2015).
- 246.Slatkin, M. Epigenetic inheritance and the missing heritability problem. *Genetics* **182**, 845–850 (2009).
- 247.Cortijo, S. *et al.* Mapping the epigenetic basis of complex traits. *Science* **343**, 1145–1148 (2014).
- 248.Guilmatre, A. & Sharp, A. J. Parent of origin effects. *Clinical Genetics* **81**, 201–209 (2012).
- 249.Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nature Reviews Genetics* **14**, 609–617 (2013).
- 250.Savory, T. H. The Mule. *Scientific American* (1970).

- 251.Loke, Y. *Life's Vital Link: The Astonishing Role of the Placenta*. 288 (OUP Oxford, 2013). at <https://books.google.com/books?id=_19oAgAAQBAJ/&pgis=1>
- 252.Walton, A. & Hammond, J. The maternal effects of growth and conformation in shire horse-shetland pony crosses. *Proceedings of the Royal Society B* **125**, 311–335 (1938).
- 253.Vrana, P. B. *et al.* Genetic and epigenetic incompatibilities underlie hybrid dysgenesis in *Peromyscus*. *Nature Genetics* **25**, 120–124 (2000).
- 254.Bruining, H. *et al.* The parent-of-origin of the extra X chromosome may differentially affect psychopathology in klinefelter syndrome. *Biological Psychiatry* **68**, 1156–1162 (2010).
- 255.Knoll, J. *et al.* Angelman and Prader-Willi Syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *American Journal of Medical Genetics* **32**, 285–290 (1989).
- 256.Dittrich, B. *et al.* Molecular diagnosis of the Prader-Willi and Angelman syndromes by detection of parent-of-origin specific DNA methylation in 15q11-13. *Human Genetics* **90**, 313–315 (1992).
- 257.Pérez Jurado, L. A., Peoples, R., Kaplan, P., Hamel, B. C. & Francke, U. Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. *American Journal of Human Genetics* **59**, 781–792 (1996).
- 258.Collette, J. C. *et al.* William's syndrome: gene expression is related to parental origin and regional coordinate control. *Journal of Human Genetics* **54**, 193–198 (2009).
- 259.Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
- 260.Lindsay, R. S., Kobes, S., Knowler, W. C., Bennett, P. H. & Hanson, R. L. Genome-wide linkage analysis assessing parent-of-origin effects in the inheritance of type 2 diabetes and BMI in Pima Indians. *Diabetes* **50**, 2850–2857 (2001).

- 261.Hanson, R. L. *et al.* Strong parent-of-origin effects in the association of KCNQ1 variants with type 2 diabetes in American Indians. *Diabetes* **62**, 2984–91 (2013).
- 262.Guo, Y.-F. *et al.* Assessment of genetic linkage and parent-of-origin effects on obesity. *The Journal of Clinical Endocrinology and Metabolism* **91**, 4001–4005 (2006).
- 263.Souren, N. Y. *et al.* Parent-of-origin specific linkage and association of the IGF2 gene region with birth weight and adult metabolic risk factors. *International Journal of Obesity* **33**, 962–970 (2009).
- 264.Hoggart, C. J. *et al.* Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS Genetics* **10**, e1004508 (2014).
- 265.Hochner, H. *et al.* Parent-of-origin effects of the APOB gene on adiposity in young adults. *PLOS Genetics* **11**, e1005573 (2015).
- 266.Hoekstra, A. S., Devilee, P. & Bayley, J.-P. Models of parent-of-origin tumorigenesis in hereditary paraganglioma. *Seminars in Cell & Developmental Biology* **43**, 117–124 (2015).
- 267.Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
- 268.Ebers, G. C. *et al.* Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *Lancet* **363**, 1773–1774 (2004).
- 269.Ramagopalan, S. V. *et al.* Parent-of-origin effect in multiple sclerosis. *Neurology* **73**, 602–605 (2009).
- 270.Chao, M. J. *et al.* Parent-of-origin effects at the major histocompatibility complex in multiple sclerosis. *Human Molecular Genetics* **19**, 3679–3689 (2010).
- 271.Nudel, R. *et al.* Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes, Brain, and Behavior* **13**,

418–429 (2014).

272.Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* **2**, 21–32 (2001).

273.Wood, A. J. & Oakey, R. J. Genomic imprinting in mammals: emerging themes and established theories. *PLoS Genetics* **2**, e147 (2006).

