Quantitative Trait Analysis in a Panel of Recombinant Inbred Rat Strains

Ian C. Grieve

Thesis submitted to Imperial College
for the degree of Doctor of Philosophy

April 2010

Physiological Genomics and Medicine Group
Division of Clinical Sciences Centre
Imperial College Faculty of Medicine
Hammersmith Hospital
Du Cane Road
London W12 0NN

Abstract

Expression quantitative trait loci (eQTLs) are generated by the combined use of microarray technology to measure gene expression and genetic linkage analysis to map the expression traits to the genome.

This thesis describes co-expression and quantitative trait transcript (QTT) analysis carried out on a dataset consisting of thousands of *cis*- and *trans*-eQTLs. These were mapped in 29 rat Recombinant Inbred strains derived from a cross between the Spontaneously Hypertensive Rat (SHR), a widely used model of the human metabolic syndrome, and the normotensive Brown Norway (BN). Gene expression data from four tissues relevant to the metabolic syndrome and cardiovascular disease were analysed: fat, kidney, adrenal gland and left ventricle.

By systematically applying a rigorous statistical methodology to the eQTL dataset, a consistent, distinct correlation structure was observed. Co-expression of groups of transcripts linked to a common region of the genome, referred to as *trans*-eQTL clusters, was investigated. Some of these cluster-forming groups were found to remain significantly correlated after the effect of genotype was accounted for, and functionally enriched.

An example of successful application of QTT analysis to the dataset is described. This contributed to the identification of *Ogn* as a regulator of left ventricular mass in

rodents and subsequent implication of the homologue of *Ogn* in a related role in humans. Correlation of a further 103 physiological traits with *cis*-eQTLs in each of the four tissues was also carried out; analysis which potentially informs a wide range of hypotheses concerning relevant phenotypes.

Together, the findings described here demonstrate the utility of a systematic computational approach using correlation-based methodologies in combination with appropriate statistical techniques to inform the genetic analysis of complex traits. These findings indicate the importance of understanding potential confounding factors in eQTL analysis, as well as the potential of the eQTL approach to stimulate gene discovery.

Acknowledgements

I am very grateful to The Wellcome Trust for generously funding this studentship (069962/Z/02/Z).

I would like to thank Professor Tim Aitman, Dr Enrico Petretto and Prof. Michael Stumpf for their supervision during my PhD.

I thank fellow members of the Physiological Genomics and Medicine group at the MRC Clinical Sciences Centre at Hammersmith Hospital for their moral and practical support at various stages of my PhD. Nick, Riz, Anuj, Liqun, Santosh, Manuela, Michelle, Catherine, Jonathan Mangion, and Helen Figueira – thanks so much to all of you.

Additionally, I'd like to acknowledge Prof. Angela McLean of the Department of Zoology, Oxford University for her support during my undergraduate degree. Without you I would not have made it this far in my academic career.

Last but certainly not least, I thank my parents for always being there for me every step of the way.

Preface

Elements of this thesis have been published, as follows:

Elements of the work described in **Chapter 4** have been published as:

Grieve, I.C., Dickens, N.J., Pravenec, M., Kren, V., Hubner, N., Cook, S.A., Aitman, T.J., Petretto, E., Mangion, J. (2008). 'Genome-wide Co-expression Analysis in Multiple Tissues'. *PLoS ONE* **3**(12):e4033

Work described in **Chapter 5** contributed to:

Petretto, E., Sarwar, R., **Grieve, I.C.**, Lu, H., Kumaran, M.K., Muckett, P.J., Mangion, J., Shroen, B., Benson, M. *et al* (2008). 'Integrated Genomic Approaches Implicate Osteoglycin (*Ogn*) in the Regulation of Left Ventricular Mass'. *Nature Genetics* **40**(5): 546-552

Statement of Contributions

All of the work described in the results chapters of this thesis was carried out by Ian Grieve. Where data was generated by others or project work was carried out in concert with external collaborators or other members of the Physiological Genomics & Medicine Group, their contributions are detailed below:

1.6 Generation of Expression Quantitative Trait Loci

The eQTL datasets were generated using Affymetrix GeneChip microarrays by Tim Aitman, Norbert Hubner and Stuart Cook.

3.1 Validation of eQTL Data

'Missing genotypes analysis' was initiated and the interpretation of the analysis assisted by Michelle Johnson. Michelle carried out the initial comparison between eQTL Reaper and MapManager for a small set of probesets, and supplied the data for ***Figure 3.1***. SCAMPA tiering (described in **2.2.2.1**) of the eQTL probesets was carried out by Nicholas Dickens.

3.2 Comparative Sequence Analysis of *cis*-eQTL Probesets

De novo sequencing of *cis*-eQTL probesets used in this analysis was performed by Michelle Johnson, Catherine Morrissey, and Chris McDermott-Roe. Sequence data provided by Fiona MacIver was also used.

3.3 Assessment of software for predicting potential 'false *cis*-eQTLs'

SHR sequencing was performed by the SHR sequence consortium (described in **2.3.1**) and data was made available for this analysis by Santosh Atanur. MAQ mapping of SNPs to the BN reference sequence (described in **2.3.2**) was performed by Liqun He and Santosh Atanur.

5.2.1 Quantitative Trait Transcript Analysis of Left Ventricular Mass and Related Traits

The expression data for Figures 5.1-5.3 were generated by the laboratory of Stuart Cook, Molecular Cardiology (CSC). Physiological trait data was supplied by Michal Pravenec. The Quantitative Trait Transcript analysis was performed and interpreted, and the plots produced, by Ian Grieve. Data from human subjects was obtained by MRI by Sanjay Prasad, Dudley Pennell and Chris Kiesewetter. Human expression data was generated by Yigal Pinto and collaborators. Quantitative Trait Transcript analyses, the results of which are shown in Table 5.2 and Figures 5.4-5.5 were carried out by Ian Grieve.

5.2.2 Quantitative Trait Transcript Analysis of non LVM-related Traits

Physiological trait data used were supplied by Daniel O'Connor of UCSD (adrenal serum and protein traits), Theodore Kurtz & Michal Pravenec (blood pressure, heart rate and pulse pressure traits), and Michal Pravenec (metabolic traits).

Table of Contents

List of Figures

List of Tables

Abbreviations

3’UTR	3-prime Untranslated Region
ACTH	Adrenocorticotropic Hormone
ADHD	Attention Deficit Hyperactivity Disorder
ANOVA	Analysis of Variance
API	Application Programming Interface
ASAT	Aspartate Aminotransferase
BN	Brown Norway (rat strain)
BP	Blood Pressure
cM	CentiMorgan
CgA	Chromogranin A
CgB	Chromogranin B
CHD	Coronary Heart Disease
CI	Confidence Interval
CRAN	Comprehensive R Archive Network
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
DAVID	Database for Annotation, Visualization and Integrated Discovery
DBH	Dopamine-β-hydroxylase
DBI	Database Interface
DBP	Diastolic Blood Pressure
DNA	Deoxyribonucleic Acid
EASE	Expression Analysis Systematic Explorer
eQTL	Expression Quantitative Trait Locus
ECM	Extracellular Matrix
EST	Expressed Sequence Tag
FCHL	Familial Combined Hyperlipidaemia
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GGM	Graphical Gaussian Model
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-wide Association Study
IBD	Identity by Descent
IQR	Interquartile Range
IRS	Insulin Resistance Syndrome
K-S	Kolmogorov-Smirnov (test)

LCMS	Likelihood-based Causal Model Selection
LOD	Logarithm of the Odds (score)
LRS	Likelihood Ratio Statistic
LVH	Left Ventricular Hypertrophy
LVM	Left Ventricular Mass
MAQ	Mapping and Assembly with Qualities
Mb	Megabase
MEMN	Macrophage-Enriched Metabolic Network
MM	Mismatch (probe)
mRNA	Messenger Ribonucleic Acid
NCEP:ATP III	The National Cholesterol Education Program's Adult Treatment Panel III Report
NEFA	Non-esterified Fatty Acids
NGS	Next Generation Sequencing
PASS	Power And Sample Size (software)
PCR	Polymerase Chain Reaction
pFDR	Positive False Discovery Rate
PM	Perfect Match (probe)
PNMT	Phenylethanolamine-N-methyltransferase
pQTL	Physiological Quantitative Trait Locus
qRT-PCR	Quantitative Reverse Transcriptase Polymerase Chain Reaction
QTL	Quantitative Trait Locus
QTT	Quantitative Trait Transcript
RI	Recombinant Inbred (strain)
RIX	Recombinant Inbred Intercross
RMA	Robust Multichip Average
RNA	Ribonucleic Acid
SCAMPA	Systematic Categorisation of Affymetrix Microarray Probeset Annotations
SBP	Systolic Blood Pressure
SDP	Strain Distribution Pattern
SgII	Secretogranin II
SHR	Spontaneously Hypertensive Rat (rat strain)
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
SVM	Support Vector Machine
WHO	World Health Organisation

Chapter 1: Introduction

1.1 Complex traits

Many of the commonest and costliest diseases, particularly in the Western world, are at least partially genetically determined, but cannot be explained by the inheritance of a single locus (i.e. they are not monogenic). Instead, the development and progression of such diseases as cancer, cardiovascular disease and diabetes are influenced by the interaction between genetic, developmental and environmental factors (Churchill et al., 2004). Therefore, these conditions are referred to as 'complex traits'.

The discovery of genes underlying complex traits is one of the biggest challenges facing researchers in the post-genomic era, and success in this area promises considerable improvements in medical knowledge and correspondingly in human health. However, the aetiologic complexity of these traits is such that progress in this area has, up until recently, been very slow (Yagil and Yagil, 2006).

1.1.1 The Metabolic Syndrome

A relationship between complex metabolic and physiological disorders including hypertension, dyslipidaemia, and insulin resistance was identified by Gerald Reaven in 1988, and termed Syndrome X (Reaven, 1988), or the 'Insulin Resistance Syndrome' (IRS). As described then, in its original form Syndrome X included (1) resistance to insulin-stimulated glucose uptake; (2) glucose intolerance; (3)

hyperinsulinemia; (4) increased very-low-density lipoprotein (VLDL) triglyceride; (5) decreased high-density lipoprotein (HDL) cholesterol; and (6) hypertension (McLaren, 1997). It was suggested that the presence of this syndrome in an individual is a significant risk factor for cardiovascular disease (CVD), and by the 1990s the name 'metabolic syndrome' was adopted to describe the condition (Ferrannini, 2007). Reaven did not include central obesity as part of his original definition of Syndrome X. Later, modified definitions such as the NCEP:ATP III and WHO definitions (Lakka et al., 2002), and the American Heart Association criteria (Grundy et al., 2004) provided new descriptions of the metabolic syndrome that included abdominal obesity in addition to increasing the specificity of definitions of components in Reaven's original criteria.

Although there is substantial evidence for links between the components of the metabolic syndrome and increased risk of CVD, there is some debate over the utility of the designation of the described combination of risk factors as a syndrome (Grundy, 2006; Kahn et al., 2005; Reaven, 2005). The facts that the described symptoms cluster together within individuals, and that they are associated with CVD, are however not in dispute – nor is the utility of rodent models which replicate features of the human metabolic syndrome, such as the Spontaneously Hypertensive Rat (SHR) (Aitman et al., 1997).

1.1.2 Hypertension

Hypertension can be defined operationally as an adult individual with a blood pressure of 140/90 mmHg or greater (Binder, 2007). As well as forming part of the criteria for diagnosis of the metabolic syndrome, hypertension is in and of itself a serious public health problem, with a prevalence of up to 40% in western populations (Zhu and Cooper, 2007). It arises as a consequence of a complex interplay between genetic and environmental factors. Estimates suggest that around 30% of the population variance can be accounted for by genetic variation (Zhu and Cooper, 2007). However, due to the numbers of different genes and pathways involved, the heterogeneity of the condition's aetiology, the high frequency of hypertension in the population, and the small effect size of causative genes, the identification of genetic variants involved in increasing risk of hypertension is not straightforward (Wellcome Trust Case Control Consortium, 2007)

In particular, because hypertension is so prevalent and the relative importance of genetic and environmental factors, such as diet and physical activity, is unclear and itself variable, the study of the genetics of hypertension in humans has proven refractory to even modern genetic strategies (Wellcome Trust Case Control Consortium, 2007). The use of animal models in laboratory conditions – i.e. a controlled environment – simplifies the investigation somewhat by minimising non-genetic variation. Rodent models such as the SHR, which is an inbred strain in which, as the name suggests, almost 100% of individuals spontaneously develop high blood pressure (Okamoto and Aoki, 1963), are a valuable resource for the study of the genetics of hypertension and related disorders.

1.1.3 Insulin Resistance

The term ‘insulin resistance’ refers to the defects in insulin action that are observed in the metabolic syndrome and other disorders, specifically defects in insulin action on carbohydrate and fat metabolism, such as glucose uptake and suppression of lipolysis (McLaren, 1997; Reaven, 1988). A common accompaniment of insulin resistance is hyperinsulinaemia, which has frequently been assumed to be a compensatory response, to maintain homeostasis of carbohydrate and fat metabolism. The aetiology and pathology of insulin resistance are still quite poorly understood, but the syndrome has been linked to numerous serious diseases including cardiovascular disease, type 2 diabetes and neurodegenerative disorders (Biddinger and Kahn, 2006).

The direct link between insulin resistance and type 2 diabetes, a disorder which is becoming increasingly prevalent across the world (Stumvoll et al., 2005), is of particular interest in the context of the metabolic syndrome since it suggests an inter-relationship between these conditions, obesity and cardiovascular disease. Links between insulin resistance syndromes and cardiovascular disease have been identified in 40-60% of coronary heart disease (CHD) patients (Aitman, 2001; Nathan et al., 1997). This means that genetic risk factors identified for insulin resistance syndromes may hold the key to improving our understanding of the causes not just of type 2 diabetes, but of cardiovascular diseases as well; in the context of the metabolic syndrome.

Type 2 diabetes, also referred to as non-insulin dependent diabetes, is diagnosed by the presence of high-levels of glucose in the bloodstream (hyperglycaemia) (Stumvoll et al., 2005). The occurrence of hyperglycaemia is a consequence of decreased insulin sensitivity as well as defects of β-cell function (Monnier et al., 2008).

1.1.4 Cardiac Hypertrophy

Hypertrophic enlargement of the ventricular wall through growth of individual myocytes is the heart's primary response to stressful physiological stimuli (Lorell and Carabello, 2000). Such enlargement, particularly of the left ventricle, occurs beneficially to health as a result of athletic exertion (Hill and Olson, 2008), but also pathologically as a response to increased blood pressure (Wilkins et al., 2004). It is a powerful predictor of cardiovascular disease, independent of other risk factors including the metabolic syndrome (Wagner et al., 2004). In particular, compensatory cardiac hypertrophy is part of a well-characterised path to heart failure (Liew and Dzau, 2004), as part of a process called 'left ventricular remodelling'.

It is hoped that gaining an improved understanding of the processes that occur during the early stages of heart failure will lead to progress in the development of new treatments. This incorporates the study of genetic variation that influences hemodynamic hypertrophy (e.g. (Cerutti et al., 2006)), and variation in left ventricular mass (LVM) that occurs independent of hypertension (Petretto et al., 2008)

1.2 Genetic Analysis of Complex Traits

The genes responsible for most of the more common monogenic diseases have been identified. However, because complex traits come about as a result of an interplay between multiple genetic and environmental factors, and each genetic locus involved explains only a small part of that interplay (Mott, 2006), studies aiming to identify contributory genes have often drawn a blank. The need for very large sample sizes and stringent case-control study design, and the difficulty of controlling for lifestyle factors, has made the discovery of complex trait genes in humans a formidable proposition (Glazier et al., 2002; Risch and Merikangas, 1996).

The approaches used in the effort to elucidate the genetic basis of complex traits have undergone many changes in the recent past and will continue to do so. The foundation upon which the varied methodologies are built is, however, consistent – the relating of region(s) of the genome with phenotypic variation. The two widely used methods of doing so are linkage analysis, used to generate quantitative trait loci (QTLs), and association analysis.

1.2.1 Genetic Markers

Genetic markers are regions of the genome that are variable, or polymorphic, between the populations under investigation. There are several types of genetic marker; the two that are most commonly used for linkage analysis are microsatellites (also known as simple sequence repeats (SSRs)) and single nucleotide polymorphisms (SNPs). A variation on SNP markers is the single functional polymorphism (SFP). Microsatellite markers are variable in length between strains or individuals, whereas SNPs and SFPs are variable in sequence. The genetic markers used in the study described in this thesis are microsatellite markers.

1.2.2 Quantitative Trait Loci (QTLs)

A quantitative trait is a phenotypic characteristic that varies in a quantifiable manner, such as blood pressure, waist circumference and left ventricular mass. The mapping of quantitative traits to the genome in the form of QTLs by linkage analysis has been performed in humans and in model organisms for many years. QTL mapping was carried out in the SHR and BN parental strains prior to the commencement of the work described in this thesis (Hubner et al., 2005); these QTLs are henceforth referred to as the physiological QTLs (pQTLs). As of 2005, almost two thousand QTLs had been mapped to the mouse genome (Flaherty et al., 2005), 708 QTLs had been mapped in the rat, and just over one thousand in humans (de la Cruz et al., 2005). QTL mapping has formed the first stage of the process of identification of numerous complex trait genetic loci (Cervino et al., 2005). Recombinant Inbred (RI) strains (described in detail in **1.4**), in particular, have enabled large numbers of QTLs to be localised (Flaherty et al., 2005). However, having identified a QTL interval, the next step - the isolation of the contributory gene has proven to be much more challenging. By 2005, only around 20 genes were identified from the thousands of rodent QTLs that had been mapped (Flint et al., 2005).

1.2.3 Genome-wide Association Studies (GWAS)

Genetic association analysis makes use of historical recombination to compare two groups (for instance, cases and controls) for the relative frequency of occurrence of a genetic trait such as a SNP. Typically, a *p*-value is found to assess the significance of the association of the phenotype with each SNP, although Bayesian methods that can quantify the probability of association are also suitable for use (Stephens and Balding, 2009).

The increased availability of sequencing has enabled the frequencies of genetic variants to be compared across the whole genome between large samples of cases and controls, an approach referred to as a Genome-wide Association Study (GWAS) (Kruglyak, 2008). The advent of GWAS has increased the rate of gene identification in higher mammalian complex traits dramatically. As of May 2008, almost 100 contributory loci for nearly 40 complex traits have been identified (Manolio et al., 2008). This has been made possible thanks to the availability of the International HapMap resource (Morton, 2008); large, well-characterised sets of clinical samples;

and the ability to genotype hundreds of thousands of markers, usually single nucleotide polymorphisms (SNPs) in samples classified according to a trait of interest (Wellcome Trust Case Control Consortium, 2007).

Although GWAS is an important innovation and has enabled several novel complex trait loci to be identified, the importance of the contribution to disease research made by rodent models has not diminished. GWAS have succeeded in identifying genomic regions – which can be large as a consequence of linkage disequilibrium (Lusis et al., 2008) – explaining only a small part of the observed phenotypic variation (Sarwar and Cook, 2009). It cannot detect rare variants and those that are too complex to associate with a SNP locus (Nathan and Orkin, 2009). Additionally, replication of GWAS findings in different populations has proven problematic in some cases (Frayling et al., 2007; Vogel et al., 2009). Animal models; particularly recombinant inbred strains and congenic and consomic strains are expected to continue to be valuable resources for geneticists (Aitman et al., 2008).

1.3 Genetic Analysis of Model Organisms

1.3.1 The Use of Model Organisms

The study of model organisms has facilitated the identification of complex trait susceptibility genes and pathways, by genetic linkage approaches and by positional cloning (Krinke, 2000). The use of inbred rodent models such as the spontaneously hypertensive rat (SHR) and recombinant inbred (RI) strains generated using it in genetic analysis eliminates the problem of genetic heterogeneity and allows almost complete control over environmental factors. It is possible to set up precise, controlled experimental crosses, making possible a wider scope of experimentation. It also enables larger sample sizes to be utilised. The most significant downside of using model organisms is that there are of course substantial developmental, physiological and genetic differences between rodents and humans.

The prioritisation of candidate genes in animal models, particularly rodents, is an approach that has been successfully applied to the study of a wide variety of complex traits (Glazier et al., 2002). Typically, the methods of analysis used in rodent models have entailed genome-wide linkage and gene expression analysis (Clee, 2005). Strategies integrating these two approaches were first used in the identification of a gene underlying a complex trait in (Aitman et al., 1999) and have been applied in rodent models to predict candidate loci for hypertension (Hubner et al., 2006; Soler et al., 2006), obesity (Mehrabian et al., 2005), and cardiac mass (Petretto et al., 2008) among others.

1.3.2 Rodents as Model Organisms for Complex Traits

The utility of rodent models for disease research has been established for almost a century, and has provided insights into the physiological and genetic basis of a wide range of disorders (Editorial, 2008). Of the two rodent species that are widely used as model organisms, the mouse (*Mus musculus*) became established as the geneticist's mammalian model of choice early on (Jacob, 1999), and was the first rodent genome to be sequenced (Waterston et al., 2002). As of 2005, almost three times as many QTLs had been identified in mouse as in rat (Flint et al., 2005). Comprehensive QTL databases exist for mice (Eppig et al., 2005) and rats (de la Cruz et al., 2005).

The rat (*Rattus norvegicus*) has been at the centre of disease research since the middle of the 19th century, and inbred strains have been in development since 1909 (Jacob, 1999). Although geneticists have historically favoured the mouse, the rat has always been preferred by physiologists because it is more amenable to laboratory experimentation. Over 500 inbred rat strains have been created in order to study common diseases (Lazar et al., 2005). As rat genetic and genomic resources continue to develop and become increasingly competitive with those available for mouse (Twigger et al., 2008), insights into rat physiology and biochemistry can be used to inform the findings of genetic analysis in the rat.

1.3.3 The Spontaneously Hypertensive Rat (SHR)

In terms of research into the metabolic syndrome and associated disorders, the most important of the inbred rat strains in terms of the volume of research carried out using it is the Spontaneously Hypertensive Rat (SHR).

The SHR inbred strain was created in 1962 from an outbred strain of *Rattus norvegicus*, the Wistar strain. The rats to be bred were selected for high blood pressure until occurrence of spontaneous hypertension approached 100% (Okamoto and Aoki, 1963).

It has since been established that, alongside displaying the consistently raised blood pressures described by Okamoto and Aoki, the SHR also models other aspects of the metabolic syndrome (Pravenec et al., 2004; Pravenec et al., 2002), including insulin resistance, dyslipidaemia and disturbances of lipid and carbohydrate metabolism. Several QTLs for metabolic syndrome-associated physiological traits have been mapped in the SHR (e.g. (Kren et al., 1997)), along with QTLs for other disorders including cardiac hypertrophy (Cerutti et al., 2006; Inomata et al., 2005; Pravenec et al., 1995) and neuro-behavioural traits such as attention deficit hyperactivity disorder (ADHD) (Vendruscolo et al., 2006).

Several other inbred lines modelling some aspects of the metabolic syndrome have also been developed. These include the Dahl salt-sensitive rat, the Lyon hypertensive rat and the Milan hypertensive rat, all of which are hypertensive (Channa et al., 2004; Pravenec et al., 2002).

1.4 Recombinant Inbred Strains

The development of rodent recombinant inbred (RI) strain panels was pioneered in the early 1970s by Donald Bailey, who by repeatedly interbreeding the same two inbred strains was able to produce a genetically distinct but fixed set of strains that are homozygous at all genetic loci (Bailey, 1971) The first ‘RI panel’, of 7 mouse RI strains – the first time this had been achieved in mammalian species – was used to study histocompatibility genes using linkage analysis.

The use of RI strains is also an established methodology in research in other species, such as the plant model organisms *Arabidopsis thaliana*, tomato and maize, the nematode *C. elegans*, and the insect model organism *Drosophila melanogaster* (Peirce et al., 2004). For obvious reasons, research into complex traits like the metabolic syndrome requires the use of a mammalian model.

RI strains provide a cumulative and permanent resource for model organism genetic research. They are of particular value, therefore, in the mapping of complex traits, because unlimited phenotypic measurements may be obtained for each line – further animals can be bred according to demand. Multiple genetically identical individuals from each line can be assayed, e.g. in gene expression studies, to reduce individual environmental and measurement variation, increasing expression trait heritability. They provide a greater resolution for genetic mapping than do intercrosses, as the breakpoints on RI chromosomes are denser (Broman, 2005). Investigation of statistical power to detect QTL has found that the number of RI strains required to achieve a given power to detect QTL is half that required in an F2 population (de Koning et al., 2005).

1.4.1 The BXH/HXB Recombinant Inbred Panel

The 29 strains making up the BXH/HXB recombinant inbred panel were created by systematic inbreeding for around 60 generations between members of the F2 generation resulting from an initial intercross of membes of the hypertensive SHR/Ola and normotensive BN.*lx*/Cub inbred strains (Pravenec et al., 1989). The breeding process used in the generation of this panel is illustrated in **Figure 1.1**.