274.Bartolomei, M. S. & Ferguson-Smith, A. C. Mammalian genomic imprinting. *Cold Spring Harbor Perspectives in Biology* **3**, a002592 (2011).

275.Barlow, D. P. & Bartolomei, M. S. Genomic imprinting in mammals. *Cold Spring Harbor Perspectives in Biology* **6**, 018382 (2014).

276.Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics* **15**, 517–530 (2014).

277.McGrath, J. & Solter, D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell* **37**, 179–183 (1984).

278.Surani, M. a, Barton, S. C. & Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* **308**, 548–550 (1984).

279.Cooper DW, VandeBerg JL, Sharman GB, P. W. Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. *Nature New Biology* **230**, 155–157 (1971).

280.Barlow, D. P., Stöger, R., Herrmann, B. G., Saito, K. & Schweifer, N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* **349**, 84–87 (1991).

281.DeChiara, T. M., Robertson, E. J. & Efstratiadis, A. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell* **64**, 849–859 (1991).

282.Ferguson-Smith, A. C., Cattanaach, B. M., Barton, S. C., Beechey, C. V. & Surani, M. A. Embryological and molecular investigations of parental imprinting on mouse

- chromosome 7. *Nature* **351**, 667–670 (1991).
283. Bartolomei, M. S., Zemel, S. & Tilghman, S. M. Parental imprinting of the mouse H19 gene. *Nature* **351**, 153–155 (1991).
284. Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Reviews Genetics* **12**, 565–575 (2011).
285. Gregg, C., Zhang, J., Butler, J. E., Haig, D. & Dulac, C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**, 682–685 (2010).
286. DeVeale, B., Kooy, D. van der & Babak, T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genetics* **8**, e1002600 (2012).
287. Kelsey, G. & Bartolomei, M. S. Imprinted genes... and the number is? *PLoS Genetics* **8**, e1002601 (2012).
288. Calabrese, J. M., Starmer, J., Schertzer, M. D., Yee, D. & Magnuson, T. A survey of imprinted gene expression in mouse trophoblast stem cells. *G3 Genes Genomes Genetics* **5**, 751–759 (2015).
289. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Research* **25**, 927–936 (2015).
290. Babak, T. *et al.* Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature Genetics* **47**, 544–549 (2015).
291. Imprinting Resource | MouseBook. at <<http://www.mousebook.org/mousebook-catalogs/imprinting-resource>>
292. Geneimprint : Home. at <<http://www.geneimprint.com/>>
293. WAMIDEX: a web atlas of murine genomic imprinting and differential expression. at <<https://atlas.genetics.kcl.ac.uk/atlas.php>>
294. Schulz, R. *et al.* WAMIDEX: a web atlas of murine genomic imprinting and differential expression. *Epigenetics* **3**, 89–96 (2008).

- 295.Morison, I. Catalogue of Imprinted Genes and Parent-of-origin Effects in Humans and Animals. at <<http://igc.otago.ac.nz/home.html>>
- 296.Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Research* **29**, 275–276 (2001).
- 297.Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends in Genetics* **21**, 457–465 (2005).
- 298.Glaser, R. L., Ramsay, J. P. & Morison, I. M. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucleic Acids Research* **34**, D29–D31 (2006).
- 299.Prickett, A. R. & Oakey, R. J. A survey of tissue-specific genomic imprinting in mammals. *Molecular Genetics and Genomics* **287**, 621–630 (2012).
- 300.Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nature Reviews Genetics* **16**, 653–64 (2015).
- 301.Chess, A. Random and non-random monoallelic expression. *Neuropsychopharmacology* **38**, 55–61 (2012).
- 302.Chess, A. Mechanisms and consequences of widespread random monoallelic expression. *Nature Reviews Genetics* **13**, 421–428 (2012).
- 303.Kelsey, G. & Feil, R. New insights into establishment and maintenance of DNA methylation imprints in mammals New insights into establishment and maintenance of DNA methylation imprints in mammals. *Philosophical Transactions of the Royal Society B* 20110336 (2013).
- 304.Hanna, C. W. & Kelsey, G. The specification of imprints in mammals. *Heredity* **113**, 176–183 (2014).
- 305.Cleaton, M. A. M., Edwards, C. A. & Ferguson-Smith, A. C. Phenotypic outcomes of