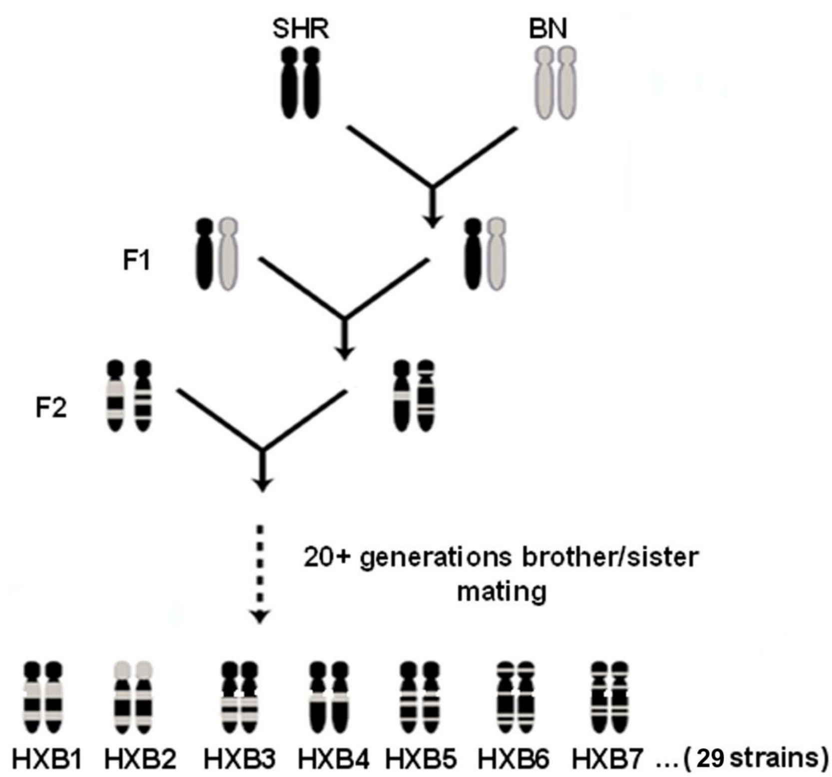


Figure 1.1: Generation of the BXH/HXB panel of recombinant inbred strains (Van Zant and de Haan, 2004)

The BXH/HXB panel was, at the time of its completion, the largest rat RI panel in the world (Pravenec et al., 1995). The genetic map derived from this panel of RI strains

(Jirout et al., 2003) has facilitated the mapping of physiological traits within the panel (Jaworski et al., 2002), and provides the basis for large-scale analysis of the genetic basis of the complex traits modelled by the SHR (Hubner et al., 2006).

1.5 Tissue Selection

The large-scale combined genetic and gene expression analysis described here was carried out on samples of four tissues that were chosen for use in the investigation because of their relevance to the complex traits manifested in the SHR and in order to build on previous findings.

As discussed previously in **1.3.3**, the SHR was long prior to the development of the BXH/HXB RI strains a widely used model of spontaneous hypertension, defective insulin action and other aspects of the human metabolic syndrome (**1.1.1**). Adipose tissue (fat) samples and kidneys were used for gene expression analysis in the initial analysis described in (Hubner et al., 2005). Adrenal glands and left ventricles were subsequently analysed in the same way; this is described in (Petretto et al., 2006a).

1.5.1 Fat

The role played by the adipose tissue in the development of the metabolic syndrome is discussed in (Eckel et al., 2005). Perhaps counterintuitively, adipocytes are now recognised as a significant endocrine organ (Biddinger and Kahn, 2006), secreting factors such as adiponectin which have been linked to metabolic syndrome and also heart disease (Eckel et al., 2005).

One gene that is highly expressed in adipose tissue and whose variation has been shown to play an important role in development of the insulin resistance syndrome in the SHR is *Cd36* (Aitman et al., 1999), which was found to underlie a QTL for hypertension in rodents.

1.5.2 Kidney

The kidney is key to the regulation of blood pressure, through the renin-angiotensin-aldosterone system (Binder, 2007; Delles et al., 2008), which regulates levels of salt reabsorption in the kidneys' tubules. Components of the system have previously been identified through genetic analysis and linked to essential hypertension (Binder, 2007). Genetic variation in the kidney has also been linked to the development of pathologies in other tissues including left ventricular hypertrophy (Ortlepp et al., 2001) and new-onset heart failure (Schocken et al., 2008).

The central role of the kidney in the development of hypertension has been demonstrated through rodent renal grafts, whereby transplantation of kidneys from genetically hypertensive donors has produced essential hypertension in previously normotensive individuals (Rettig and Grisk, 2005).

Cd36, which as previously discussed has been shown to play an important role in the development of insulin resistance, has also recently been implicated in renal regulation of blood pressure (Pravenec et al., 2008).

1.5.3 Adrenal Gland

The contribution of the adrenal gland to the development of hypertension and the metabolic syndrome is far from fully understood. It is known that rare tumours of the adrenal gland, such as pheochromocytoma, can directly produce pathologically high blood pressure (Young, 2007) by affecting levels of secreted factors such as catecholamines. At the same time, catecholamine deficiencies have previously been associated with dyslipidaemia and insulin resistance (Landsberg, 1999)

Aldosterone, a hormone produced by the adrenal cortex, forms part of the aforementioned renin-angiotensin-aldosterone system of blood pressure regulation. Levels of aldosterone are regulated primarily by angiotensin, but are also influenced by the sympathetic nervous system (SNS). Increased activity of the SNS has been associated with insulin resistance (Reynisdottir et al., 1994) as well as hypertension (Landsberg, 1999).

1.5.4 Left Ventricle

The SHR is of interest as a model of Left Ventricular Hypertrophy (LVH) as well as the metabolic syndrome. A QTL for blood pressure-independent variation in cardiac mass was identified in the BXH/HXB RI strains in 1995 (Pravenec et al., 1995), suggesting that analysis of the relationship between this phenotype and gene expression in the LV has the potential to be informative

1.6 Gene Expression Analysis

The quantity of the messenger RNA (mRNA) intermediate that is produced when a region of the genome is transcribed can be referred to as the 'transcript abundance'. The measurement of the transcript abundance and therefore of gene expression can be undertaken on a genome-wide basis through the application of microarray technology (Chee et al., 1996).

Since the advent of microarray technology, the study of differential gene expression levels as a consequence of genetic variation has made possible advances across the whole of the field of biology. Variation between individuals in gene regulation has been considered by some authors to be at least as significant as structure in terms of explaining phenotypic differences (Cotsapas et al., 2006; Whitehead and Crawford, 2006). The study of expression variation is therefore critical if we are to gain an understanding of the genetic basis of complex traits.

1.6.1 Measurement of Gene Expression

DNA microarrays can be used to measure simultaneously the expression levels of thousands of genes; this is referred to as gene expression profiling. Among other applications of this technology are comparative genomic hybridization (for assessing copy number variation) and chromatin immunoprecipitation (for investigating interactions between DNA and proteins). Microarrays can be used to measure levels of DNA as well as mRNA.

Expression profiling can be carried out using several different types of expression array which, although they vary in their design specifications, all perform essentially the same function. The two main types of DNA microarray are oligonucleotide arrays and spotted arrays (Efron et al., 2001). Both consist of 'probes' – short sequence fragments (usually either oligonucleotides or cDNAs) that complement the targets (e.g. mRNAs) whose abundance is to be measured.

The specific hybridization of complementary mRNAs to the probes on the array produces a signal whose intensity is measured to provide a quantitative assessment of the abundance of the transcript (Whitehead and Crawford, 2006). Due to the need to quantify the intensity and because of the large numbers of simultaneous tests performed, the downstream analysis and the interpretation of the data so obtained are a considerable computational and statistical challenge.

1.6.2 The Affymetrix GeneChip

The microarray system that was used to generate the rodent gene expression profile data described in this thesis is the Affymetrix GeneChip®. The GeneChip is an oligonucleotide microarray, whose design incorporates pairs of 'perfect match' (PM) and 'mismatch' (MM) probes. Probes consist of complementary oligonucleotides, 25 bases in length – the MM probes contain one mismatch (non-complementary base) at the 13th position. Each probeset normally consists of 11 probe pairs. It is intended that all probes within a probeset should map to the same gene, but sometimes this is not the case (Harbig et al., 2005).

1.6.3 Validation

The microarray is an extremely efficient, high-throughput method of measuring gene expression that allows analysis to be performed simultaneously on a genome-wide scale – something that would be utterly impractical were only single-gene techniques available for use.

That said, laboratory techniques such as the Quantitative Real Time Polymerase Chain Reaction (qRT-PCR) and Northern blotting still play an important role in gene expression studies, as a means of independently validating the data obtained through the microarray analysis. This is an integral part of the quality control process for microarray experiments (Chuaqui et al., 2002).

1.6.4 Sequence Analysis

The annotation of microarray probesets to genes, and downstream functional annotation (e.g. Gene Ontology annotation) depends on the initial mapping of a probeset to a specified transcript. This is contingent on the probeset hybridizing with the mRNA with which it is expected to hybridize.

> *"Part of the problem is that the probes on an array are identified based on what [Affymetrix] was hoping to detect, not based on what they actually detect."* (Harbig et al., 2005)

Because of this, it was considered necessary to cross-reference the probeset annotations provided by Affymetrix with those in the EnsEMBL annotation database using SCAMPA (Dickens, 2006)(**2.2.2.1)**. This has made it possible for us to use in our analysis only probesets that hybridize with the gene indicated in the probeset annotation provided by NetAffx.

Another consideration related to the probeset sequence pertains to the sequence of the region of DNA to which the probeset is intended to bind. Polymorphisms within the mRNA region that is targeted by a probe can result in hybridization differences (between two sample populations) that are not caused by differential gene expression (Alberts et al., 2005; Alberts et al., 2007b). This can lead to misinterpretations of the data later on in the analysis.

1.7 Expression Quantitative Trait Loci (eQTLs)

The eQTL study design utilises expression traits measured by DNA microarray whose genetic control points are then mapped to the genome through linkage analysis, enabling the basis of quantitative variation in gene expression to be studied. The technique, applied on a genome-wide basis, is also known as 'genetical genomics' and was first proposed by Ritsert Jansen and Jan-Peter Nap in 2001 (Jansen and Nap, 2001). Since then, the genetical genomics methodology has been applied in budding yeast (Brem et al., 2002), nematodes (Li et al., 2006), plant models (Hansen et al., 2008; West et al., 2007), rodent models (Hubner et al., 2005; Lum et al., 2006), and in humans (Goring et al., 2007; Hansen et al., 2008).

eQTL analysis can be carried out in the segregating progeny of a backcross or intercross, in recombinant inbred lines, or (typically in humans) in pairs of siblings (e.g. (Goring et al., 2007)). The methodology is similar to that used to map physiological quantitative trait loci (pQTLs), except the quantitative traits that are to be mapped are expression levels of thousands of genes. As with the mapping of pQTLs, eQTL analysis requires a genetic map, and the genotyping of genetic markers across the genome (Kendziorski and Wang, 2006).

1.7.1 The Strain Distribution Pattern (SDP)

The set of genotypes of a given genetic marker across a panel of recombinant inbred strains is referred to as that marker's Strain Distribution Pattern (SDP). Ideally, the SDP consists of a complete set of known genotypes of a vector corresponding to the number of strains in the panel. However, in practice, genotype data is often missing from the SDP. It was necessary during the generation of our eQTL dataset to take into account the statistical and computational implications of the presence in the marker data of missing genotypes (**3.1**).

1.7.2 Mapping eQTLs

There are several factors that must be taken into account when mapping eQTLs to the genome in order for the data to be meaningful and suitable for downstream analysis (Kendziorski and Wang, 2006). Summarisation, using a pre-processing algorithm such as Robust Multichip Average (RMA), and the genomic locations of all of the markers used in mapping must be verified as far as possible (e.g. (Hubner et al., 2005)). In the process of mapping itself, the number of transcripts and therefore of possible linkages means that there is a substantial problem of multiple testing. The computational methods used to generate eQTL datasets have to take these into account.

The eQTL dataset described here, in the BXH/HXB RI strain panel, was generated using a map consisting of 1,011 microsatellite markers (Hubner et al., 2005). These were genotyped in each of the 29 strains using the methods described in (Jirout et al., 2003), contiguously covering almost all of the genome and providing a clear indication of the segregation of the parental genotypes (SHR and BN) across the strains following genotyping.

Mapping of the eQTLs to the genome was carried out using QTL Reaper (Manly, 2006). Markers for which 10 or more strains lacked genotype data were excluded from the analysis. Interpolation of missing genotype data was not carried out for reasons of time.

QTL Reaper carries out genome-wide linkage analysis for each expression trait, and generates a likelihood ratio statistic (LRS) as a measure of the significance of each possible eQTL. The statistical significance of the LRS is assessed empirically by permutation, and the false discovery rate (FDR) evaluated in order to control for multiple testing (**2.5.2**), to provide a genome-wide corrected *p*-value for each probeset-marker combination. Typically, an eQTL would be used in downstream analysis if it was found to have genome-wide $p<0.05$ following correction for multiple testing (Hubner et al., 2005).

The LRS technique used by QTL Reaper, and its predecessor Map Manager QTX (Manly et al., 2001), is a weighted regression-based method of assessing linkage based on methods first described by (Haley and Knott, 1992).

SNPs, which can provide a denser marker map (STAR Consortium, 2008), became available later in the course of the project, and were available for use in fine-mapping of specific quantitative traits (Petretto et al., 2008). However, the eQTLs were not re-created using the new markers for use in this project due to a combination of statistical considerations and time constraints.

1.7.3 Visualising eQTLs

eQTL Explorer (Mueller et al., 2006) is a database-driven software application that was developed in order to facilitate the handling and interpretation, and in particular the use by non-bioinformaticians, of the eQTL dataset. eQTL Explorer enables *cis*- and *trans*-eQTLs to be visualised by the map location of the peak of linkage, and pQTLs by the locations of the flanking markers. Examples of eQTL Explorer's display of the BXH/HXB database are provided in **Figures 1.3** and **1.4**.

1.8 *Cis*- and *Trans*- Expression Quantitative Trait Loci (eQTLs)

One of the strongest aspects of the eQTL study design is its ability to delineate eQTLs into two classes, *cis*- and *trans*-, depending on the location of the expression trait compared to that of the genotyped marker.

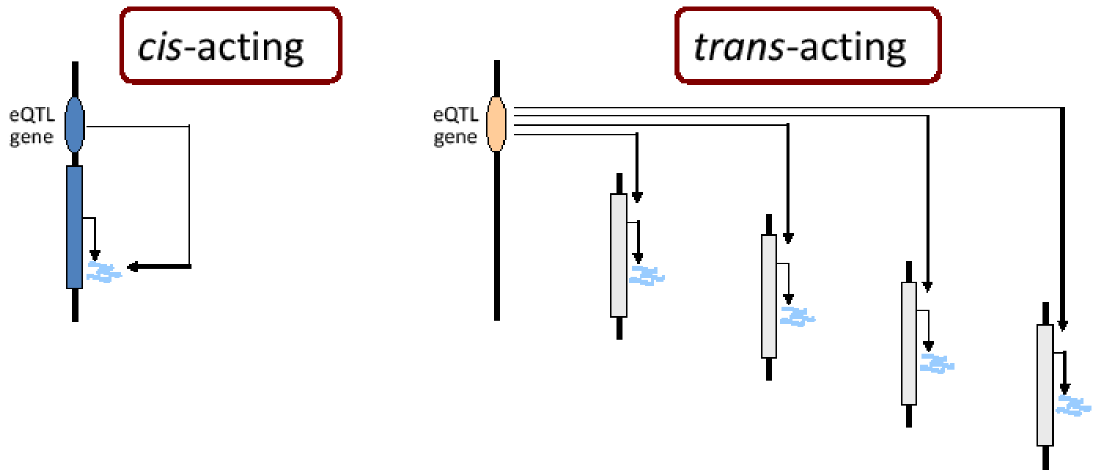


Figure 1.2: Illustration of the genetic basis of *cis*- and *trans*-acting expression Quantitative Trait Loci (eQTLs) (Petretto, E., personal communication).

As illustrated by **Figure 1.2**, *cis*-eQTLs are defined, as detailed in **1.8.1** and in (Hubner et al., 2005), so as to encompass those genes in which variation in transcript abundance is caused by genetic variation in the same gene (Doss et al., 2005). *Trans*-eQTLs (**1.8.2**) are of interest as potential remote regulators of one or more downstream genes (Petretto et al., 2006b), which may or may not be on the same chromosome as the gene in which the genetic variation giving rise to the trans-eQTL occurs.

This delineation is therefore biologically significant. The identification of an eQTL implies the presence of a genetic sequence variation that influences the levels of the transcript concerned (Drake et al., 2006). Whether that influence is local or remote has important implications for the likely biological basis of the transcript variation.

1.8.1 *Cis*-eQTLs

Cis-eQTLs are typically defined in such a way that the transcript whose abundance is being measured is, as often as possible, the result of genetic variation in the same gene, perhaps in the promoter or regulatory regions (Yvert et al., 2003). The most suitable definition of the 'window' region of the *cis*-eQTL depends on the inherent margin of error in QTL mapping and the recombination rate across the genome (Hubner et al., 2005). A *cis*-eQTL is defined here as one in which the marker and the transcript are no more than 10Mbp apart (20Mbp total window size), as selected to capture as many positional candidate genes for physiological QTLs as possible (Figure S3, (Hubner et al., 2005)). Where a substantially greater mapping precision can be achieved, such as in eQTL studies using yeast (e.g. (Yvert et al., 2003)), 'window' regions as small as 20kb have been used in the definition of a *cis*-eQTL. The guiding principle underlying the *cis-trans* delineation remains as described.

Because *cis*-eQTLs are indicative of genetic variation that has a statistically significant influence on gene expression levels, it follows that the same genetic variation may also influence physiological phenotypes. For this reason, the identification of *cis*-eQTLs that colocalise with pQTLs for a trait of interest is considered to be of particular value to geneticists (Doss et al., 2005; Flaherty et al., 2005; Wang et al., 2007). This principle underlies the analysis described in chapter **5**.

1.8.2 *Trans*-eQTLs

Trans-eQTLs are those in which variation in expression levels is controlled remotely. The eQTL (sequence variant) may be on the same chromosome as the transcript but outside the window region used to define *cis*-eQTLs, or it may be on a different chromosome.

Because the presence of *trans*-eQTLs implies that genetic variation in one gene significantly affects the expression of another, it has been hypothesised (Yvert et al., 2003) that *trans*-eQTLs are likely to represent polymorphisms in transcription factors – DNA-binding proteins that regulate gene expression (Pabo and Sauer, 1992). However, it has been observed that *trans*-eQTLs often do not map to regions of the genome containing any transcription factors (Tu et al., 2006), suggesting the involvement of *trans*-eQTL genes in more complex regulatory mechanisms. One reason for this, suggested by (Kliebenstein, 2009), may be that the pleiotropic effects of transcription factor mutations are so large as to impact the viability of organisms in which they occur.

Trans-eQTLs typically show smaller effects than *cis*-eQTLs, suggesting that they represent a lesser degree of biological variability than do *cis*-eQTLs (Petretto et al., 2006a). This finding is consistent with the aforementioned explanation of the biological basis of *trans*-eQTLs, and has been observed in other studies also (e.g. (Sladek and Hudson, 2006)).

1.8.2.1 *Trans*-eQTL Clusters

One feature of *trans*-eQTL datasets that has been the subject of much study and no little controversy is the tendency, first noted in yeast by (Yvert et al., 2003), for tens or even hundreds of expression traits across multiple chromosomes to be linked to the same genetic marker. Such groups of *trans*-eQTLs are variously referred to as *trans*-eQTL clusters, *trans*-eQTL 'hotspots', or *trans*-bands (the term '*trans*-eQTL cluster' is used to refer to them in this thesis).

Trans-eQTL clusters have been observed in eQTL studies in several species (Kirst et al., 2005; Morley et al., 2004; Petretto et al., 2006b). A *trans*-eQTL cluster identified on chromosome 4 in the BXH/HXB RI panel (Hubner et al., 2005), as visualised by eQTL Explorer (**1.7.3**), is illustrated in **Figure 1.3**, overleaf.

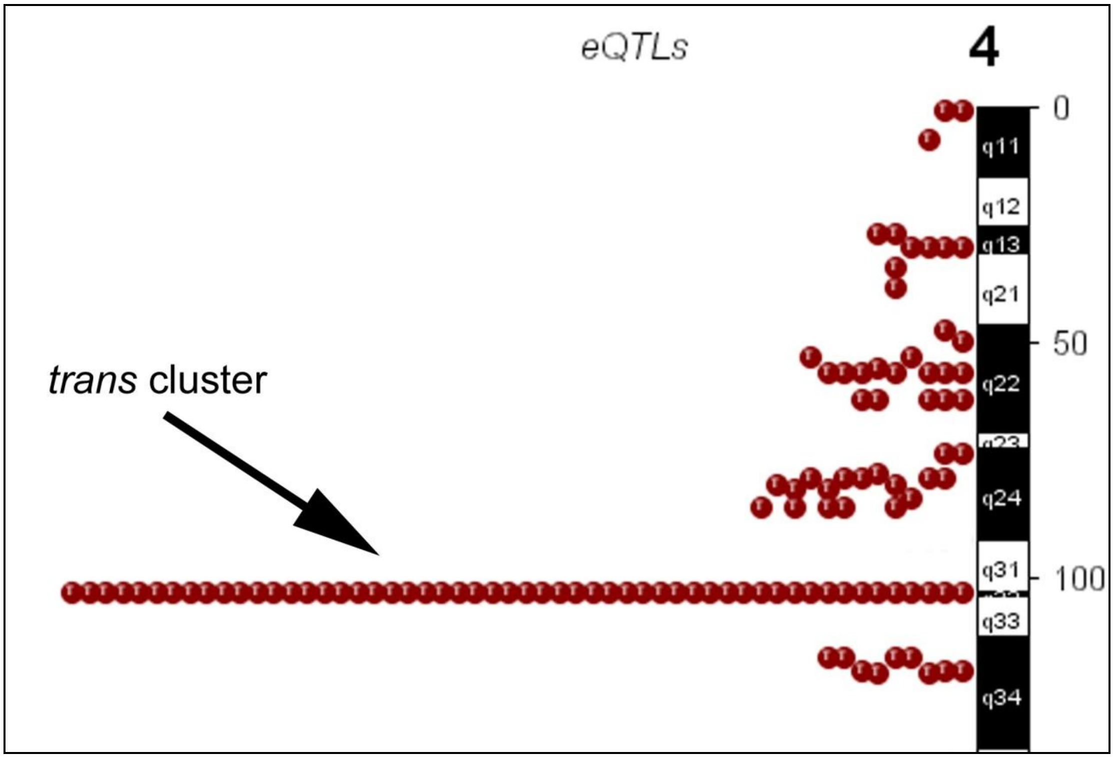


Figure 1.3: A *trans*-eQTL cluster mapped to chromosome 4 in the RI panel in kidney, as visualized by eQTL Explorer (Mueller et al., 2006)

Because *trans*-eQTL clusters consist of multiple eQTLs that all map to the same location but affect transcript levels across the genome, it has been suggested that all of these transcripts may be under the control of a single regulatory gene (Yvert et al., 2003), often referred to as a hypothetical 'master regulator' (Pastinen et al., 2006). This hypothesis is however disputed; it has been suggested that the observations of *trans*-eQTL clusters and corresponding hypothesized master regulators may be a consequence of systematic artefacts in microarray data (Kang et al., 2008), particularly since when real expression data has been combined with simulated SNP marker genotypes, 'eQTL hotspots' have arisen in the resulting dataset (Pérez-Enciso et al., 2003) suggesting that clusters observed in eQTL experiments may arise from chance linkage of a group of correlated transcripts with a region of the genome. As discussed in **1.8.3**, *trans*-eQTL clusters consist of expression traits with low effect sizes, a high proportion of which are mapped to the genome with a high probability of them being false positives, as evaluated by the expected FDR (**2.5.2**) (Petretto et al., 2006a). The statistical significance of the eQTLs forming *trans*-eQTL clusters was investigated as part of this study and found, in general, to be low (**4.2.2**).

1.8.3 Statistical power to detect *cis*- and *trans*-eQTLs

Statistical power is the power of a statistical test to reject the null hypothesis given that it is false. The statistical power to detect QTL is dependent on the sample size of the experimental population used for mapping. However, as previously mentioned (**1.4**), because RI strains are homozygous across the genome they offer a statistical power twice that achieved in F2 crosses for an equivalent number of strains in the RI panel to that of individuals in the F2 cross.

Cis-eQTLs are, on average, easier to detect because they explain a larger proportion of the variation in gene expression (their effect size is larger) than do *trans*-eQTLs (Hubner et al., 2005). Calculations of the power to detect *cis*- and *trans*-eQTLs in four tissues from the BXH/HXB panel of 29 RI strains were carried out by colleagues and the outcome of this analysis is described in (Petretto et al., 2006a). The statistical power and median expected FDR for *cis*- and *trans*-eQTLs in each of the four tissues were calculated. For *cis*-eQTLs with high ($h^2 > 0.25$) trait heritability, the statistical power was found to be ~100% in all tissues. For similarly heritable *trans*-eQTLs, the power ranges from ~10% (in adrenal) to ~60% (for the most heritable expression traits in fat). The median expected FDR for *cis*-eQTLs ranges from 4% to 8% across the four tissues, whereas for *trans*-eQTLs it ranges from ~26% (in fat and left ventricle) to ~35% (in kidney and adrenal) (Figure 4, (Petretto et al., 2006a)).

1.9 Investigating Relationships between Expression Traits

In the study of complex traits, it is important to realise that the understanding that can be obtained by studying genes in isolation is limited. The analysis of gene interactions and functional modules can provide a much more complete picture of the factors underlying phenotypic variation.

Several studies have ‘integrated’ gene expression measurements with associated data (e.g. heritability), genotype information and phenotypic measurements to predict gene interactions (Brem et al., 2005) and modules (Li et al., 2005), and build gene networks (Sieberts and Schadt, 2007). Such analyses typically make use of mathematical models; some relatively simple, others highly complex; *in silico* to generate hypotheses regarding genetic interactions. The foundations that these analysis are built on are the relationships between gene expression traits – commonly considered in terms of correlation of expression profiles across a population – and between these traits and the genetic loci that influence their genetic variation.

1.9.1 Co-expression Analysis

Correlation in gene expression is commonly used to identify pairs or groups of genes that may share functional and/or regulatory relationships. Typically, pairwise correlation is carried out by finding the Pearson (or alternatively Spearman's rank) correlation coefficient (**2.4.1**). The evaluation of the Pearson correlation coefficient across a set of genes is often the first stage in the construction of a gene network.

1.9.2 Partial Correlation Analysis

Straightforward methods of assessing the correlation of two variables (e.g. expression profiles), such as Pearson correlation, do not distinguish between direct associations between the two variables and those that arise indirectly through the effects of intermediate traits (Rockman, 2008), or between causal and responsive factors (Opgen-Rhein and Strimmer, 2007).

One method of testing whether or not a relationship between two variables of interest is direct is by carrying out partial correlation analysis, conditional on a potential confounding variable, such as the genotype (SDP) of the marker at the shared peak of linkage of two eQTLs (Mozhui et al., 2008). The partial correlation coefficient (**2.4.5**), as applied to such a pair of expression traits conditional on genotype, quantifies the correlation between between the parts of the expression profiles that are uncorrelated with the genotype.

Partial correlation analysis can thus be used to distinguish between pairwise correlations that are due to direct causal relationships, and those that arise due in whole or in part to the potential confounding variable(s) on which the correlation is conditioned (De La Fuente et al., 2004). However, it does not provide unequivocal evidence that there is a direct relationship, as it is possible that a variable not taken into account in the partial correlation model is influential (De La Fuente et al., 2004).

1.9.3 Biological Relationships Between Genes

Two main types of relationship are studied using the methods described here: Functional relationships and regulatory relationships.

Examples of functional relationships are studies in which functionally related genes are predicted based on the observation of co-ordinated expression followed by relation to functional annotation (e.g. (Fuller et al., 2007)) and analyses predicated on prior biological knowledge, such as gene set enrichment analysis (GSEA) (e.g. (Subramanian et al., 2005)). The use of such methods to elucidate pathways and networks has the potential to provide a level of insight into biological processes that is not visible at the level of individual genes (Wentzell et al., 2007). eQTL studies seeking to elucidate functional relationships typically focus on *cis*-eQTLs, as these represent polymorphisms which affect local gene expression and are therefore less likely to represent regulatory effects than are *trans*-eQTLs (Petretto et al., 2006a), in which the genetic variation underlying the eQTL is in a different gene to that whose expression has been observed to vary.

Investigations into regulatory relationships have been carried out using *cis*- and *trans*-eQTLs. In some studies, the distinction is not made prior to the analysis of expression data, while in others *cis*- and *trans*-eQTLs are considered separately. In the latter (e.g. (Ghazalpour et al., 2006; Sun et al., 2007)), the observation of *trans*-eQTL clusters (hot-spots) perhaps under the control of a 'master regulator' (Yvert et al., 2003) can be incorporated into large-scale network analysis in aid of the prediction of regulatory relationships.

Some studies combine elements of functional and regulatory investigations in pursuit of the elucidation of biological systems (e.g. (Lan et al., 2006)). The use of these methods to predict networks of genes acting on a complex trait such as obesity in rodents (Fuller et al., 2007; Lum et al., 2006) or flowering time in *Arabidopsis* (Keurentjes et al., 2007) has led to the testing of regulatory hypotheses and the characterisation of new pathways. It is anticipated that quantitative 'systems' (Hall, 2004) approaches to the understanding of complex traits like obesity and the metabolic syndrome will grow in importance as the sophistication of the analyses that are possible – both in terms of biology and in computational power – increases (Wu et al., 2009).

1.9.4 Gene Networks

The construction of gene networks is often an intermediate stage in the elucidation of functional and/or regulatory relationships between genes. Networks consist of nodes, usually representing genes or proteins, linked by edges, indicating associations between the nodes (Sieberts and Schadt, 2007).

Methods of network reconstruction used in the analysis of gene expression data range from relatively straightforward co-expression networks built on correlations (defined mathematically in a variety of ways) between expression profiles (e.g. (Schadt et al., 2005; Zhang and Horvath, 2005)). In some cases, these analyses take into account the effects of linkage disequlibrium (Doss et al., 2005) and/or long-range allelic association (non-independence of genetic markers) (Cervino et al., 2005). More sophisticated methods of co-expression network modelling, such as the construction of Graphical Gaussian Models (GGM), use partial correlations to produce models that include only direct linear associations between nodes (Opgen-Rhein and Strimmer, 2007).

Networks that use gene expression data can be roughly divided into two groups: ‘a priori’ network analysis and ‘a posteriori’ (Hansen et al., 2008). The ‘a priori’ approach, exemplified by (Wentzell et al., 2007)’s study of glucosinolate biosynthetic pathways in *Arabidopsis*, makes use of prior knowledge of biological relationships. The ‘a posteriori’ approach, of which the aforementioned GGM network inference methodology is an example, does not require prior biological information but begins with uninformed correlations. As such, ‘a posteriori’ methods have the potential to identify truly novel networks and processes – but are somewhat more sensitive to the assumptions made in their construction (Drake et al., 2006) and the interpretation of the findings – for instance the implications for the network of *trans*-eQTL clusters (‘hot-spots’) (Tu et al., 2006).

Some studies have made use of still more complex mathematical models, such as Bayesian network reconstruction, a probabilistic method which provides the option of weighting of the expression profiles (Li et al., 2005; Zhu et al., 2004). This and other techniques enable factors related to the original linkage analysis (e.g. LOD score) and/or functional information to be made use of in the construction of the network. This increases the flexibility of the technique, but can also make the interpretation of the output more challenging.

1.9.5 Modules

Many integrated studies make use of the concept of ‘modules’ – groups of highly correlated genes identified following co-expression analysis (Ghazalpour et al., 2006). The definition of a module used varies in its specifics from study to study, but the aim is the same: to find groups of genes which relate to a common biological function and may combine to form pathways (Sieberts and Schadt, 2007).

The study of the patterns of co-expression has subsequently been applied to the modules identified – often following network construction - to predict structures within these modules (Lum et al., 2006), allowing relationships between component genes to be inferred. A separate study (Sun et al., 2007) investigated expression profiles of *cis*- and *trans*-eQTLs forming modules to analyse the transcriptional regulation of the networks identified.

These modules can related to physiology through the use of functional annotation methods such as Gene Ontology (GO) (**2.6.1**), e.g. (Lum et al., 2006). Although ontological analysis is not without its limitations (Khatri and Draghici, 2005), it remains a well-established and useful method of identification of functionally related genes in a large set, such as biological pathways from modules identified through 'a posteriori' network analysis (Ghazalpour et al., 2005).

1.10 From eQTLs to Genes

eQTL datasets have been generated in a wide range of different species, making use of various experimental designs (de Koning and Haley, 2005). One of the commonest motivations for generating eQTLs is to serve as a potential resource for gene discovery and/or the investigation of interactions between genes, particularly in the study of complex traits (Drake et al., 2006). The aforementioned delineation of eQTLs into *cis*- and *trans*-regulated categories is a precursor to more function-focused downstream analysis.

Cis-eQTLs are assumed to arise as a consequence of genetic variation within a small region, specifically the genomic region containing the gene giving rise to the eQTL. Hence, somewhat more success has been achieved in the identification of genes underlying *cis*-eQTLs than in the study of *trans*-regulation which typically, necessarily focuses more on identifying pathways and interactions between genes (Tu et al., 2006). Candidate genes for disease traits have been prioritised and in some cases validated following the mapping of *cis*-eQTLs in many model organisms (e.g. (Brem et al., 2005; Meng et al., 2007; Petretto et al., 2008)) and in humans (Goring et al., 2007).

1.10.1 Prioritisation of Candidate Genes

The mapping of eQTLs in related populations (such as an intercross of inbred strains) that differ with regard to a complex trait provides an opportunity to scour the entire genome for expression variants that may relate to that trait.

One way of doing so that we have made use of in our analysis is to make use of previously mapped (Hubner et al., 2005) and new (Petretto et al., 2008) physiological

QTLs (pQTLs) . *Cis*-eQTLs that coincide with pQTLs, as illustrated by the eQTL Explorer visualisation shown in **Figure 1.4**, are considered to be particularly strong candidates for the gene that underlies the pQTL and therefore directly influences the phenotypic component concerned (Doss et al., 2005; Flaherty et al., 2005).

As can be observed in the illustrative figure (Figure 1.4), the pQTL intervals are often wide, encompassing many eQTLs and even more genes. The prioritisation of potential candidate genes can begin *in silico*, including through the use of functional analyses and comparative genomics (Wang et al., 2005), and by making use of higher-density SNP data to map QTLs with a substantially higher degree of precision (Cervino et al., 2005). The development of these techniques and associated network resources, such as EnsEMBL (Hubbard et al., 2002), has the net effect of reducing the amount of laboratory work entailed in gene discovery and decreasing the total time it takes and money it costs to do so.

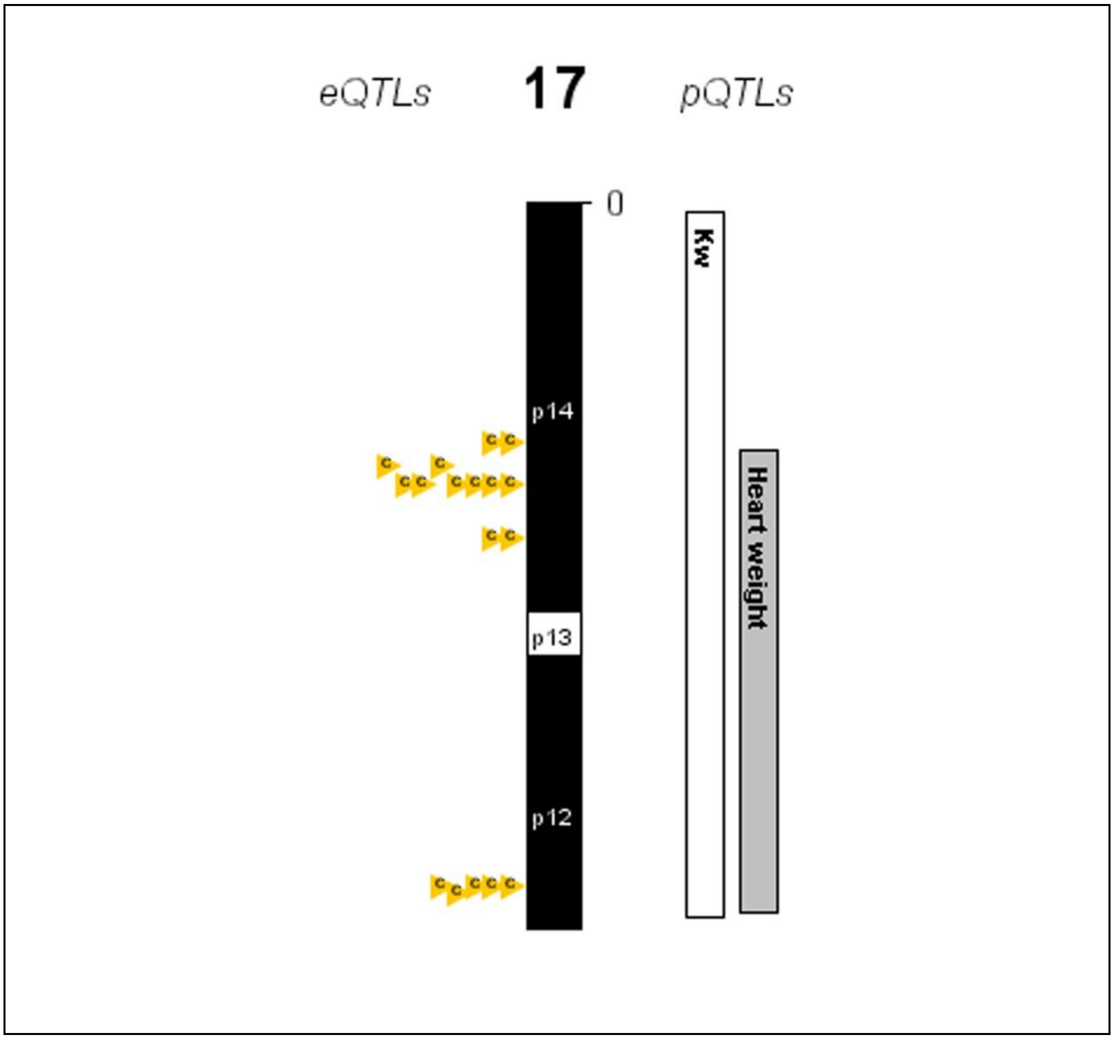


Figure 1.4: *Cis*-eQTLs, mapped to chromosome 17 in the RI panel in left ventricle, that are coincident with pQTLs for heart weight and kidney weight (Kw), as visualized by eQTL Explorer (Mueller et al., 2006)

1.10.1.1 Quantitative Trait Transcripts

Quantitative Trait Transcript (QTT) analysis, first described in (Passador-Gurgel et al., 2007), is a means of directly relating quantitative phenotypic variation with variation in gene expression across a population such as an RI panel using correlation analysis. Where the expression profiles of *cis*-eQTL transcripts are used in QTT analysis, the method effectively combines conventional QTL mapping (as by definition their expression profiles correlate with the genotype of a local marker) with correlation of the expression and physiological traits. The QTT method is however not limited to *cis*-eQTLs and can be applied to any transcripts for which expression profiles have been generated in the same population as the quantitative physiological trait has been measured, as was performed in humans and described in (Emilsson et al., 2008).

1.10.2 Validation of Candidate Genes

Although computational genomic methods have revolutionised the prioritisation of candidate genes for complex traits, there is a limit to the level of evidence of gene function that can be gathered *in silico*. The gold standard for validation of genes hypothesized to underly QTLs remains to take place in the laboratory, using strategies such as transgenic expression of the candidate gene (e.g. (Meng et al., 2007; Schadt et al., 2005)) and/or gene knockouts (e.g. (Mehrabian et al., 2005; Petretto et al., 2008)) in combination with physiological and biochemical assessments.

1.10.3 Moving from Model Organisms to Humans

Applying genetical genomics in model organisms, particularly rodents, has enabled discoveries to be made that have directly contributed to our understanding of the biology of complex traits in humans, as well as that of the model organisms themselves. Comparative genomics and studies of homology, aided by resources such as EnsEMBL's Compara species comparison database (Flicek et al., 2008), facilitate the translation of rodent study outcomes to humans. The investigation of species concordance of QTLs (Wang et al., 2005) and identification of orthologous human genes to those whose functions have been investigated in rodents (Petretto et al., 2008) have enabled genes underlying atherosclerosis and cardiac mass to be characterised.

One of the most significant rationales for the continued use of animal models in biomedical research is, as discussed in **1.3**, the relative ease (compared to working with human subjects) with which a wide variety of physiological traits can be characterised. Before physiological QTLs can be mapped, it is necessary to identify and quantify the phenotype concerned. Inbred and recombinant inbred strains facilitate this by providing a genetically homogeneous sample population, minimising individual variation and increasing statistical power to detect QTLs. They can also be kept in a much more controlled environment than can human subjects (although this is less of a problem with cell-lines, there are only some physiological traits that can be studied using cell-lines). This increases the reliability of the QTLs that are then generated.

1.11 Aims of This Work

The eQTL dataset that has been used for this computational genetics study was generated in multiple tissues from the BXH/HXB panel of RI strains, the generation of which from SHR and BN is described in (Hubner et al., 2005; Petretto et al., 2006b) and which provides a platform for the study of the genetic basis of the metabolic syndrome in the rat model organism.

The availability for analysis of this dataset, along with extensive prior physiological characterisation of the RI panel, provides an opportunity to analyse in detail the relationships between expression traits, and between gene expression and physiological phenotypes. The 'genetical genomics' study design, integrating genetics and gene expression enables *cis*- and *trans*-eQTLs to be differentiated, and informed hypotheses relating to the processes underlying the observed expression variation to be generated.

The prior mapping of pQTLs for a variety of phenotypic traits in the SHR and BN strain combination enabled functional analysis of the dataset to be performed. Further physiological measurements were obtained in the parental strains and in the RI strains during the course of this study, and these were used in combination with the eQTL data to generate more specific functional hypotheses.

The specific aims presented in this thesis are to:

1. Verify the quality of the eQTL data in the RI strains that is the main material for the study described in the remainder of this thesis, including validation of the method used by the linkage software that mapped the eQTLs.
2. Assess a statistical model that aims to predict potential ‘false *cis*-eQTLs’.
3. Investigate the correlation structure of the eQTL dataset, and the relationship between pairwise co-expression of eQTLs and the genotypes at the eQTLs’ peaks of linkage in *cis*- and *trans*-eQTLs.
4. Assess the consistency or otherwise of the relationships observed in Specific Aim 3 across the four tissues of study.
5. Analyse the patterns of co-expression observed in *trans*-eQTL clusters (hot-spots), thereby testing the hypothesis that clusters represent co-ordinately regulated groups of genes.
6. Use computational methods to prioritise candidate ‘master regulators’ of *trans*-eQTL clusters, i.e. the *cis*-eQTLs in the vicinity of the cluster’s peak of linkage.
7. Perform quantitative trait transcript (QTT) analysis using physiological traits that have been measured in the RI strains, in order to generate hypotheses regarding the genetic basis of these traits.
8. For one candidate gene in particular that was identified in part through QTT analysis, *Ogn*, perform further statistical analysis in the RI strains and on data from humans.

Chapter 2 describes the materials and methods that are used in the results chapters of the thesis (chapters 3-5), including the statistical approaches used.

Chapter 3 describes quality control and statistical validation of the methods of linkage analysis used to generate the eQTL dataset in four tissues from the BXH/HXB RI panel. The investigations that resulted in the discovery and subsequent correction of a flaw in the software that was used for linkage analysis are outlined, as well as the implications for analysis and interpretation of eQTLs that were detected using this software. The assessment of a methodology for the detection of potentially spurious *cis*-eQTLs in the eQTL dataset using newly available SHR genome sequence data is described.

Chapter 4 contains the details of a co-expression analysis that was performed in *cis*- and *trans*-eQTLs following quality control and filtering. Statistically significantly correlated pairs of eQTLs were studied in all four tissues. The genetic basis of co-expression in *trans*-eQTL clusters was investigated, including testing for functional enrichment and reassessment of co-expression after taking into account the effects of genotype.

Chapter 5 describes the QTT analysis that was performed using physiological traits for which measurements were obtained in each of the RI strains. This entails the correlation of expression traits with physiological traits. The process and outcome of the QTT analysis of one trait, augmented by the incorporation of related analyses, that contributed to the results of a major publication are described. The outcomes of the application of the described methods to other traits are summarised. A QTT analysis performed in a human dataset is described.

Chapter 2: Materials and Methods

2.1 The eQTL Dataset

The eQTL dataset that formed the basis for this computational and statistical study was generated prior to the commencement of this thesis by Tim Aitman (MRC Clinical Sciences Centre, London; fat tissue) Norbert Hubner (Max-Delbruck-Center, Berlin; kidney and adrenal tissues) and Stuart Cook (MRC Clinical Sciences Centre, London; left ventricle) The process that was undertaken in each case is outlined below.

2.1.1 Generation of the Recombinant Inbred Strains

The derivation of the RI strains is as described in (Hubner et al., 2005; Pravenec et al., 1989). Thirty-six RI strains were bred by systematic inbreeding and sibling mating for at least 20 generations (Pravenec, 1996) from the initial cross of BN.Lx/Cub and SHR/Ola inbred parental strains by Michal Pravenec and Vladimir Kren at the Czech Academy of Sciences in Prague. Of these 36 strains, four went extinct, two were not available for generation of eQTL data, and one was subsequently removed from the analysis due to loss of homozygosity, possibly due to breeding error. The remaining 29 were used in downstream analysis.

2.1.2 Generation of the eQTL Dataset

The eQTL dataset consists of thousands of expression traits that have been measured in the RI strains and mapped to the genome by linkage analysis and been found to be statistically significant following correction for multiple testing. The procedures that were undertaken to generate the dataset were undertaken prior to the commencement of this thesis, and are described in detail in (Hubner et al., 2005) and (Petretto et al., 2006a). They are summarized in **2.1.2.1-2.1.2.3**.

2.1.2.1 Microarray Expression Profiling

Measurements of mRNA abundance were obtained in four tissues (fat, kidney, adrenal and left ventricle), as described in (Hubner et al., 2005) from four biological replicates in each of the original 30 RI strains by microarray (as discussed above, one strain was subsequently removed from the analysis). Each of the 120 microarrays used for each tissue experiment (4 replicates × 30 RI strains) was processed and normalized separately (Petretto et al., 2006a). Outliers were removed using the Nalimov outlier test ($p < 0.001$), and the raw values averaged across the four biological replicates in each strain using the Robust Multichip Average (RMA) algorithm to give expression summary values. The expression value used in downstream analysis is the anti-log of the RMA output values (Hubner et al., 2005). For the fat, kidney and adrenal tissues RAE 230A Affymetrix GeneChips, which assess the expression levels of 15,923 transcripts across the genome, were used. For left ventricle Rat230_2 Affymetrix GeneChips, which assess the expression levels of

31,099 transcripts, were used.

2.1.2.2 Marker Genotyping

1,011 autosomal microsatellite markers were used in the construction of the genetic linkage map used in the subsequent generation of the eQTL dataset. The construction of the linkage map in the BXH/HXB RI strain panel was as described in (Jirout et al., 2003; Pravenec, 1996). To summarize, DNA was obtained from each strain, and genotyping was achieved by PCR analysis and/or Southern blotting (Pravenec, 1996).

2.1.2.3 Linkage Analysis

eQTLs represent regions of the genome that are linked to an expression trait. The marker map was constructed using MapMaker (Lincoln et al., 1993), incorporating input from Ensembl (Hubbard et al., 2002) and BLASTN alignment in order to position the genotyped markers as accurately as possible on the genetic map. Following the construction of the map, the genome-wide linkage analysis of expression traits was performed using QTL Reaper (Manly, 2006). This software, whose analysis is based on that implemented in the earlier Map Manager QTX system, also by Ken Manly (Manly et al., 2001) assesses linkage between each expression trait and each marker by regression, finding a likelihood ratio statistic (LRS) for each probeset-marker combination. The genome-wide statistical significance was found empirically by permutation as described in (Hubner et al., 2005). Quoted *p*-values of eQTL significance are empirical *p*-values obtained by comparing the LRS score with the distribution of at least 1,000 null (permuted) LRS scores for that probeset-marker combination.

2.1.3 The eQTL Database

The eQTL data, including probeset and linkage mapping information, eQTL statistics (e.g. LOD score and heritability), probeset expression, and marker genotype data was stored in a MySQL database, held on the Codon server maintained by the Imperial College Centre for Bioinformatics. eQTL Explorer (Mueller et al., 2006) was used to produce visual representations of the *cis*- and *trans*-eQTL data held in the MySQL database.

2.1.4 Interpretation of eQTL Data

A variety of data indirectly related to the eQTL mapping process summarized in **2.1** was recorded in the eQTL database and is used to inform downstream analysis of the data therein:

2.1.4.1 Fold Change

The term ‘fold change’ refers to a ratio in transcript abundance (such as measured by a microarray) between two populations. Fold change between the parental strains was found for each probeset by dividing the average expression across the biological replicates in SHR with that in BN. The fold change can be used as an indicator of the size of the eQTL effect (Petretto et al., 2006a).

2.1.4.2 Heritability Analysis

Heritability (h^2) refers to the proportion of the total variance in a population that can be attributed to genetic factors, as opposed to environmental factors (Visscher et al., 2008). In the RI strains, the measure of heritability (h^2_{trait}) was found for each probeset as described in (Petretto et al., 2006a) using the following formula:

$$h^2_{trait} = 0.5 \frac{V_A}{0.5V_A + V_R}$$

Where V_A represents the variance of the mean expression levels across the strains, and V_R represents the mean of the variances within each strain.

2.2 Validation of Data

A wide variety of factors must be taken into consideration in order to achieve meaningful outcomes from the downstream analysis of an eQTL dataset (Kendziorski and Wang, 2006). Some validation of the biological and computational components of the dataset and its construction, including quality control of the microarray gene expression data (described in (Hubner et al., 2005)) and accounting for multiple testing of the eQTLs by calculating the False Discovery Rate (described in (Petretto et al., 2006a)), was carried out prior to the commencement of this thesis. Methods used in the computational validation and filtering of the eQTL dataset discussed in Chapter 3 (**3.1**) are described in more detail in this section.

2.2.1 Validation of the eQTL Mapping

The methods described below were used to validate the eQTL mapping through comparison of the output of QTL Reaper with those of other methods. The structure of the data is described in more detail in order that the methods used and the motivations for doing so can be explained as clearly as possible.