- imprinted gene models in mice: elucidation of pre- and postnatal functions of imprinted genes. *Annual Review of Genomics and Human Genetics* **15**, 93–126 (2014).
- 306.Tomizawa, S.-I. *et al.* Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**, 811–820 (2011).
- 307.Keverne, E. B. & Curley, J. P. Epigenetics, brain evolution and behaviour. *Frontiers in Neuroendocrinology* **29**, 398–412 (2008).
- 308.Keverne, E. B. Importance of the matriline for genomic imprinting, brain development and behaviour. *Philosophical Transactions of the Royal Society B* **368**, 20110327 (2013).
- 309.Moore, T. & Haig, D. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics* **7**, 45–49 (1991).
- 310.Haig, D. The kinship theory of genomic imprinting. *Annual Review of Ecology Evolution and Systematics* 9–32 (2000).
- 311.Haig, D. Coadaptation and conflict, misconception and muddle, in the evolution of genomic imprinting. *Heredity* **113**, 96–103 (2014).
- 312.Tunster, S. J., Jensen, A. B. & John, R. M. Imprinted genes in mouse placental development and the regulation of fetal energy stores. *Reproduction* **145**, R117–R137 (2013).
- 313.Wang, X., Miller, D. C., Harman, R., Antczak, D. F. & Clark, a. G. Paternally expressed genes predominate in the placenta. *Proceedings of the National Academy of Sciences* **110**, 10705–10710 (2013).
- 314.Perez, J. D. *et al.* Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. *eLIFE* **4**, e07860 (2015).
- 315.Varmuza, S. & Mann, M. Genomic imprinting—defusing the ovarian time bomb. *Trends in Genetics* **10**, 118–123 (1994).

- 316.Day, T. Intralocus sexual conflict can drive the evolution of genomic imprinting. *Genetics* **167**, 1537–1546 (2004).
- 317.Wolf, J. B. & Hager, R. A maternal-offspring coadaptation theory for the evolution of genomic imprinting. *PLoS Biology* **4**, 2238–2243 (2006).
- 318.Spencer, H. G. & Clark, A. G. Non-conflict theories for the evolution of genomic imprinting. *Heredity* **113**, 112–118 (2014).
- 319.Patten, M. M. *et al.* The evolution of genomic imprinting: theories, predictions and empirical tests. *Heredity* **113**, 119–128 (2014).
- 320.Wolf, J. B. & Brandvain, Y. Gene interactions in the evolution of genomic imprinting. *Heredity* **113**, 129–137 (2014).
- 321.Butler, M. G. Genomic imprinting disorders in humans: a mini-review. *Journal of Assisted Reproduction and Genetics* **26**, 477–486 (2009).
- 322.Tomizawa, S.-I. & Sasaki, H. Genomic imprinting and its relevance to congenital disease, infertility, molar pregnancy and induced pluripotent stem cell. *Journal of Human Genetics* **57**, 84–91 (2012).
- 323.Ishida, M. & Moore, G. E. The role of imprinted genes in humans. *Molecular Aspects of Medicine* **34**, 826–840 (2013).
- 324.Kalish, J. M., Jiang, C. & Bartolomei, M. S. Epigenetics and imprinting in human disease. *International Journal of Developmental Biology* **58**, 291–298 (2014).
- 325.Lazaraviciute, G., Kauser, M., Bhattacharya, S., Haggarty, P. & Bhattacharya, S. A systematic review and meta-analysis of DNA methylation levels and imprinting disorders in children conceived by IVF / ICSI compared with children conceived spontaneously. **20**, 840–852 (2014).
- 326.Hu, Y., Rosa, G. J. M. & Gianola, D. A GWAS assessment of the contribution of genomic imprinting to the variation of body mass index in mice. *BMC Genomics* **16**,

576 (2015).

327.Wolf, J. B. & Wade, M. J. What are maternal effects (and what are they not)? *Philosophical Transactions of the Royal Society B* **364**, 1107–1115 (2009).

328.Kim, K.-H. & Lee, K.-A. Maternal effect genes: findings and effects on mouse embryo development embryonic arrest. *Clinical and Experimental Reproductive Medicine* **41**, 47–61 (2014).