2.2.1.1 Statistical Methods of Comparative Linkage Analysis

The validation of the QTL Reaper output was carried out by comparison with two other methods of linkage analysis: Map Manager QTXb – a software package by the same authors as QTL Reaper (Manly et al., 2001), and an implementation of the Mann-Whitney-Wilcoxon two-sample test.

Map Manager QTXb is used to map gene expression to a marker map. It enables a wide variety of analysis methods and assessments of statistical significance to be used to map QTLs. However it lacks the functionality to perform a batch analysis, meaning that full-scale validation of QTL Reaper findings using this software was impractical.

A pointwise linkage analysis method was developed using Perl and R. This enabled a measures of linkage for each probeset-marker combination to be quickly and efficiently obtained, with minimal complications. This method applies the Mann-Whitney-Wilcoxon test and provides a nominal, indicative *p*-value for each combination.

2.2.1.2 The Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon test, sometimes referred to as the Mann-Whitney *U* test or Wilcoxon rank-sum test, is a non-parametric equivalent of the two-sample *t*-test. It is an established independent test of linkage of transcript expression levels to genetic markers (e.g. (Yvert et al., 2003)). A test statistic (*U*) is found by ordering the expression levels, then comparing their rank sums corresponding to each of the two groups n_1 and n_2 where R_1 is the sum of the ranks of the expression levels in group 1. The formula used to obtain *U* is given below:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

In the analysis described in **3.1**, the Mann-Whitney-Wilcoxon test is used to compare the expression levels for each of the two groups corresponding to the SHR and BN genotype across the 29 RI strains. The R implementation *wilcox.test* was used for this analysis.

2.2.2 Validation of Probeset Annotations

The accuracy of the mapping of microarray probesets to the genome is critical to the analysis of the eQTL dataset that is described in chapters 4 and 5. The annotations provided by Affymetrix were compared to the alignment of the sequence with the EnsEMBL (Hubbard et al., 2002) assembly. The software used to do this was developed in-house, led by Nicholas Dickens and Jonathan Mangion (then of MRC Clinical Sciences Centre), and is called SCAMPA (Systematic Categorisation of Affymetrix Microarray Probeset Annotations) (Dickens, 2006).

2.2.2.1 SCAMPA

SCAMPA is a Perl pipeline that categorises probesets on the basis of two criteria: The consistency of the alignment of the probeset sequence provided by Affymetrix to the genome in EnsEMBL, and the consistency of the gene annotation of the probeset between the Affymetrix and EnsEMBL. The probesetes were categorised by alignment into tiers:

1) All probes map to the same region of the genome
2) A minority of the probes map to a different region of the genome to the rest
3) Probes map to multiple genomic locations
4) The majority of the probes do not map to the genome at all

The annotation categories, from A to E, indicate the degree of agreement between the Affymetrix and EnsEMBL annotations.

2.2.3 Filtering of the eQTL Dataset

Only probesets found to map consistently to the genome (i.e. those in SCAMPA tiers 1 and 2) were used in the analyses described in chapters 4 and 5. Probesets that did not meet this criterion, had missing expression data or for which the SDP of the marker at the peak of linkage had more than 10 missing genotypes were removed from the analysis (**Table 3.2**). Reannotation of these probesets was not attempted for reasons of time.

2.3 Detecting Polymorphisms in Probe Binding Regions

The measurements provided by using a microarray such as an Affymetrix GeneChip to study gene expression are produced by quantifying the hybridisation of oligonucleotide probes on the chip with mRNA in samples of each tissue. The underlying assumption is that the level of hybridisation is representative of the quantity of mRNA present. However, it has been shown that probe signals can vary as a consequence of polymorphisms (e.g. SNPs) in the regions of the transcribed target gene that bind to the probe, resulting in a misleading indication of differential expression between the sample populations (Alberts et al., 2007a). Methods used in the study of polymorphisms in probe binding regions discussed in Chapter **3** (**3.2** and **3.3**) are described in more detail in this section. It should be noted that the findings of this study were not used to filter the sets of eQTLs used in chapters **4** and **5**.

2.3.1 Sequencing of SHR Genome

The SHR genome was sequenced by Santosh Atanur and collaborators (the SHR Genome Sequencing Consortium) using massively parallel sequencing technology from Illumina/Solexa (Atanur et al., 2010); the sequencing methodology is as applied in (Bentley et al., 2008). At the time of commencement of the analysis described in this thesis, sequencing provided 4.5X coverage of the SHR genome.

2.3.2 Identification of SNPs in Regions of Interest

SHR sequence data was related to the EnsEMBL (Flicek et al., 2008) BN reference sequence by Santosh Atanur using MAQ, a freely available software package for building mapping assemblies (*http://maq.sourceforge.net/*). 2.3 million SNPs were identified between SHR and BN in this way. On validation by capillary sequencing, 96% of a sample of predicted SNPs were confirmed and the false positive rate of SNP prediction was found to be less than 1%.

Affymetrix probes were mapped to the SHR genome in collaboration with Santosh Atanur and Liqun He using RMAP v0.2 (*http://rulai.cshl.edu/rmap/*), an R package which maps sequencing reads to their genomic location. Probesets in which all of the probes mapped consecutively to the same chromosome were used in the analysis described in **3.3**. Probesets whose probes mapped to the genome with more than 2 mismatches or mapped to multiple locations on the genome were not used in this analysis.

2.3.3 AffyGG

AffyGG is a computational method of predicting *cis*-eQTLs that occur due to polymorphisms in the probe binding region and are therefore not indicative of variation in gene expression (Gilad et al., 2005), sometimes referred to as potential false *cis*-eQTLs (Alberts et al., 2007b). It was developed by Rudi Alberts, Gonzalo

Vera and Ritsert Jansen. The method, implemented in an R package, is based on comparing the signals from individual probes with those across the whole probeset (Alberts et al., 2008).

AffyGG's *probeElimination* procedure works by combining individual probe signals and genotype data in an ANOVA-based statistical 'multiprobe' model (Alberts et al., 2005) to identify probes within the probeset having a disproportionate effect on the QTL linkage. The probe having the greatest such effect (referred to as the 'probe by QTL' effect) is eliminated from the probeset ('dropped') in an iterative fashion. AffyGG repeats this process until 9 of the 11 probes have been 'dropped'.

The principle underlying the AffyGG methodology is that because 'false *cis*-eQTLs' are caused by polymorphisms in the probe binding regions of one or more of the probes in the probeset, the discounting of these probes will cause the QTL effect to evaporate. However, it has not previously been possible to objectively assess whether or not the probes that are eliminated by AffyGG do in fact have polymorphisms in their probe binding regions, and the outputs produced by the AffyGG *probeElimination* procedure are therefore subject to interpretation.

2.3.4 Support Vector Machine (SVM)

The Support Vector Machine is a flexible binary classification method that applies a 'support vector', which is not necessarily linear, to achieve optimal categorisation

(Cortes and Vapnik, 1995). The R implementation *svm* contained within package *e1071* (Karatzoglou et al., 2006) was used with the linear kernel in the analysis described in **3.3.2**, in order to discriminate between two groups of *cis*-eQTL transcripts that had previously been identified (**Figure 3.6**) as displaying an approximately linear relationship with respect to two variables.

2.4 Co-expression Analysis

Methods used in the co-expression analysis of eQTLs that was performed in the sets of *cis*-eQTLs, *trans*-eQTLs and in *trans*-eQTL clusters in all four tissues in the eQTL dataset and is described in Chapter **4** are described in more detail in this section..

2.4.1 Correlation Methods

Correlation analysis provides a means of quantifying the strength of the linear relationship between two variables. In statistical genetics, correlation-based methods are widely applied to investigate possible relationships between genetic and phenotypic traits.

2.4.1.1 Pearson Correlation

The Pearson correlation coefficient (r) – sometimes referred to as the product moment correlation coefficient – is calculated in order to quantify the linear relationship between two sets of paired variables *x* and y, such as a pair of expression profiles obtained from the same panel of RI strains, by the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The Pearson correlation coefficient can range from -1 to 1, where -1 indicates the maximum possible negative linear relationship between the variables, 0 indicates no linear relationship, and 1 indicates the maximum possible positive linear relationship. The significance of this relationship can be assessed by application of a *z*-test to the coefficient given the sample size, or (as was done in the work described here) empirically by permutation. The implementation of Pearson correlation contained within MatLab function *corrcoef* was used in the co-expression and QTT analyses described in chapters **4** and **5** of this thesis.

2.4.1.2 Point Biserial Correlation

Point biserial correlation is a method of quantifying the linear relationship between a continuous and a dichotomous (binary) variable, such as an expression profile and a set of homozygous genotypes. The point biserial correlation coefficient r_{pb} is a special case of the Pearson correlation coefficient (Kline, 2004), and is found by applying the following formula, where *X* is the continuous variable, *Y* is the dichotomous variable, and π is the proportion of values where Y=1:

$$r = \frac{(\overline{X}_1 - \overline{X}_0)\sqrt{\pi(1-\pi)}}{S_X}$$

The implementation of point biserial correlation *biserial.cor* contained within the R package *ltm* was used to obtain coefficients for use in the partial correlation analysis (**2.4.5**) described in **4.2.3**.

2.4.2 Genome-wide Correlation of Gene Expression Levels

The application of Pearson correlation to test for linear relationships between the transcript expression profiles of all pairs of distinct *cis*-eQTL and *trans*-eQTL transcripts was carried out using a pipeline developed in Perl and MatLab. The Perl elements of the pipeline utilise the standard DBI module to obtain data from the eQTL database, and filter for SCAMPA mapping tier as described in **2.2.3**.

MatLab was used to carry out the pairwise Pearson correlation of probeset expression levels and perform the permutation analysis required to generate the empirical *p*-value for the correlation of each pair. The methodology used for the permutation analysis is outlined in **2.5.3**.

2.4.3 Assessment of the Similarity of Genotypes

It was considered that the relationships between eQTLs identified through co-expression analysis are at least in part a consequence of relationships between the underlying genotypes. To assess this, correlation of the genotype data was performed for each pair of co-expressed eQTLs. This was carried out using a simple matching coefficient method implemented in Perl. For a pair of SDPs, a ‘match’ (i.e. a given strain having the same genotype in SDP 1 as in SDP 2) was scored as 1, a ‘mis-match’ was scored as 0, and the total across the 29 strains was output as the matching coefficient. Missing genotypes were disregarded, since it would not be possible to differentiate a ‘match’ from a ‘mismatch’ in strains in which one of the two genotypes being compared was not available. The empirical *p*-value for the observed number of matches, given the number of genotypes in the SDPs being compared (excluding any strain in which there was a missing genotype in either SDP) is assessed as described in **2.5.4**.

2.4.4 Distance Analysis of Candidate Regulators of *trans*-eQTL clusters

The *trans*-eQTL clusters observed in the eQTL dataset were found to be co-expressed and predicted to arise as a result of co-regulation by a regulatory gene at or near the peak of linkage of the cluster.

The *cis*-eQTLs in the ‘window region’ of the peak of linkage, conservatively defined as 50Mb either side of the peak in order to account for inaccuracies in mapping due to linkage disequilibrium, were tested for correlation with the cluster using the same method as was used for the global co-expression analysis.

Genetic distance (cM) was incorporated into this analysis as a covariate. For each cluster, the genetic distance between the peak of linkage and each of the *cis*-eQTLs in the window region was found. This distance was plotted against the averaged absolute correlation coefficients of the expression profile of each of the *cis*-eQTL transcripts with each of the eQTLs making up the *trans*-eQTL cluster. Simple linear regression of the distance against the averaged absolute correlation coefficient was then performed on the data in SPSS. An R^2 value, which gives the proportion of the total variance between the two variables that is explained by the regression model, was found:

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

The analysis as described was also performed using all probesets mapping to the conservative window region (as opposed to just *cis*-eQTLs).

2.4.5 Partial Correlation Analysis

First order partial correlation analysis is used to find the correlation of two variables *x* and *y* after the effect of a third variable *z* has been taken into account. The method implemented is as described in (Mozhui et al., 2008) and was performed by finding the partial correlation coefficient $r_{xy.z}$ in the formula given below:

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

In the above partial correlation formula, as applied to a pair of expression profiles within a *trans*-eQTL cluster, r_{xy} is the Pearson correlation coefficient (**2.4.1.1**) of the expression profiles and r_{xz} & r_{yz} are the point biserial correlation coefficients (**2.4.1.2**) of each expression profile with the SDP of the marker at the cluster's peak of linkage.

A two-tailed *t*-test was used to assess the significance of each pairwise partial correlation coefficient conditioned on the genotype at the peak of linkage of the *trans*-eQTL cluster. *t*, with *n-2-k* degrees of freedom was found using the following formula.

$$t_{n-2-k} = r\sqrt{\frac{n-2-k}{1-r^2}}$$

In this formula, *r* is the partial correlation coefficient, *k* is the number of conditioning variables (in first order partial correlation analysis, *k* = 1), and *n* is the number of data points (here, *n* is the number of RI strains with genotype data in the SDP of the marker at the *trans*-eQTL cluster's peak of linkage). Where *n* = 29 (i.e. no missing genotypes in the SDP), the two-tailed *t*-test is performed with 26 degrees of freedom. The significance threshold corresponding to the relevant number of degrees of freedom was then found using a lookup table.

2.5 Assessment of Statistical Significance

The methods used to assess the statistical significance of the findings of the co-expression and genotype correlation analyses described in the previous section, including methods used to correct for multiple testing, are outlined in this section. Their application to the datasets produced by the correlation analyses is described.

2.5.1 Permutation Testing

Permutation testing is a method of finding a threshold of statistical significance of QTL linkage by randomly shuffling the trait values a large number of times and comparing the observed data to the null data (Churchill and Doerge, 1994). It enables an empirical *p*-value for the significance of an observation that inherently takes into account the distribution of the traits to be found, rather than a *p*-value, subject to assumptions, being estimated on the basis of a statistical model. The method by which permutation testing was applied in the co-expression analysis is outlined in **2.5.3**.

2.5.2 Multiple Testing Correction Methods

The genome-wide analyses described here involve the simultaneous testing of hundreds of thousands of hypotheses. The application of standard tests, such as Pearson correlation, in such a way results in the prediction of large numbers of false positive findings (Manly et al., 2004). This is accounted for in statistically sound studies by the application of a multiple testing correction method.

Multiple testing correction methods used in genome-wide studies typically control either the familywise error rate (FWER) – using methods such as the Bonferroni correction – or the false discovery rate (FDR). The FWER-controlling methods, which control for the probability of making one or more false discoveries for a given trait, have been found to be overly conservative for genome-wide studies, causing variants having all but the most extreme effect sizes to go undetected (Benjamini and Yekutieli, 2005; Kendziorski and Wang, 2006), so FDR-controlling methods are generally preferred (Manly et al., 2004).

2.5.2.1 The False Discovery Rate (FDR)

The FDR was originally described in (Benjamini and Hochberg, 1995) as *"the proportion of the rejected null hypotheses that are erroneously rejected"*, i.e. the proportion of false positives among all features called significant.

The Benjamini-Hochberg procedure, a step-down method described in the 1995 paper, was the first devised to control for FDR. An empirical Bayesian equivalent to the Benjamini-Hochberg was described by (Efron and Tibshirani, 2002).

An adjustment to the approach to controlling for FDR described by Benjamini & Hochberg was proposed by John Storey (Storey, 2002). This method, along with others proposed around the same time (Manly et al., 2004), differ from that of

Benjamini & Hochberg in that they make use of a method of estimating the proportion of true null hypotheses (often referred to as π_0 or m_0).

The method described in full in (Storey, 2003a) and implemented in the Q-value procedure makes use of a concept called the positive False Discovery Rate (pFDR), which is conditioned on at least one positive finding having occurred. This procedure for controlling the FDR, which estimates π_0 from the distribution of the *p*-values has been shown to give substantially greater power to reject the null hypothesis when it is false when the proportion of true alternative hypotheses is high (Manly et al., 2004).

2.5.2.2 The Q-value Method

The *q*-value procedure provides a significance threshold that corresponds to the proportion of significant features that turn out to be false positives (Storey and Tibshirani, 2003). If the assumptions implicit to the methodology, e.g. that *p*-values will be approximately evenly distributed from 0 to 1 (Xiao, 2004), are taken into account then Q-value is a robust and straightforward approach to multiple testing correction. It is particularly suited to genome-wide studies as, although all FDR-based methods use a working assumption of independence of the underlying variables (Storey, 2002), it been shown to be conservative under 'weak dependence' of the considered features (Petretto et al., 2006b). 'Weak dependence' occurs when the number of correlated features is small relative to the total number of queried features (Pounds, 2006). Small groups of genes that behave nonrandomly (e.g. *trans*-eQTL clusters) can be considered to fit into this category (Petretto et al., 2006b). Software used to implement this method, QVALUE, is available as a stand-alone package (Storey, 2003b) or as an R package from CRAN (Comprehensive R Archive Network) (Dabney and Storey, 2008)

2.5.3 Assessment of statistical significance in co-expression analysis

An implementation of the permutation method described by (Churchill and Doerge, 1994) was used to determine the significance of all Pearson correlation coefficients of pairs of expression profiles in co-expression analysis, and also in QTT analysis (**2.7**). The MatLab script used to implement this method, along with sample input files, are provided in ***Appendix I***.

Briefly: Pearson correlation was performed, within a matrix, of all pairs of interest (e.g. all *cis*-eQTLs in a tissue), using MatLab's *corrcoef* function. Following this, for each of the pairs, 10,000 'null' co-expression matrices were generated by repeatedly performing Pearson correlation, again using *corrcoef*, after the expression traits were randomly shuffled across the RI strains. Following this, for each pair the observed correlation coefficient was ranked among the null correlation coefficients. A matrix of actual correlation coefficients, referred to henceforth as the 'observed matrix' is constructed, whose dimensions correspond to the number of eQTLs under analysis. A matrix of nominal *p*-values is also created by the MatLab script, but these were not used in further analysis.

Following the construction of the ‘observed matrix’ of correlation coefficients, the correlation analysis is repeated many times with the expression measurements having been shuffled across the strains, to generate a set of ‘null’ matrices of correlation coefficients.The empirical pairwise *p*-value corresponding to a Pearson correlation coefficient corresponds to the ranking of the observed correlation coefficient as follows:

$$p = \frac{rank}{10{,}000}$$

Thus, a ranking of 500th among the 10,000 (actually 10,001, but for simplicity’s sake 10,000 was used in calculating the empirical *p*-values by this method) would correspond to an empirical *p*-value of 0.05.

In order to correct for multiple testing across the genome and threshold the *p*-values obtained as above in a manner that takes into account the multiple testing across the genome, an FDR-based multiple testing correction, the QVALUE method (**2.5.2.2**) (Storey and Tibshirani, 2003), was applied to these empirical *p*-values. A *q*-value threshold of $q = 0.05$ was used in this analysis.

2.5.4 Assessment of statistical significance in correlation of genotypes

Significant correlation or anti-correlation of genotypes was defined on the basis of a probability table (***Appendix II***), in the construction of which the probability of each possible coefficient occurring by chance was calculated. The significance of a matching coefficient is dependent on the number of matches that have been tested (i.e. the greater of the two numbers of missing genotypes in the two SDPs being correlated). Significance of correlation and/or anti-correlation was defined as a two-tailed empirical probability of $p<0.1$.

2.5.5 Assessment of statistical power of co-expression analysis

The statistical power and precision of the co-expression analysis were assessed for *cis*- and *trans*-eQTLs in the four tissues using the tests for power and confidence interval (CI) of correlation implemented in PASS (Power And Sample Size) 2008. Full descriptions of the test procedures used are given in *http://www.ncss.com/pdf/manuals/PASSUG3.pdf*, chapters 800 and 801 respectively; the methods are summarised below:

2.5.5.1 Power Calculations for Co-expression Analysis

The statistical power (**1.8.3**) of a correlation analysis can be found by the formula given below, where r is the sample correlation coefficient; $R(r|n,\rho)$ is the area under the correlation density curve to the left of r, N is the sample size, ρ is the population correlation, and α is the significance level for one correlation.

The power of the significance test $\rho_1 > \rho_0$ is given by first solving:

$$1 - R(r_\alpha \mid N, \rho_0) = \alpha$$

Followed by computing the power (1-β):

$$1 - \beta = 1 - R(r_\alpha \mid N, \rho_1)$$

In the analysis using this method that is presented in **4.2.1**, the sample correlation coefficient used for the power and CI analyses was chosen to be approximately equivalent to the threshold $q = 0.05$ for each co-expression analysis. The significance level for one correlation $\alpha = 0.05$ was used.

The confidence interval (CI) is an interval defined, based on a specified probability α, sample size and sample correlation coefficient, such that the population correlation coefficient will occur within the interval 1-α% of the time. Where α=0.05, the proportion of intervals that will include the true population coefficient is 95%.

2.6 Functional Analysis of Co-expressed or Co-incident Genes

Computational analyses were performed on sets of transcripts that the co-expression analysis described in chapter **4** found to be related by expression correlation and colocalisation, in order to investigate the possibility of functional relatedness among these transcripts.

2.6.1 Gene Ontology (GO)

The Gene Ontology is a system of annotation, which facilitates the analysis of genes by providing a controlled vocabulary, meaning that related genes can easily be identified by a computer program through the presence of shared 'GO terms' (Ashburner et al., 2000).

The GO is divided into three categories: Biological Process, which describes the biological objective to which the gene contributes (e.g. 'transport'); Molecular Function, which describes the biochemical activity to which the gene contributes (e.g. 'aminomethyltransferase activity'; and Cellular Component, which describes the cellular location in which the gene is active (e.g. 'nuclear membrane'). Not all GO annotated genes have annotations for all three categories.

The ontology takes the form of a directed acyclic graph (DAG), which relates terms of different degrees of specificity, linking 'low-level' (more general) to high-level (more specific) terms. Annotation to a specific term implies annotation to all of the 'parent', more general terms (Lewin and Grieve, 2006).

2.6.2 DAVID

DAVID (Database for Annotation, Visualization and Integrated Discovery) is a web-based tool for functional annotation of gene/protein sets that makes use of Gene

Ontology annotations as well as providing options of taking into account other forms of functional information such as protein and pathway annotation (Dennis et al., 2003).

In this project, DAVID was used to assess functional enrichment in *trans*-eQTL clusters by testing for statistically over-represented GO terms, using the DAVID Gene Functional Classification Tool (Huang et al., 2007). This tool tests for functional enrichment of a list of gene identifiers (Affymetrix probeset identifiers were used in the analysis presented in **4.2.4**) relative to a background gene set (the DAVID *Rattus norvegicus* population background set was used in the analysis). Enrichment is quantified by the Expression Analysis Systematic Explorer (EASE) score (Hosack et al., 2003). This is found by the Functional Classification Tool by applying a conservative modification to the Fisher's Exact Test score following the comparison, within a 2x2 contingency table, of the proportion of genes in the supplied list annotated with a given GO term against the population background set (*http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#E3*).

2.6.3 The Kolmogorov-Smirnov (K-S) Test

The two-sample Kolmogorov-Smirnov (K-S) test is a non-parametric test of the hypothesis that two samples are drawn from the same distribution. The K-S test statistic *D* is found by computing the difference between the two sample distribution functions $F_{n2}(x)$ and $F_{n2}(x)$ as per the below formula (Young, 1977):

$$D = \max | F_{n_1}(x) - F_{n_2}(x) |$$

In **4.2.4**, this method was used to compare levels of functional enrichment found in 15 large *trans*-eQTL clusters with those found in random groups of *trans*-eQTL genes. The SPSS v14 implementation of the two-sample Kolmogorov-Smirnov test, which computes an exact *p*-value, was used for this analysis.

2.7 Quantitative Trait Transcript (QTT) analysis

As described in chapter **5**, QTT analysis was performed in order to test for correlations between physiological trait measurements in the RI strains and expression traits that have been mapped to the genome. This can be carried out by genome-wide correlation analysis, similarly to that performed for pairs of eQTLs and described in **2.4.2**. Specifically, for each physiological trait of interest, the measurements of the expression of genes underlying each *cis*-eQTL were correlated with the set of trait measurements by Pearson correlation, and *p*-values for each combination of trait and *cis*-eQTL transcript expression profile were found using the permutation method used for co-expression analysis described in **2.5.3** and implemented in MatLab (Petretto et al., 2008).

Chapter 3: Validation and Exploration of a Large Biological Dataset

The analyses described in Chapters 4 and 5 are dependent on the integrity of the BXH/HXB eQTL dataset. The validation and filtering of the dataset described in **3.1** was carried out prior to the commencement of the genome-wide co-expression and quantitative trait transcript studies described in chapters 4 and 5. The work described in **3.2** and **3.3** was exploratory and not used to filter the eQTL dataset.

3.1 Validation of eQTL Data

3.1.1 Introduction

The eQTL dataset was generated in each of the four tissues using 1,011 microsatellite markers. In fat, kidney and adrenal, 15,923 probesets were analysed. In left ventricle, 31,099 probesets were analysed through the use of a more advanced GeneChip (**2.1.2.1**). The set of eQTLs used in downstream analysis is delineated in **Table 3.1**. These, totalling 8,269 eQTLs across all four tissues, were identified at a genome-wide significance of $p < 0.05$ from a total of 16,098,153 possible significant linkages (eQTL) in fat, kidney and adrenal and 31,441,089 in left ventricle. At this level of significance, the tissue-specific expected FDR for *cis*-eQTLs is 4–8%, and the tissue-specific expected FDR for *trans*-eQTLs is 26-35% (Petretto et al., 2006a).

Tissue	No. of *cis*-eQTLs	No. of *trans*-eQTLs
Fat	558	923
Kidney	718	1,033
Adrenal	602	933
Left Ventricle (LV)	1,362	2,140
Total	3,240	5,029

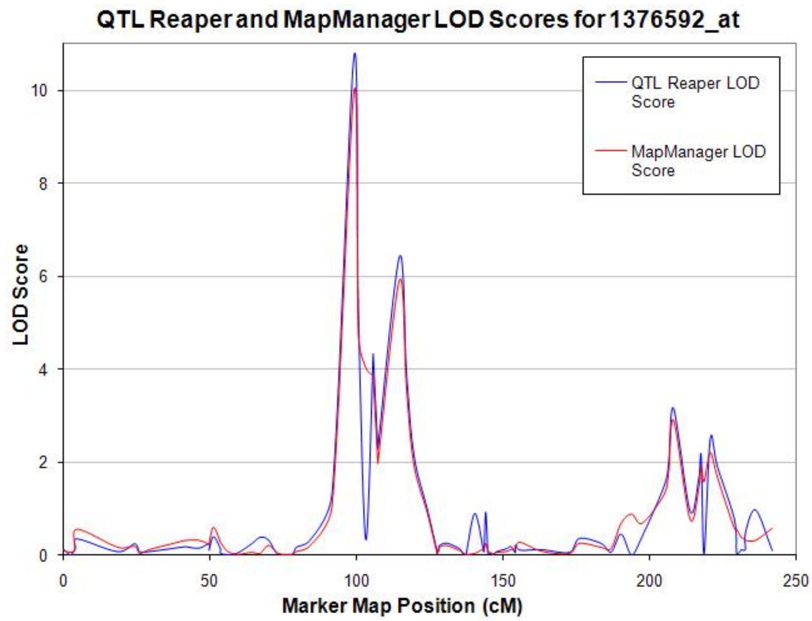
Table 3.1: Expression QTLs identified in each of the 4 rat tissues at genome-wide $p<0.05$

The size and nature of this dataset makes its analysis statistically challenging. Manual curation of a dataset this large is impossible, which means that computational analysis is essential not only to interpret and analyse the dataset, but also to ensure that the quality of the data is robust enough to enable hypothesis generation. This was achieved through the application of a series of filters and verification procedures to the gene expression and linkage data.

The linkage analysis was carried out using QTL Reaper, a software package developed by Ken Manly of the University of Tennessee (Manly, 2006). Michelle Johnson, a colleague in the Physiological Genomics & Medicine group, observed that some of the linkages identified by QTL Reaper could not be replicated by MapManager QTX, a software program that uses a similar method to perform linkage analysis.

A sample outcome from the initial comparative linkage analysis using QTL Reaper and MapManager QTX, is provided in **Figure 3.1**. Some probesets showed good concordance in LOD scores between the outputs of the two methods, as in *a)*, while others showed poor concordance; specifically, a high LOD score was generated by QTL Reaper but no such linkage was predicted by MapManager. An example of such a probeset is shown in *b)*. This probeset 1375894_at, along with several others like it, was included in the dataset at that time as a significant eQTL because of this spurious QTL Reaper finding of ‘linkage’.

a)



b)

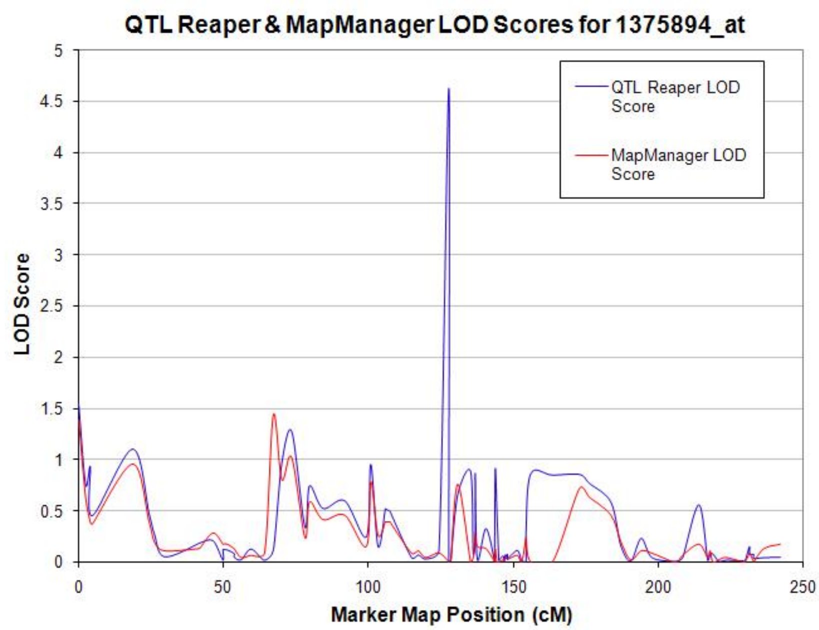


Figure 3.1: Corroboration of eQTL Reaper and MapManager QTX LOD scores on Chromosome 1 for *a)* an eQTL linked to a marker with no missing genotype data in its SDP and *b)* an eQTL linked to a marker with missing genotype data in its SDP.

The data for this figure was provided by Michelle Johnson, and it is included for illustrative purposes

I followed up the observation illustrated in **Figure 3.1** by commencing a full-scale review of the eQTL data and the generation of a new dataset, the current set, which is used in all of the statistical analyses described in Chapters **4** & **5**.

Separately from the verification of linkage analysis outcomes, study has been undertaken of the accuracy of the probeset annotations. This involved the application of a technique, SCAMPA (**2.2.2.1**) developed by a colleague, Nicholas Dickens, to test the concordance of the mapping of the probesets against EnsEMBL annotations. This was used after linkage analysis, to filter the eQTLs identified by the re-run linkage analysis. The results of this filtering are shown in **Table 3.2**.

TISSUE	CIS		TRANS	
	No. before Filtering	No. after Filtering	No. before Filtering	No. after Filtering
Fat	785	558	1149	923
Kidney	961	718	1320	1033
Adrenal	860	602	1201	933
Left Ventricle (LV)	1927	1362	2735	2140
TOTAL	4533	3240	6405	5029

Table 3.2: Outcome of filtering of eQTLs by SCAMPA 'tier' classification based on EnsEMBL mapping.

SCAMPA tiers 3 and 4 were excluded from downstream analysis.

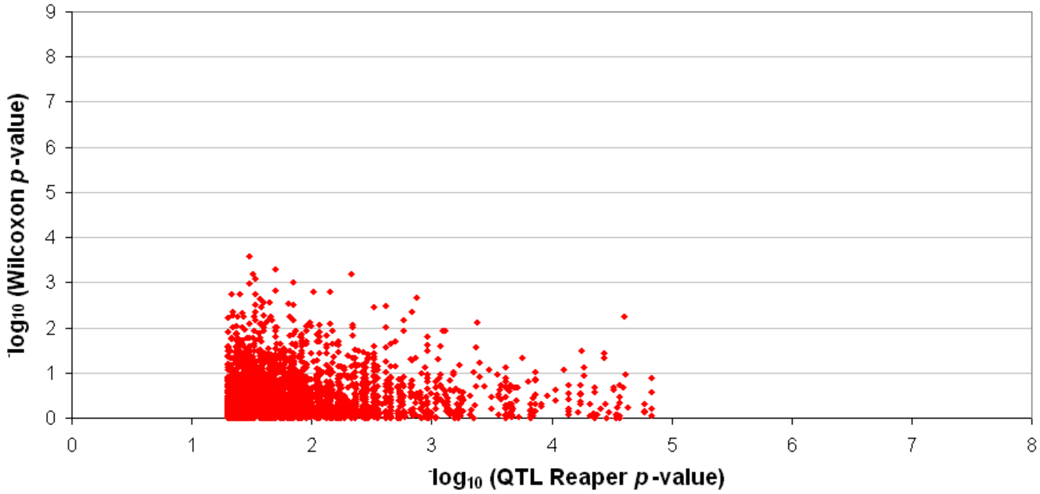
3.1.2 Results

Probesets whose linkage data failed to corroborate between the two software packages used were observed to correspond to those in which the SDP of the marker at the peak of linkage (peak marker) contained one or more missing genotypes; that is, where at least one of the then 30 RI strains could not be genotyped at that marker locus, and therefore no genotype information was provided for this strain. This led to the formulation of the hypothesis that the presence of missing data in the SDP underlies the non-corroboration of these QTL Reaper-identified linkages, and the prediction that these linkages may be spurious.

The ‘missing genotypes hypothesis’ was tested across the whole dataset by direct comparison of the *p*-value of linkage output by QTL Reaper with that produced by performing a Mann-Whitney-Wilcoxon test – a non-parametric equivalent of the two-sample *t*-test, between each eQTL probeset and peak marker. This method was used because MapManager is unsuited to large-scale ‘batch’ analysis.

The results of this analysis, illustrated in **Figure 3.2**, were clear. These scatterplots show in *a)* the outcome of this comparative analysis in eQTLs whose peak marker SDPs contained at least one missing genotype, and in *b)* the outcome of comparative analysis for those eQTLs whose peak marker SDPs contained no missing genotypes. Log *p*-values are given in order that the relationship between Wilcoxon and QTL Reaper *p*-values may be observed.

a)



b)

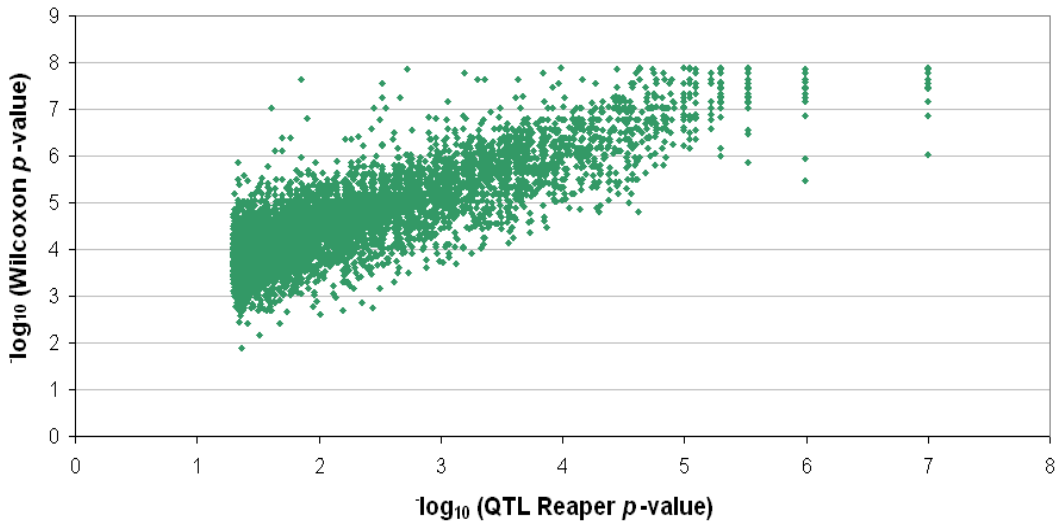


Figure 3.2: QTL Reaper *p*-values plotted against Wilcoxon Test *p*-values for all eQTLs linked to markers *a)* with missing genotype data and *b)* without any missing genotype data

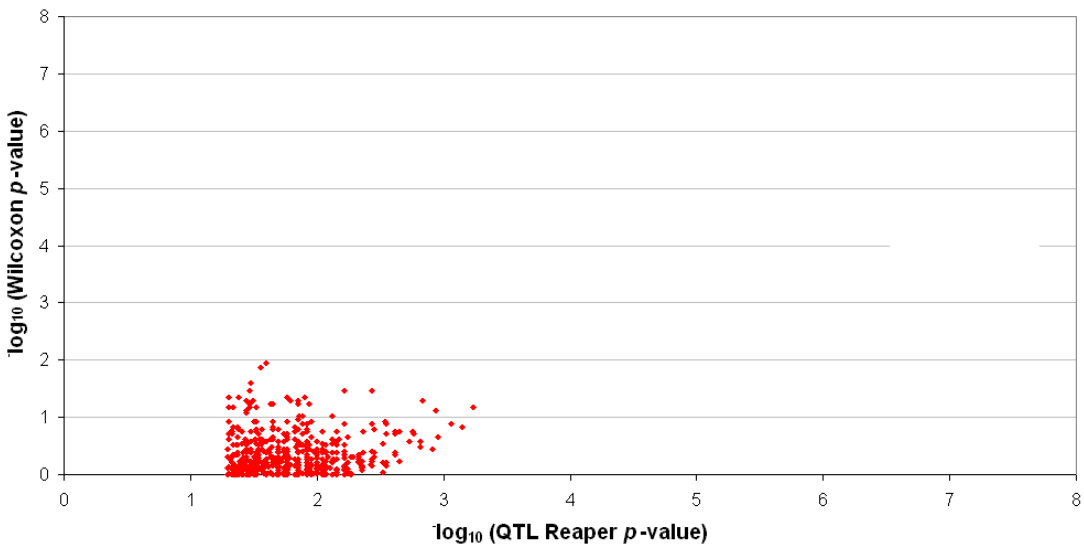
A cumulative total of 3,765 eQTLs across the four tissues had at least one missing genotype and are shown in *a)*. Of these, 254 (6.7%) had Wilcoxon $p < 0.05$, and 62 (1.6%) had Wilcoxon $p < 0.01$.

A cumulative total of 10,813 eQTLs across the four tissues had no missing genotypes and are shown in *b)*. Of these, 10,813 (100%) had Wilcoxon $p < 0.05$, and 10,812 (99.99%) had Wilcoxon $p < 0.01$. All totals are prior to any other filtering of eQTLs e.g. tiering filter.

All eQTLs have a QTL Reaper *p*-value of $p<0.05$ ($^{-}$log10 $p > 1.3$). As is apparent in *a)*, the eQTLs linked to markers with missing genotypes are irreproducible by Mann-Whitney-Wilcoxon test just as they were by MapManager; the distribution of Wilcoxon *p*-values is clearly indicative of a serious problem with the linkage data being output by QTL Reaper. In *b)*, it is similarly clear that this is not a universal problem with the QTL Reaper linkage data, but can indeed be explained by the presence of missing genotypes.

The ‘missing genotypes hypothesis’ was also tested with regard to ‘groups’ of eQTLs. These were defined differently to the *trans*-eQTL clusters that were introduced in **1.8.2.1** and will be discussed in depth in Chapter 4 – all sets of 10 or more probesets linked to a marker (regardless of whether they were *cis* or *trans*) were included in the analysis. There were 711 such sets, (termed ‘groups’ to avoid confusion with *trans*-eQTL clusters) in the dataset across the four tissues. The Wilcoxon *p*-value of linkage of each of the eQTLs in each group was found, and the percentage of corroborating (i.e. significant by Wilcoxon test) eQTLs was found for each group. The outcome of this analysis was similarly conclusive to those described previously. Of the 711 groups, 614 had 100% corroboration between QTL Reaper and Wilcoxon *p*-values, defined as a significant eQTL (all eQTLs by definition have QTL Reaper $p \leq 0.05$) with nominal Wilcoxon $p<0.05$. The other 97 had 40% corroboration or less. This fully corresponded to the set of groups whose peak markers contain missing genotypes, as illustrated in **Figure 3.3**.

a)



b)

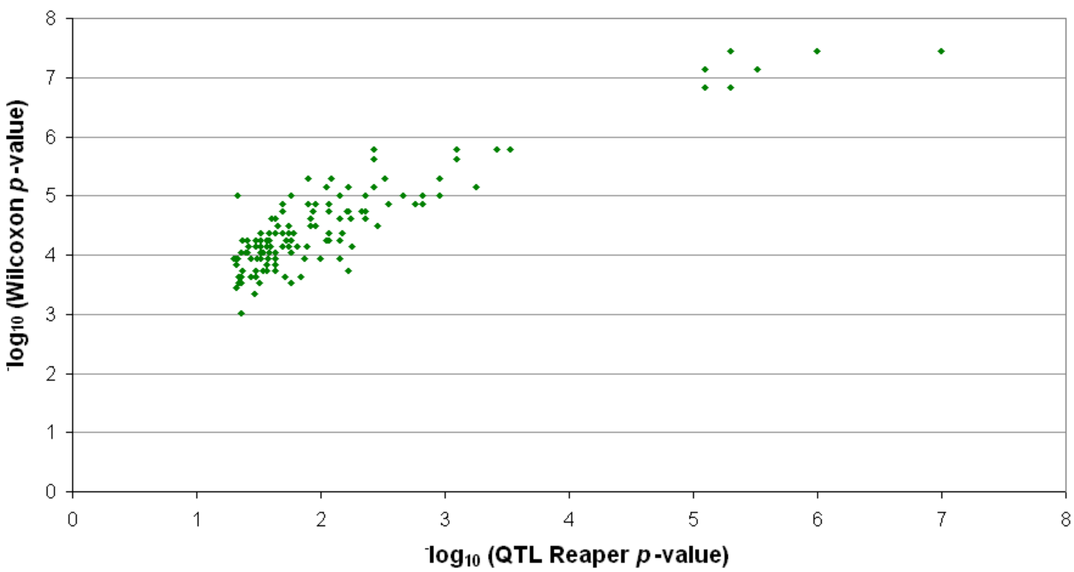


Figure 3.3: QTL Reaper *p*-values vs Wilcoxon *p*-values for eQTLs forming groups for which the marker at the peak of linkage has missing genotype data in *a)* one or more strains and *b)* none of the strains

The group shown in *a)* is linked to marker D1Cebr103s1, *b)* is linked to marker D17Rat144. Note the differing *y*-axis scales in *a)* and *b)*. The maximal Wilcoxon *p*-value for any eQTL shown in *b)* is 0.001.

Figure 3.3 shows the data used in the previous figure for two specific 'groups'. One of these, shown in *a)*, had 2 missing genotypes in its peak marker's SDP. The other, shown in *b)*, had no missing genotypes in its peak marker's SDP. Again, Log *p*-values are used for clarity of the scatterplots.

The combined outcomes of these statistical analyses provided conclusive evidence for the 'missing genotypes hypothesis', showing that the presence of any number of missing genotypes in marker SDPs led to the indication of spurious linkages by QTL Reaper which were incorporated into the eQTL dataset.

Upon investigation of the QTL Reaper software, these spurious linkages were found to have arisen as a consequence of a flaw in an algorithm handling missing genotypes, which was a discarded part of the MapManager software from which QTL was derived but became important when the software was adapted into iterative form (as in QTL Reaper). The symptions of this flaw, which caused SDPs with missing genotypes to be interpreted incorrectly and had previously gone un-noticed, were communicated to Ken Manly, the author of the software, and resolved. A patched version of QTL Reaper was made available to the host group and to other software users.

Following re-analysis of the linkage data using the new, patched version of QTL Reaper, the comparative analysis shown in **Figure 3.2** was repeated on the new dataset. The outcome of this, shown in **Figure 3.4**, demonstrated full corroboration between QTL Reaper and the Wilcoxon-Mann-Whitney results, indicating that spurious linkages were no longer present in the dataset following the correction of the problem with the linkage analysis software.

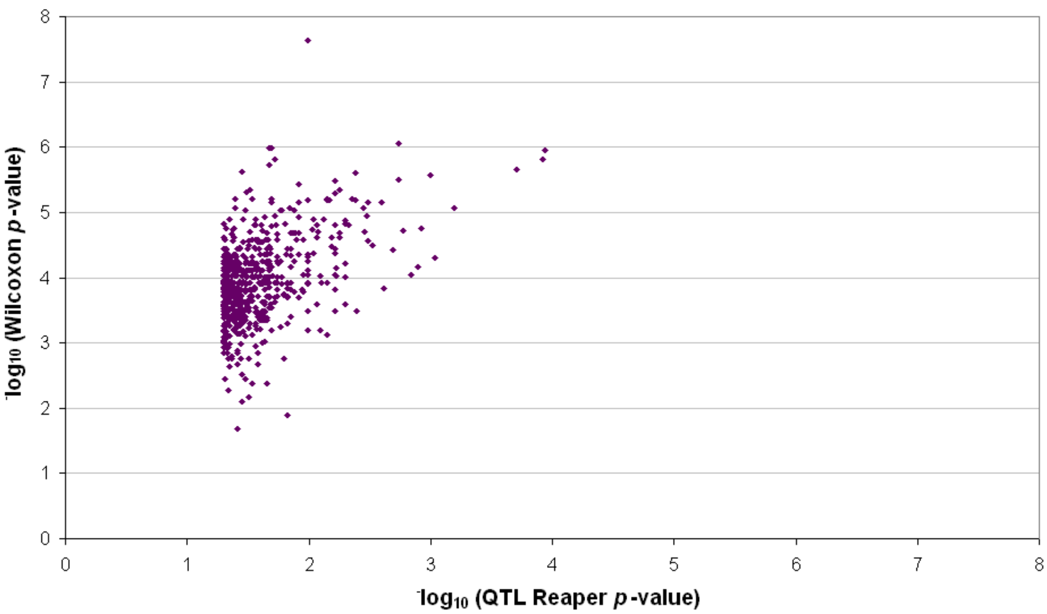


Figure 3.4: QTL Reaper vs Wilcoxon *p*-values for all eQTLs, including those with missing genotype data for one or more strains, in the eQTL dataset following the repair of the bug in QTL Reaper and the re-running of the analysis.

The outcomes of the process of data validation outlined here exemplify the crucial role that the undertaking of systematic computational quality control has played in the eQTL analysis project. The discrepancies in linkage analysis output between the original (un-patched) version of QTL Reaper and other methods concerned a substantial proportion of the eQTL dataset (25.8% of cumulative total eQTLs across the four tissues).

3.1.3 Discussion

The discoveries about the dataset made in the course of the missing genotypes analysis exemplify the importance of systematic validation of large, multifactorial datasets. The flaw in QTL Reaper that produced the spurious linkage *p*-values was identified only because the initial observation of discrepancy between the QTL Reaper linkages and those given by other methods of analysis.

Without these discrepancies having been observed and their cause identified through a combination of critical assessment of the dataset and large-scale statistical analysis, these spurious eQTLs may have gone un-noticed. Needless to say, this would have had a serious impact on the work described in the next two chapters and on that of others using the QTL Reaper software or this particular eQTL dataset.

Following these observations of the mishandling of missing data by the original version of QTL Reaper, the fully corroborated dataset generated using the patched version of QTL Reaper was used to produce the set of eQTLs used in all studies described in this thesis – the dataset delineated in **Table 3.1**.

3.2 Comparative Sequence Analysis of *cis*-eQTL Probesets

3.2.1 Introduction

Cis- and *trans*-eQTLs are classified in such a way as to reflect local and distant regulation of gene expression respectively. Gene expression is measured, as described in **1.6.1**, by assaying transcript abundance through the hybridization of complementary mRNAs to probes on a microarray, followed by the measurement of signal intensity. While this methodology is well-established and has led to many novel discoveries, it is not flawless. One potential pitfall in the interpretation of the output of DNA microarray experiments was described in (Alberts et al., 2007b).

Alberts et al found that it is possible for differences in probe signal intensity to be produced as a result of polymorphisms (e.g. SNPs) in the region targeted by the probe, even in the absence of differential gene expression between the two sample populations (i.e. those with SHR or BN genotype at the linkage region). These spurious linkages show up in the dataset as *cis*-eQTLs, and because they are not actually indicative of differentially expressed genes they are referred to as 'false *cis*-eQTLs'. Validation in mouse datasets has found around a third of the *cis*-eQTLs to be irreproducible by qRT-PCR (Doss et al., 2005).

We sought to analyse our eQTL dataset in order to identify potential false *cis*-eQTLs. To this end, we investigated *cis*-eQTLs whose target regions had been sequenced in

SHR and BN and identified those which may have been caused by polymorphisms in one or more probe binding regions. Later, work on the sequencing of the SHR genome (Atanur et al., 2010) enabled a much more thorough assessment of the levels of polymorphism in probe binding regions and therefore of the potential proportion of spurious linkages in the *cis*-eQTL dataset.

3.2.2 Results

25 *cis*-eQTLs probesets had been sequenced in BN and SHR, in order that the SNPs in the probe binding regions could be observed and quantified. These are listed in **Table 3.3**. Just under half of this (admittedly small) sample – many of which were originally sequenced for other reasons – were found to contain a SNP in at least one of the probes' target regions.

The findings from this preliminary analysis suggested that, as in the mouse datasets discussed in (Doss et al., 2005) and the datasets discussed in (Alberts et al., 2007b), a substantial proportion of the *cis*-eQTLs in the rat BXH/HXB dataset contain SNPs in at least one of their probes' binding regions.

Probeset	Tissue	No. Polymorphisms	Polymorphic Probes
1368225_at	Adrenal	2	4, 7
1373592_at	Adrenal	1	3
1374583_at	Fat	2	4, 8
1374713_at	Left Ventricle	2	1, 3
1374915_at	Fat	1	3
1376732_at	Kidney	1	4 & 5 (overlap)
1383303_at	Kidney	1	6
1384309_at	Left Ventricle	2	8, 9
1388909_at	Adrenal	2	1, 9
1392961_at	Left Ventricle	1	11
1393902_at	Kidney	1	3
1398387_at	Kidney	1	6
1370065_at	Adrenal	0	--
1370269_at	Adrenal	0	--
1370401_at	Kidney	0	--
1371137_at	Kidney	0	--
1375343_at	Kidney	0	--
1377329_at	Left Ventricle	0	--
1379398_at	Left Ventricle	0	--
1385205_at	Left Ventricle	0	--
1387376_at	Fat	0	--
1387819_at	Kidney	0	--
1388617_at	Adrenal	0	--
1392074_at	Left Ventricle	0	--
1392972_at	Left Ventricle	0	--

Table 3.3: *Cis*-eQTL probesets whose target regions were sequenced in BN and SHR, showing probes whose target regions were found to contain SNP(s).