329.Carlier, M., Roubertoux, P. L. & Pastoret, C. The Y chromosome effect on intermale aggression in mice depends on the maternal environment. *Genetics* **129**, 231–236 (1991).

330.Wolff, G. L., Kodell, R. L., Moore, S. R. & Cooney, C. A. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB Journal* **12**, 949–957 (1998).

331.Heijmans, B. T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences* **105**, 17046–17049 (2008).

332.Gleason, G. *et al.* The serotonin 1A receptor gene as a genetic and prenatal maternal environmental factor in anxiety. *Proceedings of the National Academy of Sciences* **107**, 7592–7597 (2010).

333.Godfrey, K. M. *et al.* Epigenetic gene promoter methylation at birth is associated with child's later adiposity. *Diabetes* **60**, 1528–1534 (2011).

334.Hager, R., Cheverud, J. M. & Wolf, J. B. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics* **178**, 1755–1762 (2008).

335.Babak, T. *et al.* Global survey of genomic imprinting by transcriptome sequencing. *Current Biology* **18**, 1735–1741 (2008).

336.Wang, X. & Clark, A. G. Using next-generation RNA sequencing to identify imprinted genes. *Heredity* **113**, 156–166 (2014).

- 337.Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* **47**, 353–360 (2015).
- 338.Luedi, P. *et al.* Computational and experimental identification of novel human imprinted genes. *Genome Research* **17**, 1723–1730 (2007).
- 339.Connolly, S. & Heron, E. a. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Briefings in Bioinformatics* **16**, 429–448 (2015).
- 340.Feng, R., Wu, Y., Jang, G. H., Ordovas, J. M. & Arnett, D. A powerful test of parent-of-origin effects for quantitative traits using haplotypes. *PLoS ONE* **6**, (2011).
- 341.Howey, R. *et al.* Increased power for detection of parent-of-origin effects via the use of haplotype estimation. *American Journal of Human Genetics* **97**, 419–434 (2015).
- 342.Mott, R. *et al.* The architecture of parent-of-origin effects in mice. *Cell* **156**, 332–342 (2014).
- 343.Smith, Z. *White Teeth*. (Hamish Hamilton, 2000). at <https://books.google.co.uk/books/about/White/_Teeth.html?id=BJtud-Wbt9YC/&pgis=1>
- 344.Beijsterveldt, C. E. M. van *et al.* The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics* **16**, 252–267 (2013).
- 345.Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin research and human genetics* **16**, 271–281 (2013).
- 346.Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics* **13**, 231–245 (2010).
- 347.Wright, F. a *et al.* Heritability and genomics of gene expression in peripheral blood. *Nature Genetics* **46**, 430–437 (2014).

- 348.Boomsma, D. I., Orlebeke, J. F. & Van Baal, G. C. M. The Dutch Twin Register: growth data on weight and height. *Behavior Genetics* **22**, 247–251 (1992).
- 349.Boomsma, D. I. *et al.* Netherlands Twin Register: a focus on longitudinal research. **5**, 401–406 (2002).
- 350.Boomsma, D. I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* **9**, 849–857 (2006).
- 351.Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* 1–7 (2013). doi:[10.1038/ejhg.2013.118](https://doi.org/10.1038/ejhg.2013.118)
- 352.Boomsma, D. I. *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *European Journal of Human Genetics* **16**, 335–342 (2008).
- 353.Clayton, D. Package ‘snpStats’. (2015).
- 354.Clayton, D. *Data input vignette reading genotype data in snpStats*. (2014).
- 355.Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575 (2007).
- 356.Wright, F. a *et al.* Heritability and genomics of gene expression in peripheral blood. Supplementary information. *Nature Genetics* **46**, 430–437 (2014).
- 357.Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
- 358.Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research* **22**, 1760–1774 (2012).
- 359.Gelman, A. & Hill, J. in *Data analysis using regression and multilevel/Hierarchical models* 529–546 (2006).