The SHR genome was sequenced in 2008 by Solexa paired-end sequencing technology, as described in (Atanur et al., 2010). The availability of this data enabled us to systematically identify *cis*-eQTL probesets with polymorphisms in one or more of the probe binding regions.

Around half (52.2%) of the total number of *cis*-eQTL probesets in our dataset were tested – the rest were filtered according to the following criteria: *a)* all 11 probes in the probeset were required to map consistently to the same region of the genome; *b)* genetic marker data at the site of the eQTL was required to be complete across the 29 strains in the panel; and *c)* complete sequence information was required to be available in BN and SHR.

The results of this analysis are shown in **Table 3.4**, below:

Tissue	**No. *cis*-eQTLs Tested**	**% with ≥1 BN SNP**	**% with ≥1 SHR SNP**	**% with ≥1 SNP (BN and/or SHR)**
Fat	298	27.2	38.3	49.7
Kidney	377	27.1	35.0	47.5
Adrenal	316	27.5	37.0	50.0
Left Ventricle	703	27.0	35.8	48.6
TOTAL	1694			

Table 3.4: Findings of SNP analysis of *cis*-eQTLs using EnsEMBL BN and Solexa SHR sequence data

The findings of the SNP analysis were consistent across the four tissues. Around 27% of the *cis*-eQTLs contained at least one polymorphism with respect to the BN sequence, and 35-38% contained at least one SNP in SHR. In total (because some eQTLs had SNPs in both BN and SHR when compared to the Affymetrix probe sequence) just under half of the *cis*-eQTLs across the four tissues were found to contain a polymorphism in at least one of the probes' binding regions.

3.2.3 Discussion

Preliminary sequence analysis of probe binding regions suggested that around half of the *cis*-eQTLs in the dataset may have SNP(s) in at least one of the probes' target sequences. This 'prediction' was borne out by the much more thorough analysis of SNPs that was possible once the SHR sequence became available.

The number of SNPs found in BN may appear surprising, since the GeneChips used here are ostensibly based on this inbred strain. However Affymetrix probesets are designed on the basis of EST (expressed sequence tag) sequences as well as from the rat genome assembly. ESTs are typically somewhat more prone to sequencing errors (Jain et al., 2007), and their use in the development of probes may explain many of these discrepancies.

Although the proportion of *cis*-eQTL probesets with probe-binding region polymorphisms may appear large, it is important to note that not all SNPs affect hybridization - not all probesets with SNP(s) in their probe binding regions are 'false *cis*-eQTLs' (Alberts et al., 2007b). At the same time, it does suggest that levels of irreproducible *cis*-eQTLs in the BXH/HXB dataset may be similar to those found in other eQTL datasets generated using Affymetrix GeneChips, e.g. (Doss et al., 2005).

3.3 Relating predictions of software for predicting potential 'false *cis*-eQTLs' to Solexa sequence data

3.3.1 Introduction

The ANOVA-based method of predicting 'false *cis*-eQTLs' implemented in AffyGG (2.4.3) and described in (Alberts et al., 2008) has the potential to provide a means of filtering eQTL datasets to identify potential 'false *cis*-eQTLs' which are due to SNP(s) in probe binding regions rather than differential expression. However, despite this potential, it has not been adopted as a means of validating *cis*-eQTLs prior to functional or integrative analysis such as quantitative trait transcript (QTT) analysis.

One reason for this may be that the algorithm has not previously been objectively validated on a dataset in which the SNPs have been fully characterised, meaning that the predictions made by AffyGG are taken somewhat on faith. Another is that the interpretation of the numerical output (a series of *p*-values) produced by AffyGG is not straightforward and again somewhat subjective.

Here is described an attempt to address both of these issues, by making use of the SHR sequencing work carried out by the SHR Genome Sequencing Consortium (Atanur et al., 2010). The sequencing of the SHR genome (**2.3.1**) and the subsequent mapping of the Affymetrix GeneChip probes to the new sequence and to the EnsEMBL BN reference sequence (**2.3.2**) have enabled the polymorphisms in the probe binding regions of the *cis*-eQTL probesets that passed initial filtering to be fully characterised. This data was used to assess the ability of AffyGG to identify probes within *cis*-eQTL probesets that contain SNPs, as indicated by the order in which probes are ‘dropped’ (eliminated from the probeset) by the iterative *probeElimination* process (**2.3.3**).

3.3.2 Results

The analysis described here was performed in the 1,694 *cis*-eQTLs across four tissues from the BXH/HXB RI panel listed in **Table 3.4**, following the same filtering as described in **3.2.2**.

The effect on the ‘QTL by probe’ effect of the elimination of the first probe in the probeset was quantified. The ‘QTL by probe’ *p*-value output by AffyGG prior to the elimination of any probes is referred to as the ‘startint’ *p*-value, and the *p*-value following the elimination of the first probe is referred to as the ‘first pinteraction’ *p*-value. For example, for the *cis*-eQTL whose probe intensity data are shown in **Figure 3.5 a)** and which contains a polymorphism in probe #5, the ‘startint’ *p*-value given by AffyGG is $p = 1.87 \times 10^{-62}$. Once this probe is eliminated, the QTL effect evaporates: the ‘first pinteraction’ *p*-value is $p = 0.36$.

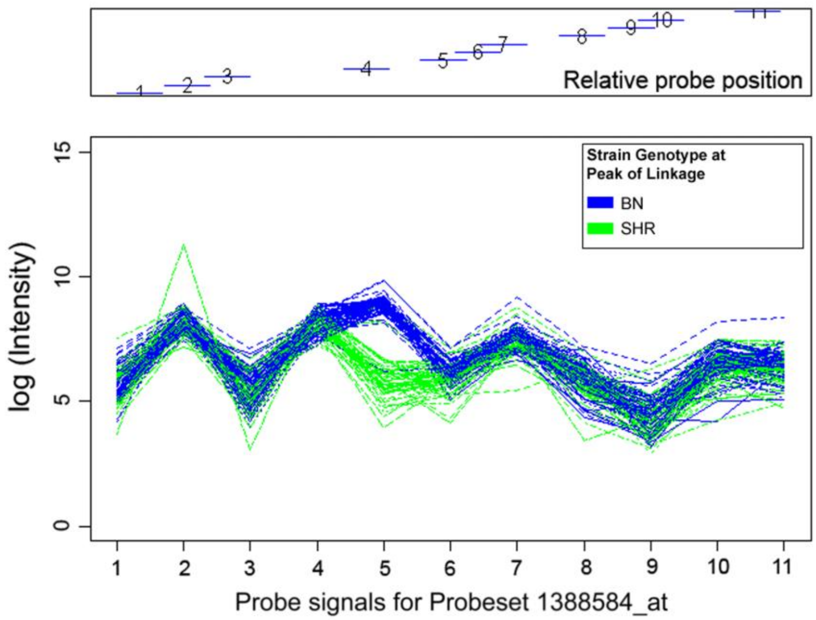


Figure 3.5 a): Probe signals plot for a *cis*-eQTL probeset containing a polymorphism in probe #5, and not displaying differential expression

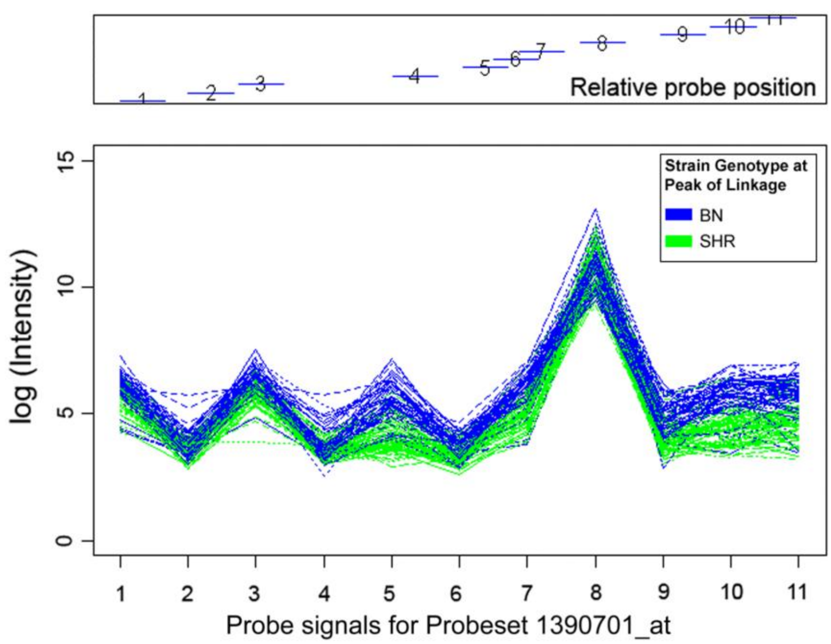


Figure 3.5 b): Probe signals plot for a *cis*-eQTL probeset containing no polymorphisms across the 11 probes and displaying differential expression across the majority of the probes

By way of contrast, the *cis*-eQTL shown in **Figure 3.5 b)** has no polymorphisms in any of its 11 probes' binding regions. Note also that the log (intensity) for the SHR samples is lower than that for the BN samples in most of the 11 probes. This indicates that this *cis*-eQTL is probably due to differential gene expression. In other words, it is likely to be a true *cis*-eQTL.

The 'startint' *p*-value for the likely 'true' *cis*-eQTL shown in **Figure 3.5 b)** is p = 1.033021e-26, and the 'first pinteraction' *p*-value is p = 2.986986e-20. The ratio between these two is far lower than for the likely 'false' *cis*-eQTL shown in **Figure 3.5 a)**. Through the interpretation of these *p*-values, AffyGG can be used to identify probes that deviate substantially from the probeset average and potentially give a misleading indication of linkage. *Trans*-eQTLs can not be 'false' in this way because the genomic location of the transcript is not the same as that the probeset maps to, therefore any probe interactions that AffyGG might 'find' are due to chance. The *trans*-eQTL datasets in each tissue therefore make ideal, tissue-specific, controls. Thresholding of the *cis*-eQTLs in each tissue as described in (Alberts et al., 2007b), based on the 99th percentile of 'startint' and 'first pinteraction' following the application of AffyGG to the *trans*-eQTLs in our dataset (**Table 3.5**) was applied to the *cis*-eQTLs in each tissue (**Table 3.6**).

Trans StartInt Percentiles			
	No. *Trans*-eQTLs	StartInt 99th Percentile	FirstInt 99th Percentile
FAT	882	2.60E-20	2.86E-12
KIDNEY	945	1.29E-24	2.14E-15
ADRENAL	1039	6.35E-24	2.00E-15
LV	1972	1.32E-29	2.65E-10
Total/Average	4838	6.50E-21	6.70E-11

Table 3.5: 99th percentile threshold 'startint' and 'firstpinteraction' AffyGG *p*-values obtained by applying AffyGG algorithm to *trans*-eQTLs in each of the four tissues.

No. Cis Above Threshold			
	No. *cis*-eQTLs	No. (%) StartInt > 99th Percentile	No. (%) FirstInt > 99th Percentile
FAT	298	81 (27.2%)	66 (22.1%)
KIDNEY	377	60 (15.9%)	46 (12.2%)
ADRENAL	316	62 (19.6%)	41 (13.0%)
LV	703	101 (14.4%)	133 (18.9%)
Total	1694	304 (17.9%)	286 (16.9%)

Table 3.6: Numbers of *cis*-eQTLs in each of the four tissues with 'startint' and 'first pinteraction' *p*-values exceeding the threshold

These *cis*-eQTLs were predicted as potential 'false *cis*-eQTLs' at the specified threshold of significance

As shown in **Table 3.6**, using the threshold of significance suggested, the 99th percentile of the probe by QTL effect before any probes have been dropped (i.e. 'startint'), almost 18% of *cis*-eQTLs were flagged as potential 'false' *cis*-eQTLs. Using the 99th percentile of the 'first pinteraction' effect instead, almost 17% of *cis*-eQTLs were so flagged.

However, this analysis was uninformed by sequence data. With knowledge of the SNP content of each probeset's probes' binding regions, it was possible to consider just how meaningful these thresholds are in terms of identifying potential 'false' *cis*-eQTLs.

We found that, of the 14.4-27.2% (across the four tissues) of *cis*-eQTLs flagged up by applying the thresholds given in **Table 3.5**, 73.3-80.2% contained at least one polymorphism between BN, SHR or both and the reference sequence (Data Not Shown). However, 39.1-43.4% of probesets that were not flagged up also contained at least one polymorphism. It was subsequently observed that the ratio of the QTL by probe *p*-value prior to the dropping of any probes and following the dropping of the first probe may be a more consistent indicator of potentially false *cis*-eQTLs. This is shown in **Figure 3.6**, upon which the ratio between 'startint' and 'first pinteraction' *p*-value is plotted for each of the 703 *cis*-eQTLs in the left ventricle.

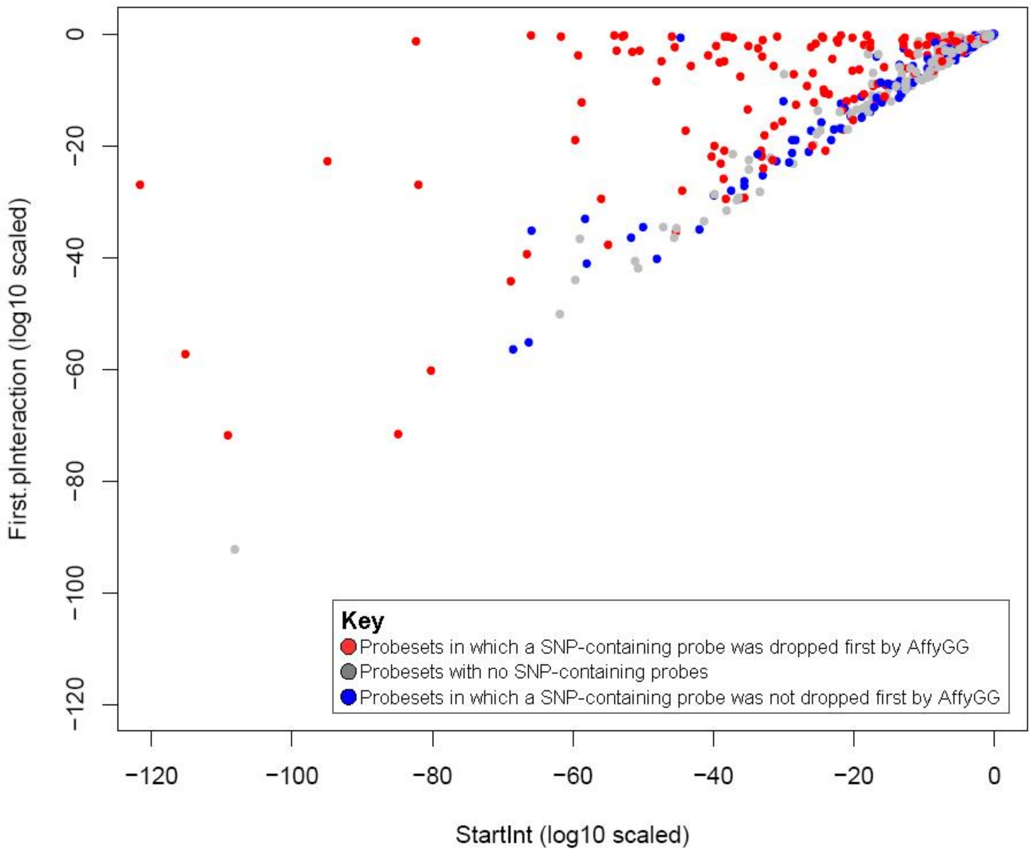


Figure 3.6: *Cis*-eQTL probesets in left ventricle by AffyGG ‘startint’:‘first pinteraction’ ratio

All *cis*-eQTLs in left ventricle are plotted according to AffyGG output; specifically the ratio between *startint* (*x*-axis, the interaction *p*-value prior to elimination of the first probe) and *first pinteraction* (*y*-axis, the interaction *p*-value following the elimination of the first probe).

It can clearly be seen in this figure that the probesets shown in red (those in which a probe with a polymorphism in its probe binding region was dropped first by AffyGG) are substantially different in terms of this ratio to the other two sets. As such, the probesets with polymorphisms that were not identified by AffyGG (perhaps because the probeset is actually a true (i.e. differentially expressed) *cis*-eQTL) were combined with those containing no SNPs (and therefore assumed to be a true *cis*-eQTL) in the next stage of this exploration of AffyGG's predictive model (**Figures 3.7-8**).

This ratio between 'startint' and 'first pinteraction' was systematically related to findings from comparative sequence analysis in the SHR and BN, by means of a Support Vector Machine (SVM) classification method (Cortes and Vapnik, 1995), implemented in R as described in **2.3.4**. The outcome of the initial SVM analysis is shown in **Figure 3.7**, below.

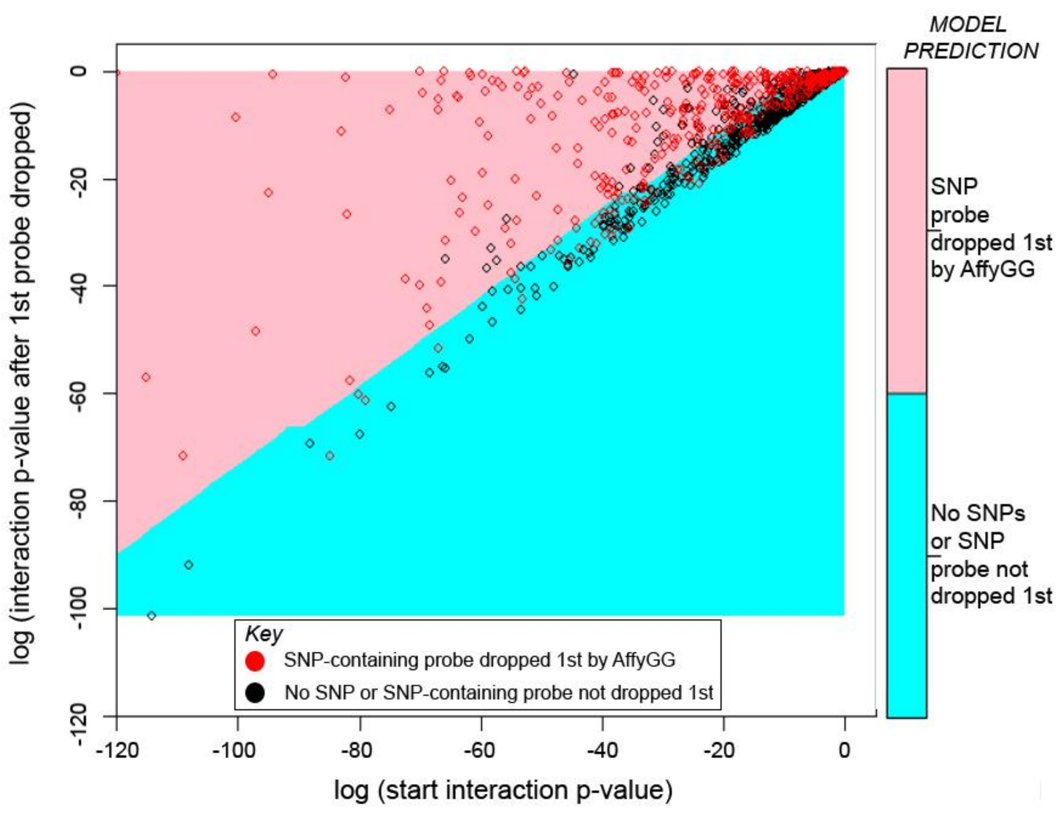


Figure 3.7: Classification of *cis*-eQTL probesets according to AffyGG prediction of probes with polymorphisms in probe binding regions

The accuracy of the SVM's discrimination between those probesets for which AffyGG dropped a SNP-containing probe first and those for which it did not (including those probesets with no SNPs) was assessed by 10-fold cross-validation. The overall accuracy was found to be 86%. The Type I error rate of the SVM was very low – only 35 of 1293 probesets predicted not to have a SNP containing probe dropped first were misclassified. However, the rate of correct identification of probesets in which a SNP probe was dropped first was somewhat poorer – 197 of 401 (49%) were correctly classified by the SVM.

It was noted that as the startint or first pinteraction *p*-values approach 1 (log $p \sim 0$), classification becomes highly challenging. In **Figure 3.8 *a)* and *b)***, probesets for which log 'startint' < -10 and log 'startint' < -20 respectively were removed from the analysis. Following this, the rate of correct prediction of 'SNP probe dropped first' probesets increased to 67% (of 293) and 76% (of 212) respectively.

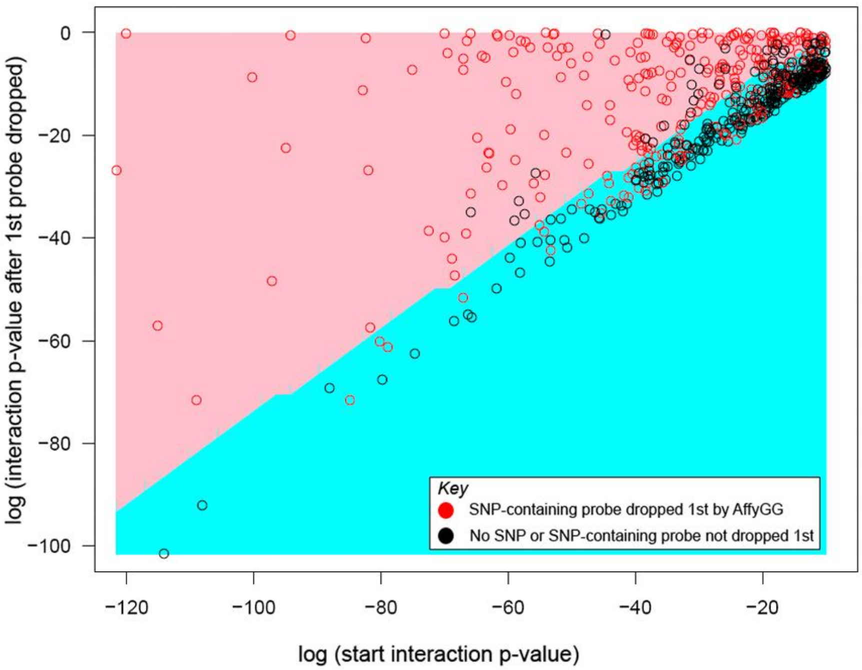


Figure 3.8 a): SVM classification of *cis*-eQTL probesets according to AffyGG prediction of probes with polymorphisms in probe binding regions, excluding those with log 'startint' < 10.

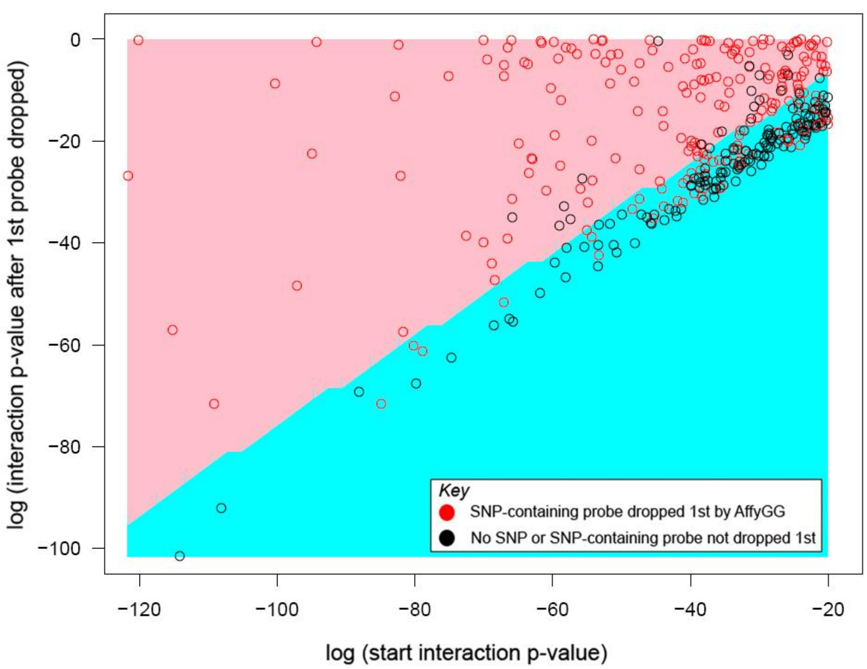


Figure 3.8 b): SVM classification of *cis*-eQTL probesets according to AffyGG prediction of probes with polymorphisms in probe binding regions, excluding those with log 'startint' < 20.

Details of the SVM's classification of probesets by ratio of 'startint' to 'first pinteraction' are provided in **Table 3.7**. As can be seen from the table; as a general rule, the higher this ratio the more likely the probeset is to have a SNP-containing probe. These probes are identified by AffyGG in 87.7% of cases where the ratio is greater than 3.

Ratio (Greater than)	Cumulative Total Probesets	No. (%) predicted to have a SNP-containing probe dropped first by AffyGG	No. (%) of overall correct SVM predictions
11.369	68	68 (100%)	67 (98.5%)
4	136	128 (94.1%)	117 (86%)
3	179	157 (87.7%)	147 (82.1%)
2	260	195 (75%)	199 (76.5%)
1.5	474	229 (48.3%)	368 (77.6%)
1.124	789	231 (29.3%)	637 (80.7%)

Table 3.7: SVM classification of 789 probesets by ratio of 'startint' to 'first pinteraction'. Probesets for which the sum of 'startint' and 'first pinteraction' is less than 10 were removed.

3.3.3 Discussion

The combining of newly-available SHR sequence data, Affymetrix probe sequence information and the EnsEMBL BN reference sequence enabled a comprehensive assessment of the prevalence of SNPs in probe binding regions and their effect on *cis* linkage to eQTLs to be made. This could then be related to the outputs of the AffyGG ANOVA model, enabling a more objective interpretation of the statistics produced by the software than has previously been possible to be made.

It was found by assessing the thresholding method discussed in (Alberts et al., 2007b), which considers only the ‘startint’ *p*-value, against the sequence findings that the rate of identification of SNP containing probes of the suggested threshold was low. It was observed following the production of a scatterplot like that shown in **Figure 3.6** that a ratio of ‘startint’ to ‘first pinteraction’ is a somewhat better predictor of SNP-containing probes identified by AffyGG, and this was explored further through the application of the SVM.

It is important to note that the methodology used in this exploration of the interpretation of the AffyGG model’s output evaluates AffyGG’s rate of identification of SNP-containing probes, not the raw numbers of probesets with polymorphisms in their probe binding regions. This is because a probeset with a SNP in one or more of its probes’ binding regions is not necessarily a ‘false’ *cis*-eQTL – it is anticipated that many truly differentially expressed probesets will also have polymorphisms in one or more probe binding regions. Therefore, the observation that the AffyGG model has not eliminated a SNP-containing probe first does not indicate a Type II error on the part of AffyGG. (That is, SNP-containing probes are not always influential as regards the QTL effect). This complicates the interpretation of the AffyGG output, since it means that sequence cannot be used to ‘judge the accuracy’ of predictions, only to contribute to improving our understanding of those predictions.

Another factor that complicates the interpretation of the output is that the numbers of SNPs in a probeset vary. It is relatively straightforward to view the change in QTL effect upon dropping a probe (the change in ratio of ‘startint’ to ‘first pinteraction’) as an indicator of the likelihood of an eQTL being due to polymorphism rather than differential expression. However, some *cis*-eQTLs were found to have as many as 7 probes with SNPs in their binding regions, making it very difficult to determine whether or not they are potentially ‘false’ by analysing this ratio. Unfortunately, it is impossible to know (unless full sequence data is available) how many polymorphisms there are across a probeset, so the interpretation of the AffyGG output cannot be so informed.

The use of SHR sequence data to inform statistical analysis of the predictions of AffyGG has the potential to make this software much more useful. However, new angles on this analysis may be required to tackle the problem of uninformative *cis*-eQTLs.

Chapter 4: Genome-wide Co-expression Analysis Incorporating Investigation of the Effect of Genotypic Variation

4.1 Introduction

Each eQTL in the dataset is described as either *cis* or *trans*, as defined in **1.8**. These classes are anticipated to represent biological differences in the modes of regulation represented by the two types of eQTLs. Specifically, *cis*-eQTLs are defined as they are (**1.8.1**) so as to identify local gene effects on expression across the strains; conversely, *trans*-eQTLs are expected to identify remote gene effects, i.e. genes whose expression is affected by a polymorphism at a different locus.

The only forms of data about regulation actually contained within the eQTL dataset are *a)* the microarray-measured expression levels of the probesets in each of the RI strains and *b)* the genotype information contained within the genetic markers' SDPs. These can be combined in computational analysis with other information to enable informed biological hypotheses about the regulation of gene expression to be formulated.

This eQTL dataset and others can and have been studied with a view to isolating the genes underlying specific eQTLs or groups of eQTLs (Hubner et al., 2005), an approach that has facilitated gene discovery in several model organisms (e.g. (Meng et al., 2007)) and in humans (Goring et al., 2007). However, while it has been possible to identify and characterise the genes underlying *cis*-eQTLs, the study of *trans*-regulation using eQTL data has proven more challenging, undoubtedly at least in part due to the relatively smaller effect size of *trans*-eQTLs (Petretto et al., 2006a).

Genome-wide, or 'global' investigations into the eQTL dataset have the potential to improve our understanding of the structure of the data, and therefore the quality of the predictions made using it. Several genome-wide analyses of eQTL data have been carried out, however the data is generally not fully explored before progression to the application of functional enrichment, pathway prediction and/or network construction algorithms. In some studies, such as (Cowley et al., 2008; Yvert et al., 2003), co-expression analysis is carried out in a set of transcripts without making use of the *cis*-*trans* delineation.

The application of network construction methods to *cis*-eQTLs has not met with significant success in identifying functional sets (Lum et al., 2006). One possible explanation for this, noted by (Doss et al., 2005), is the confounding effect of linkage disequilibrium. The ambiguity as to whether the correlation of closely linked transcripts is due to pleiotropy (one gene affecting the expression of both of the correlated transcripts) or linkage makes the inference of biological significance of such relationships difficult (Chen et al., 2007).

Another possible factor affecting the interpretation of *cis*-eQTL co-expression analysis is long-range allelic association among un-linked markers. In (Cervino et al., 2005) it was suggested that this allelic association may lead to false linkages in QTL mapping studies, and it is possible that this may apply also to co-expression analysis.

The systematic co-expression analysis of the sets of *cis*-eQTLs and *trans*-eQTLs in each of the four tissues, combined with the testing for correlation of the SDPs of their associated markers, was intended to enable the structures of all of these datasets to be compared and contrasted, including class of eQTL and tissue.

It has been shown that genes show highly tissue-specific patterns of regulation (Yang et al., 2006) and eQTL detection rates show tissue-specificity across the genome (Petretto et al., 2006a). The latter analysis is extended here through the investigation of the tissue-specificity of the statistical significance of co-expression relationships between eQTLs in the same dataset. In particular, the *trans*-eQTL clusters in the study dataset are highly tissue-specific – a large cluster identified in one tissue is completely absent in the next (**Table 4.4**, ***Appendix III***).

The 81 (**Table 4.4**) *trans*-eQTL clusters, defined as a group of 10 or more expression traits linked to the same genetic locus in *trans*, were identified purely through linkage analysis. They consist of transcripts with highly correlated expression levels, however the significance of this correlation in informing about biological relationships between the transcripts is not clear from genome-wide co-expression analysis alone, because of the confounding effect of the genotype at the common peak of linkage. This can be accounted for by the use of partial correlation, conditioned on the genotype (**4.2.3**).

Trans-eQTL clusters have been hypothesised (Yvert et al., 2003) to represent co-ordinated regulation of the transcripts comprising the cluster (**1.8.2.1**), and previous attempts to identify the 'master regulator' (Morley et al., 2004) explaining the linkage of the cluster genes to the same locus have utilised co-expression methods but with little success (Keurentjes et al., 2007; Tu et al., 2006). The possibility of functional enrichment within *trans*-eQTL clusters has been explored in other datasets (e.g. (Lan et al., 2006; Wu et al., 2008)), with some studies having outcomes suggesting that such clusters are predictive of biological relationships.

In (Fuller et al., 2007), coexpression network construction is performed prior to categorisation to identify 'modules' (**1.9.5**), hypothesised to consist of highly interconnected genes. This approach to 'module' identification is used by (Lan et al., 2006) to identify highly correlated gene sets from a set of 5,293 *trans*-mapping mouse transcripts. These gene sets, like the *trans*-eQTL clusters, map to the same genomic location. Such 'modules' were also observed by (Emilsson et al., 2008) following analysis of a large human expression dataset, however these did not map to the genome by linkage-analysis. It is disputed whether the *trans*-eQTL clusters identified by linkage analysis and described in e.g. (Hubner et al., 2005) represent truly biologically meaningful groupings of transcripts (Breitling et al., 2008). It has been suggested that these clusters may instead indicate genes with highly correlated expression (Pérez-Enciso et al., 2003), perhaps due to inter-sample technical variation (Kang et al., 2008), that link to the genome by chance because of the relatively small number of strains in the panel (**1.8.3**).

The genetics and gene expression of eQTLs forming *trans*-eQTL clusters were studied in detail, and the previously hypothesised relationship between correlation of co-localising *cis*-eQTLs and the cluster *trans*-eQTLs was explored. Additionally, the hypothesis that the correlation in expression levels observed in *trans*-eQTL clusters ceases to be significant after the effect of genotype has been taken into account, and the prediction of functional enrichment of *trans*-eQTL clusters, were assessed in all large (≥30 *trans*-eQTLs) clusters in the BXH/HXB dataset through the application of partial correlation and Gene Ontology (GO) analysis respectively.

4.2 Results

4.2.1 Co-expression analysis of *cis*- and *trans*-eQTLs

The methods used to carry out co-expression analysis and assess statistical significance are described in **2.4.2** and **2.5.3**. The methods used to test for correlation of marker genotypes are described in **2.4.3** and **2.5.4**. Methods used in the assessment of the statistical power of co-expression analysis are described in **2.5.5**.

Figure 4.1 and **Table 4.1** display the results of the co-expression analysis of *cis*-eQTLs. They show that over 80% of the pairs of *cis*-eQTLs found to have significantly correlated expression profiles could potentially be explained by either genetic linkage (i.e. the eQTLs are less than 50cM apart) or long-range allelic association. In either case, testing for correlation of their peak marker's SDPs identifies a statistically significant effect.

In **Figure 4.1**, all significantly correlated pairs of *cis*-eQTLs in each tissue are shown, plotted by their positions on the genetic map.

In **Table 4.1**, the percentage of all pairs of *cis*-eQTLs in each of the datasets found to be significantly correlated is given. This ranged from 0.6% to 0.9% of pairs tested. Of these, 67.9-81.6% could potentially be explained by linkage disequilibrium. These findings indicate that significant correlation of unlinked *cis*-eQTLs is very rare.

In order to give context to these findings, **Figure 4.2** shows the pattern of allelic association for all markers with no missing genotype data (n = 671) markers.

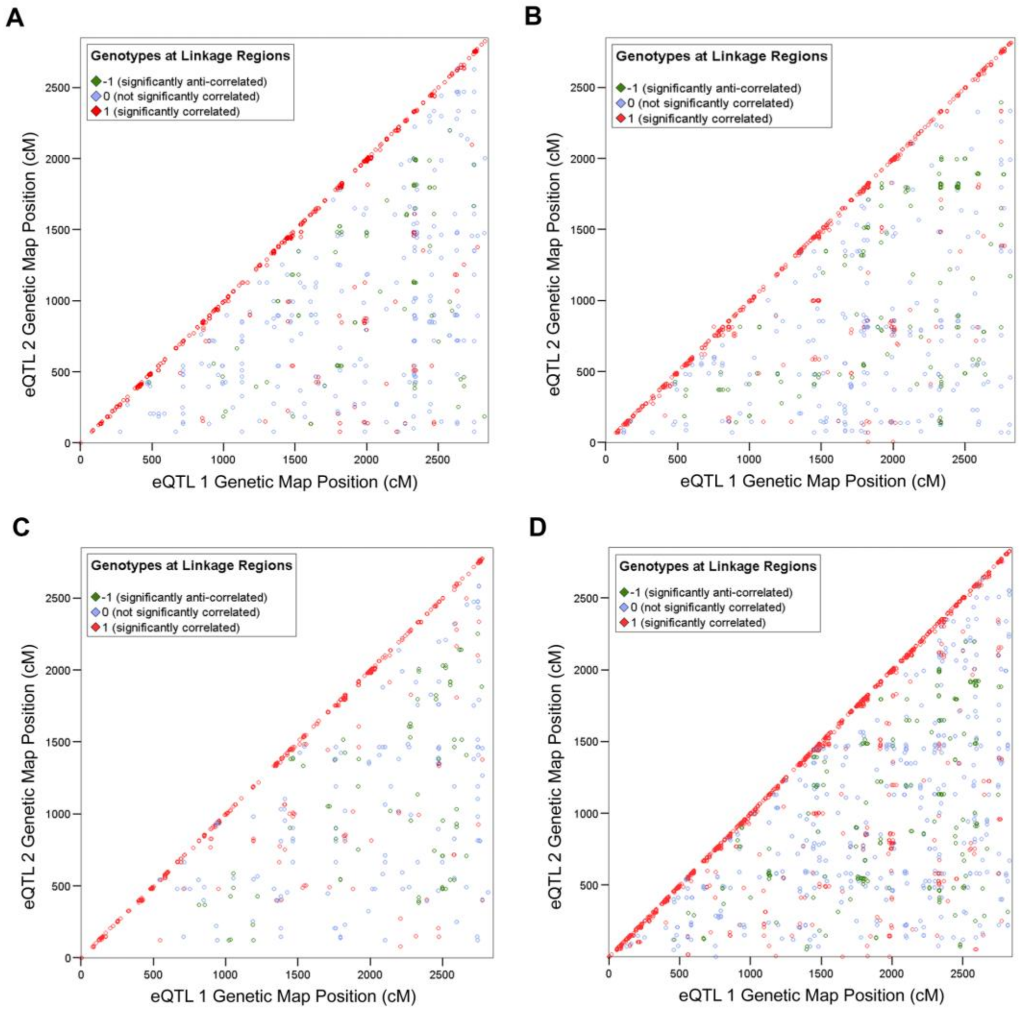


Figure 4.1: Scatterplots showing pairs of *cis*-eQTLs with significantly correlated gene expression by genetic map location in *A)* Fat, *B)* Kidney, *C)* Adrenal and *D)* Left Ventricle tissues.

Those with significantly correlated SDPs are plotted in red; those with significantly anti-correlated SDPs in green; and those whose SDPs were not found to be correlated are shown in blue.

Tissue	Total Number of Pairs	No. Significant Correlated Pairs (*q = 0.05*)	% Significant Correlated Pairs	No. (%) Significantly Correlated Pairs located less than 50cM apart	No. (%) Significantly Correlated Pairs with Correlated Genotypes at Peaks of Linkage
Fat	155,403	1,386	0.9	941 (67.9%)	1,114 (80.4%)
Kidney	257,403	2,284	0.9	1,654 (72.4%)	1,979 (86.6%)
Adrenal	180,901	1,443	0.8	1,123 (77.8%)	1,294 (89.7%)
LV	926,841	5,371	0.6	4,382 (81.6%)	4,912 (91.5%)

Table 4.1: Summary of *cis*-eQTL co-expression statistics in each of the four tissues.

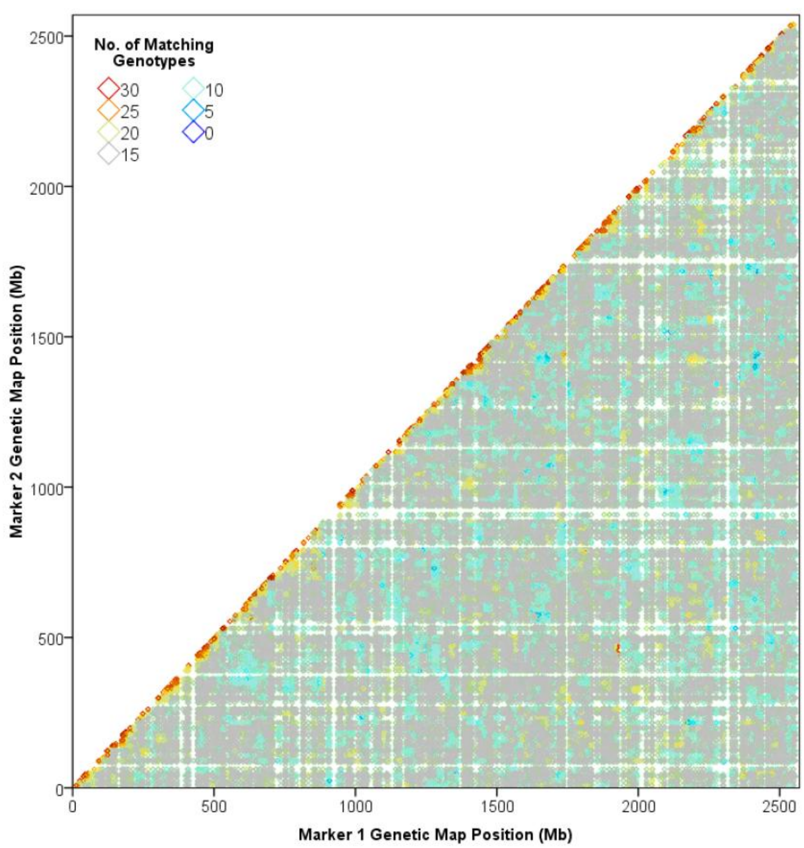


Figure 4.2: Scatterplot showing numbers of matches in SDP of all pairs of markers (excluding those with missing genotype data), genome-wide.

The data shown in **Figure 4.1** and **Table 4.1** can be considered in view of the genome-wide view of genotypic correlation of marker SDPs given in **Figure 4.2**, and contrasted with the outcome of the correlation of *trans*-acting eQTLs. This was performed using the same methods as was the *cis*-eQTL co-expression analysis.

Because *trans*-eQTL probesets, by definition, do not map to the same genetic location as their peaks of linkage, two sets of genotype correlation statistics were produced for each significantly correlated pair of *trans*-eQTLs. These provide information on the genotypes of the markers at the eQTLs' peaks of linkage, and of the nearest marker to the genetic map position of the probeset.

Figure 4.3 shows all significantly correlated pairs of *trans*-eQTLs, plotted by the genetic map position of the probesets. **Figure 4.4** shows the same correlated pairs, but plotted by the genetic map position of the peaks of linkage rather than the probesets. A summary of the findings of the genome-wide correlation of the *trans*-eQTL dataset in each of the four tissues is given in **Table 4.2**.

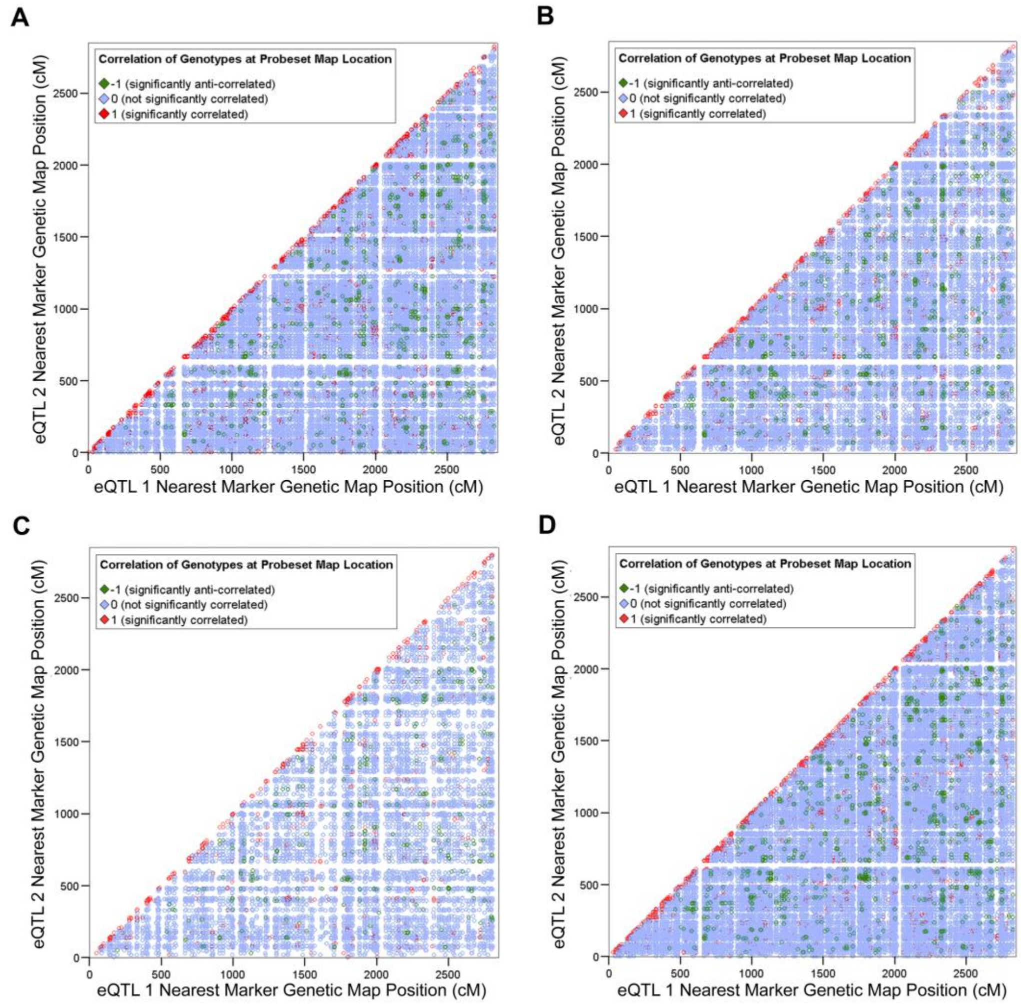


Figure 4.3: Scatterplots showing pairs of *trans*-eQTLs with significantly correlated gene expression by genetic map location of transcript in *A)* Fat, *B)* Kidney, *C)* Adrenal and *D)* Left Ventricle tissues.

Colour-coding as in Figure 4.1

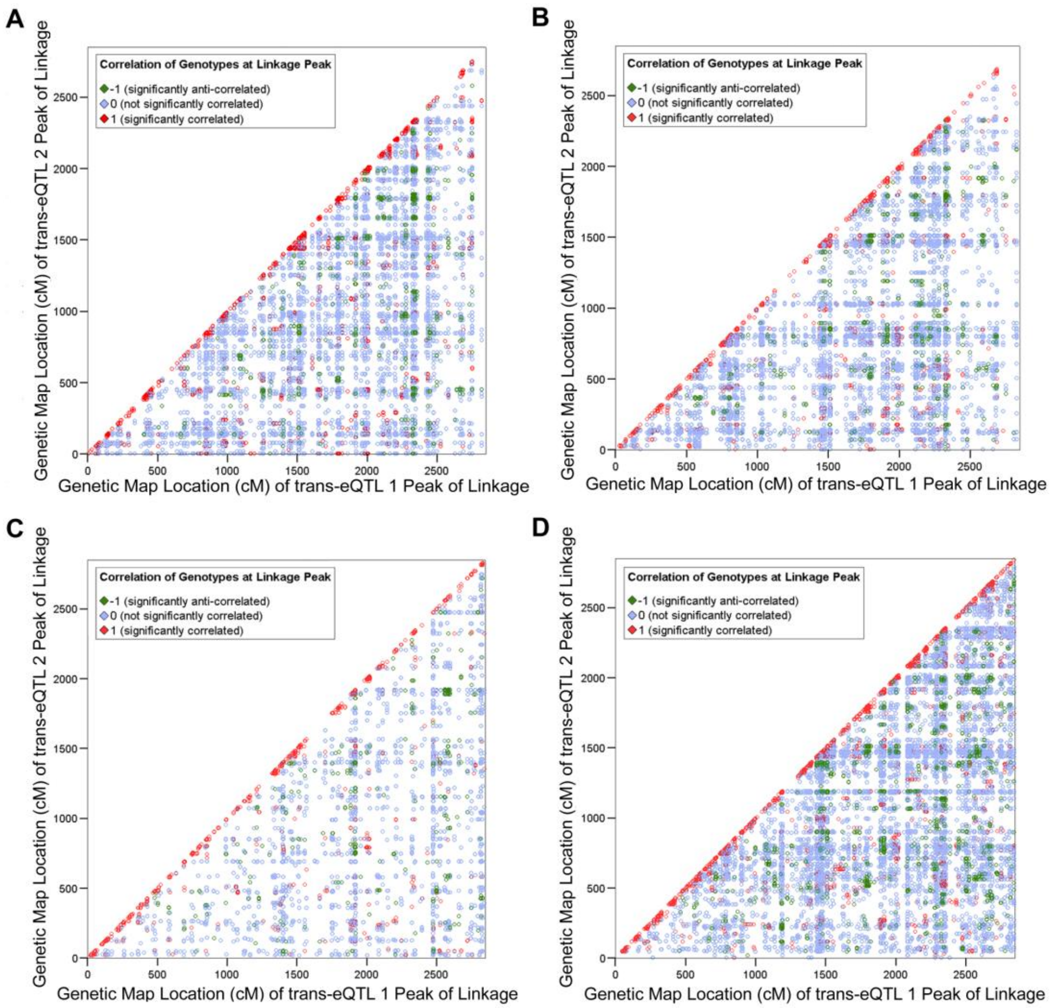


Figure 4.4: Scatterplots showing pairs of *trans*-eQTLs with significantly correlated gene expression by genetic map location of peak of linkage in *A)* Fat, *B)* Kidney, *C)* Adrenal and *D)* Left Ventricle tissues.