- 360.Honaker, J., King, G. & Blackwell, M. AMELIA II: a program for missing data. *Journal Of Statistical Software* **45**, 1–54 (2011).
- 361.Honaker, A. J., King, G., Blackwell, M. & Blackwell, M. M. Package ‘Amelia’. (2015).
- 362.Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, e161 (2007).
- 363.Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703–714 (2011).
- 364.Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73**, 1162–1169 (2003).
- 365.Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084–1097 (2007).
- 366.MACH 1.0 - Markov Chain Haplotyping. at <<http://csg.sph.umich.edu//abecasis/MaCH/>>
- 367.Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* **5**, e1000529 (2009).
- 368.Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2011).
- 369.Livne, O. E. *et al.* PRIMAL: fast and accurate pedigree-based imputation from sequence data in a founder population. *PLOS Computational Biology* **11**, e1004139 (2015).
- 370.SHAPEIT. at <https://mathgen.stats.ox.ac.uk/genetics/_software/shapeit/shapeit.html>

371. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).
372. O’Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* **10**, e1004234 (2014).
373. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955–959 (2012).
374. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5**, 3934 (2014).
375. 1000GP Phase 1 haplotypes 9 Dec 2013. at <https://mathgen.stats.ox.ac.uk/impute/data/_download/_1000G/_phase1/_integrated/_SHAPEIT2/_9-12-13.html>
376. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *American Journal of Human Genetics* **93**, 687–696 (2013).
377. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
378. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
379. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
380. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818–825 (2014).
381. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* **22**, 221–227 (2014).
382. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in

European samples using the 'Genome of The Netherlands'. *European Journal of Human Genetics* **22**, 1321–6 (2014).

383.DuoHMM. at <https://mathgen.stats.ox.ac.uk/genetics/_software/duohmm/duohmm.html>

384.Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics* **11**, 800–805 (2010).

385.Lawson, D. J. & Falush, D. Population identification using genetic data. *Annual Review of Genomics and Human Genetics* **13**, 337–361 (2012).

386.Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**, 1011–1021 (2012).

387.Vinkhuyzen, A. A., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and partitioning of heritability in human populations using whole genome analysis methods. *Annu Rev Genet* **47**, 75–95 (2013).

388.Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16**, 33–44 (2014).

389.Forni, S., Aguilar, I. & Misztal, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics, Selection, Evolution* **43**, 1 (2011).

390.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).

391.Price, A. L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics* **7**, e1001317 (2011).

392.Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics* **44**, 1084–1089 (2012).

393. Yang, S. *et al.* Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics* **15**, 13 (2014).
394. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519–525 (2011).
395. Visscher, P. M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104–1110 (2007).
396. Torres, J. M. *et al.* Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *American Journal of Human Genetics* **95**, 521–534 (2014).
397. Fedko, I. O. *et al.* Estimation of genetic relationships between individuals across cohorts and platforms: application to childhood height. *Behavior Genetics* **45**, 514–528 (2015).
398. Visscher, P. M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genetics* **10**, e1004269 (2014).
399. GCTA-GREML Power Calculator. at <<http://cnsgenomics.com/shiny/gctaPower/>>
400. Krishna Kumar, S., Feldman, M. W., Rehkopf, D. H. & Tuljapurkar, S. Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences* **113**, E61–E70 (2016).
401. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Commentary on ‘Limitations of GCTA as a solution to the missing heritability problem’. *bioRxiv* **1**, (2015).
402. Gamazon, E. R. & Park, D. S. SNP-based heritability estimation: measurement noise, population stratification and stability. *bioRxiv* (2016).
403. Kumar, S. K., Feldman, M. W., Rehkopf, D. H. & Tuljapurkar, S. Response to ‘Commentary on limitations of GCTA as a solution to the missing heritability problem’. *bioRxiv* (2016). doi:[10.1073/pnas.1600634113](https://doi.org/10.1073/pnas.1600634113)

- 404.Jansen, R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC Genomics* **15**, 33 (2014).
- 405.Shen, S. Q., Turro, E. & Corbo, J. C. Hybrid mice reveal parent-of-origin and Cis- and Trans-regulatory effects in the retina. *PLoS ONE* **9**, e109382 (2014).
- 406.Tukey, J. W. We need both exploratory and confirmatory. *The American Statistician* **34**, 23–25 (1980).
- 407.Yalcin, B., Flint, J. & Mott, R. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* **171**, 673–681 (2005).
- 408.Alon, U. How to choose a good scientific problem. *Molecular cell* **35**, 726–8 (2009).
- 409.White, E. P. *et al.* Nine simple ways to make it easier to (re)use your data. *PeerJ PrePrints* **1**, e7v2 (2013).
- 410.Wickham, H. Tidy data. *Journal of Statistical Software* **59**, 1–23 (2014).