Colour-coding as in Figure 4.1

Tissue	Total Number of Pairs	No. Significant Correlated Pairs (q = 0.05)	% Significant Correlated Pairs	No. (%) Significantly Correlated Pairs with Significantly Correlated Genotypes at Peaks of Linkage	No. (%) Significantly Correlated Pairs with Significantly Correlated Genotypes at Probeset Locations
Fat	425,503	63,477	14.9	31,288 (49.3%)	6,375 (10.0%)
Kidney	533,028	39,257	7.4	18,628 (47.5%)	4,601 (11.7%)
Adrenal	434,778	12,482	2.9	7,945 (63.7%)	1,370 (11.0%)
LV	2,288,730	78,826	3.4	44,110 (56.0%)	8,450 (10.7%)

Table 4.2: Summary of *trans*-eQTL co-expression statistics in each of the four tissues.

It is clear that there are many more significantly correlated pairs of *trans*-eQTLs than there are pairs of *cis*-eQTLs. One explanation for this finding is the fact that there are somewhat more *trans*-eQTLs in the eQTL dataset than there are *cis*-eQTLs (see **Table 3.1**).

However, as well as there being a greater number of significantly correlated pairs of *trans*-eQTLs, a much greater proportion (2.9 – 14.9%) of possible pairs of *trans*-eQTLs were found to be significantly correlated, relative to *cis*-eQTLs (0.6 – 0.9%). Unlike *trans*-eQTLs, the search-space for *cis*-eQTL transcripts is limited to the defined window region (**1.8.1**), meaning that many more probesets can form a *trans*-eQTL in a given region of the genome than can form a *cis*-eQTL. Additionally, the fact that *trans*-eQTL genes can be located anywhere in the genome and be linked to the same region means that highly correlated groups of *trans*-eQTLs (*trans*-eQTL clusters) can occur, as was observed here. This explains the contrasting patterns of correlation observed in **Figures 4.3 and 4.4**.

Across the four tissues, 47.5-63.7% of significantly correlated pairs of *trans*-eQTLs were found by the matching coefficient analysis of SDPs to have significantly correlated genotypes at the peak of linkage. Less than 12% of the transcripts, however, were co-located. This is indicative of correlation of transcript expression profiles located in different regions of the genome, perhaps suggesting that – for these transcripts - gene expression (as quantitified by probeset hybridization) is influenced to a much greater degree by the segregation of a region of the genome that is at least 20Mb away from the gene and is often on a different chromosome than by the segregation of the gene itself. Alternatively, it could be indicative – as suggested by (Breitling et al., 2008) – of chance linkage of the co-expressed *trans*-eQTLs. If one of a group of highly correlated *trans*-eQTL genes is linked to a given region, it is very likely that the others will be too, giving rise to a *trans*-eQTL cluster.

The power and precision of the Pearson correlation analysis itself, given the sample size (n = 29), was assessed by finding the statistical power (estimated probability of rejecting the null hypothesis of no significant correlation given that it is false) and confidence intervals (range in which the true correlation coefficient will occur with 95% probability) as described in **2.5.5**. In each tissue, a Pearson correlation coefficient for this analysis was chosen to be approximately equivalent (due to the use of empirical p-values, there is not an exact, consistent correspondence) to the threshold q = 0.05. The confidence interval and statistical power for each co-expression analysis are given in **Table 4.3**.

Tissue	Approx. Threshold Pearson	95% C.I. Lower Bound	95% C.I. Upper Bound	Statistical Power
***Cis*-eQTLs**				
Fat	0.61	0.314	0.798	96%
Kidney	0.61	0.314	0.798	96%
Adrenal	0.62	0.328	0.804	96%
Left Ventricle	0.63	0.343	0.810	97%
***Trans*-eQTLs**				
Fat	0.47	0.125	0.714	75%
Kidney	0.58	0.271	0.781	93%
Adrenal	0.55	0.230	0.763	89%
Left Ventricle	0.54	0.216	0.757	88%

Table 4.3: 95% confidence intervals and statistical power for the identification of significantly correlated pairs of *cis*- and *trans*-eQTLs

Significant Pearson correlation coefficient defined as coefficient approximately equivalent to threshold $q = 0.05$

The results given in **Table 4.3** show that the statistical power and precision of co-expression analysis is, like that of the eQTL mapping itself (**1.8.3**), both greater and more variable across the tissues in *cis*-eQTLs than in *trans*-eQTLs. This is a consequence of a combination of the small size of the RI panel used for the analysis, and the relatively low threshold correlation coefficients defined by the *q*-value procedure as significant in *trans*-eQTLs.

4.2.2 Co-expression analysis within *trans*-eQTL clusters

The pairwise correlation structure within *trans*-eQTL clusters was analysed, and the results are shown in **Figure 4.5** and **Table 4.4**. As expected, there was found to be a much greater degree of co-expression within the clusters than in the set of *trans*-eQTLs as a whole.

Figure 4.5 shows the outcome of this analysis, indicating the percentage of significantly correlated pairs (as defined in **2.6**) in *trans*-eQTL clusters, grouped by tissue. The boxplots show the mean, interquartile range and range of this percentage in each of the four tissues. The horizontal lines give the percentage of significant pairwise correlation of expression profiles across the tissue as a whole. **Table 4.4** summarises the underlying data.

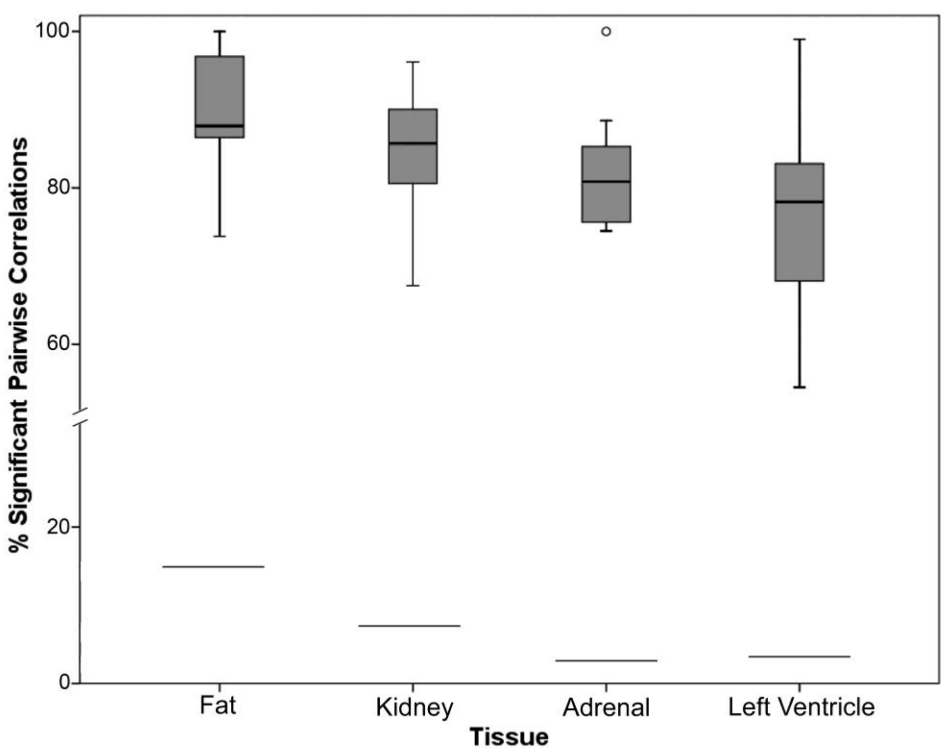


Figure 4.5: Boxplots showing percentage within-cluster significant correlation of genes underlying *trans*-eQTLs forming *trans*-eQTL clusters relative to percentage significant correlation across tissue (indicated by horizontal lines).

Tissue	% Trans-eQTL Significant Correlation (overall)	No. Trans-eQTL Clusters (≥10 co-linked eQTLs)	% Significant Correlation Within Clusters (average)
Fat	14.9	11	90.2
Kidney	7.4	23	84.2
Adrenal	2.9	9	82.4
LV	3.4	38	77.2

Table 4.4: Significant co-expression of *trans*-eQTLs within *trans*-eQTL clusters and globally, by tissue

Much of the statistically significant co-expression observed in the genome-wide analysis of *trans*-eQTLs is due to the presence in the dataset of these highly correlated groups of cluster-forming transcripts. The variability in levels of statistically significant correlation of *trans*-eQTLs between tissues can also largely be explained by the presence in the data of these groups. In fat, a large *trans*-eQTL cluster contains 146 *trans*-eQTLs with 96.8% significant pairwise correlation (see ***Appendix III***), for a total of 10,250 significantly correlated pairs. This means that just under a third (32.7%) of the total number of significantly correlated *trans*-eQTLs in fat can be explained by the presence of this one *trans*-eQTL cluster.

Kidney and adrenal do not contain such large *trans*-eQTL clusters (the largest cluster across these two tissues occurs in kidney and consists of 57 *trans*-eQTLs). The number of clusters observed in left ventricle is much higher than in any of the other three tissues – this is not surprising since the expression of almost twice as many transcripts was profiled and double the number of *trans*-eQTLs were detected (**Table 3.1**)

As discussed in **1.8.3**, an analysis of the statistical power to detect *trans*-eQTLs in each of the four tissues was carried out by others working on the dataset and published (Petretto et al., 2006a). The power to detect *trans*-eQTLs was found to be substantially higher in fat and left ventricle than in kidney and adrenal. One possible reason for this relates to the greater heterogeneity of cell types in the kidney and adrenal tissues relative to the other two; as many eQTLs have been found to be cell-type specific (Gerrits et al., 2009), the signals in these tissues may be noisier.

Because the effect size of *trans*-eQTLs is generally small (Hubner et al., 2005), relative to *cis*-eQTLs they are difficult to detect and the false positive rate is much higher. The *trans*-eQTLs in the 81 clusters listed in ***Appendix III*** were identified at genome-wide $p < 0.05$, at which the median expected FDR ranges from 26-35% across the tissues. This means that genome-wide, around one third of all *trans*-eQTLs are false positives, particularly as most (>75% across the four tissues) *trans*-eQTLs are not significant at $p < 0.01$. In **Table 4.5**, the numbers of *trans*-eQTL clusters in each tissue at more stringent levels of genome-wide significance are provided.

Tissue	**No. *trans*-eQTL clusters at:**		
	***p* < 0.05**	***p* < 0.01**	***p* < 0.001**
Fat	11	1	0
Kidney	23	1	0
Adrenal	9	1	0
Left Ventricle	38	3	0

Table 4.5: Numbers of *trans*-eQTL clusters at different genome-wide significance thresholds for mapping of *trans*-eQTLs.

Trans-eQTL cluster is defined as a group of 10 or more *trans*-eQTLs that map to the same region of the genome

Of 81 *trans*-eQTL clusters, six remain in the dataset when only the set of *trans*-eQTLs that are significant at $p < 0.01$ is considered. These are bolded in ***Appendix III***, which also lists the numbers of *trans*-eQTLs in all clusters found at $p < 0.05$ remaining in the dataset at the more stringent genome-wide significance threshold.

4.2.3 Partial correlation analysis of large *trans*-eQTL clusters

The *trans*-eQTL clusters, by definition, consist of groups of genes linked to the same region of the genome. As such, it is logically implied that there will be some correlation of the transcripts' expression profiles across the 29 RI strains, as these transcripts would not otherwise associate with the SDP of the marker at a common peak of linkage. Partial correlation analysis of the expression profiles, conditioned on the genotype of the marker, is a means of assessing the degree to which the effect of genotype gives rise to the observed correlation.

Partial correlation analysis was carried out on the fifteen clusters, across the four tissues, consisting of 30 or more *trans*-eQTLs (**Table 4.6**). Significance of the partial correlations was assessed using a two-tailed *t*-test. A *t*-statistic was found for each pair as described in **2.4.5** and significance at *n*-3 degrees of freedom (where *n* is the number of genotypes in the SDP of the marker at the cluster peak of linkage, excluding missing genotypes) was found. For each cluster, the percentage of pairs of expression traits whose partial correlation coefficients were found to be significant by this method is given.

Tissue	Marker at Cluster Peak of Linkage	No. *trans*-eQTLs in cluster	No. pairs with significant partial correlation	% pairs with significant partial correlation	% significant Pearson correlation
LV	D15Rat29	54	2664	93.1	99.0
Fat	D17Rat1	146	17700	83.6	96.8
LV	D15Ucsf1	46	1508	72.9	92.9
Fat	Cacna1s	33	646	61.2	96.8
LV	D15Rat98	30	488	56.1	88.0
LV	D15Utr2	51	1416	55.5	85.5
Kidney	D15Rat69	57	1554	48.7	91.2
LV	Cyp45c	35	564	47.4	87.9
Fat	D4Rat240	31	372	40.0	84.7
Kidney	Igk@	49	924	39.3	89.6
Adrenal	D11Rat16	31	312	33.5	75.5
Adrenal	D17Rat144	47	656	30.3	88.6
LV	Crabp1	165	7870	29.1	77.1
LV	D8Mit12	77	1492	25.5	68.8
LV	Ckb	43	328	18.2	79.2

Table 4.6: Outcome of partial correlation analysis of the *trans*-eQTL genes forming clusters consisting of 30 or more eQTLs.

Nominal significance of partial correlations defined by applying a two-tailed *t*-test to the coefficient with *n*-3 degrees of freedom. The clusters are ordered by % significance of the partial correlations.

The *trans*-eQTL clusters were found to vary substantially in levels of significant partial correlation. As expected, there is a positive relationship between the proportion of pairs of transcripts found to be significantly correlated by uninformed Pearson correlation, and those found to be significantly correlated once the potential confounding effect of genotype is taken into account through the application of partial correlation. These percentages are however not statistically equivalent, as the threshold used to assess significance of the partial correlation coefficients is not corrected for multiple testing across the genome-wide set of *trans*-eQTLs.

4.2.4 Functional enrichment analysis of large *trans*-eQTL clusters

One hypothesis for the explanation of the presence of these cluster-forming groups of co-expressed genes in eQTL datasets such as that described here is that they represent co-ordinated regulation of multiple genes perhaps by a ‘master regulator’ close to the peak of linkage of the *trans*-eQTL cluster (Yvert et al., 2003). It was considered that these genes may well share functional characteristics that would explain the occurrence of such co-regulation in a wider biological context. This hypothesis was tested by assessing the levels of functional enrichment within the sets of co-expressed transcripts by comparison of the gene ontology (GO) annotations of the probesets using DAVID functional classification tool (Huang et al., 2007). The biological process and molecular function ontologies were used. As with the partial correlation analysis, only those clusters consisting of 30 or more *trans*-eQTLs were used in the analysis. The outcome of this analysis is shown in **Table 4.7**. For each GO term found to be significant at *p*<0.01 by the DAVID functional enrichment tool, the term and the number of genes annotated for that term are provided in ***Appendix IV***.

Tissue	Marker at Cluster Peak of Linkage	No. Cluster *trans*-eQTLs	Significant GO terms at *p*<0.001	Significant GO terms at *p*<0.01	Significant GO terms at *p*<0.05
LV	D15Rat98	30	0	0	1
Adrenal	D11Rat16	31	0	2	7
Fat	D4Rat240	31	2	4	8
Fat	Cacna1s	33	0	1	4
LV	Cyp45c	35	0	0	0
LV	Ckb	43	3	4	12
LV	D15Ucsf1	46	0	2	12
Adrenal	D17Rat144	47	0	1	8
Kidney	Igk@	49	1	3	14
LV	D15Utr2	51	2	13	20
LV	D15Rat29	54	8	15	28
Kidney	D15Rat69	57	8	12	19
LV	D8Mit12	77	0	0	3
Fat	D17Rat1	146	3	20	52
LV	Crabp1	165	0	4	14

Table 4.7: Outcome of GO analysis of *trans*-eQTL clusters consisting of 30 or more eQTLs using DAVID, a functional annotation software tool.

Twelve (80%) of the large *trans*-eQTL clusters were found to be significantly enriched for at least one GO term at *p*<0.01, and 7 at *p*<0.001.

To make it possible to assess the significance of these findings in the context of the *trans*-eQTL dataset as a whole, simulation studies were performed. As the median size of the 15 large *trans*-eQTL clusters is 47 eQTLs, 200 random sets of 47 *trans*-

eQTL genes were selected and tested for functional enrichment using DAVID in the same way as were the large *trans*-eQTL clusters.

At all *p*-value thresholds, fewer significant GO terms were found in random groups than in the *trans*-eQTL clusters. Kolmogorov-Smirnov (K-S) tests (**2.6.3**) were used to assess the statistical significance of these tests. At all three DAVID *p*-value thresholds (p<0.001, p<0.01, p<0.05) tested, the degree of functional enrichment found in the *trans*-eQTL clusters was found to be significantly greater than in the random *trans*-eQTL sets (K-S Exact test *p*-values: p<0.001, p<0.002, p<0.004).

4.2.5 Investigation of the Genetic Basis of Co-regulation in *trans*-eQTL Clusters

The outcomes of the co-expression and functional investigations of the *trans*-eQTL clusters described in the previous sections suggest that at least some of these groups of transcripts may represent biologically significant sets of genes within the tissue in which they were identified. As the defining feature of a *trans*-eQTL cluster is that all of the eQTLs therein are linked to the same region of the genome, this region is intuitively the first place to look in the search for the genetic basis of this co-regulation.

The region of the genome 50Mb either side of the peak of linkage of each cluster was designated the cluster's 'window region'. A wide region was deliberately chosen in order that all possible 'master regulators' (**1.8.2.1**) were taken into consideration, taking into account the possibility of mapping error and the effects of linkage disequilibrium. The expression profiles of the *cis*-eQTLs in this ‘window region’ were then tested for co-expression with those of the *trans*-eQTLs in the cluster. It was anticipated that this would enable the prioritisation of *cis*-eQTLs as master regulator candidates.

This analysis was carried out, and *cis*-eQTLs with expression profiles highly correlated to those of the cluster *trans*-eQTLs were identified for several clusters. However, when the outcomes of these analyses (**Figure 4.6**) were inspected, the confounding effect of linkage disequilibrium was found to produce a strong negative relationship between the average correlation coefficient (R^2: 0.15-0.43) with the distance of the potential regulator from the peak of linkage of the *trans*-eQTL cluster.

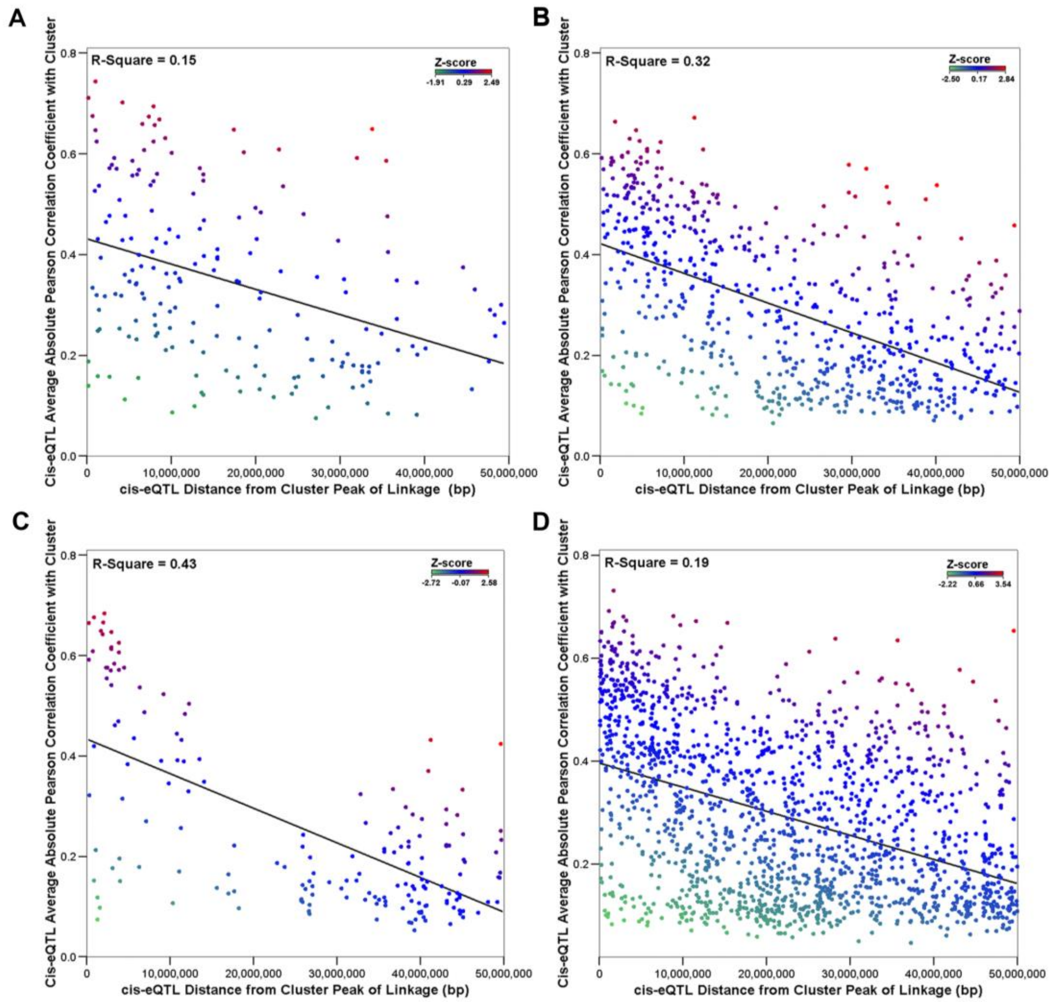


Figure 4.6: Scatterplots showing correlation of genes underlying cluster *trans*-eQTLs and *cis*-eQTLs located in the 50Mb window-region by distance of the *cis*-eQTL from the cluster peak of linkage in *A)* Fat, *B)* Kidney, *C)* Adrenal and *D)* Left Ventricle tissues.

The hypothesis that genetic linkage is the cause of this distance effect on correlation of *cis*- and *trans*-eQTLs linked to the same genetic region was tested by analysing the correlation of genes located in the region but not linked to it. **Figure 4.7** shows the outcome of this analysis.

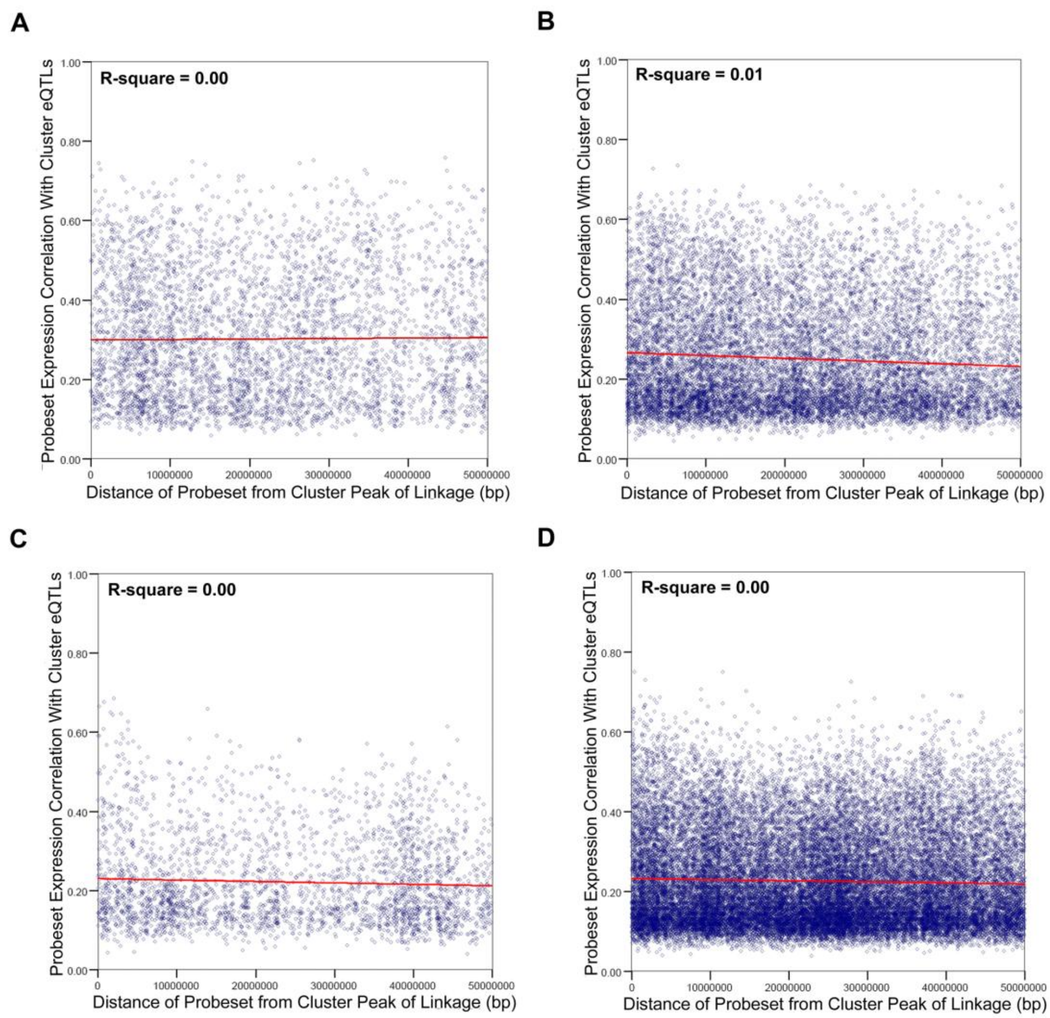


Figure 4.7: Scatter plots showing correlation of genes underlying cluster *trans*-eQTLs with all probesets (not just eQTLs) located in the 50Mb window-region by distance of the *cis*-eQTL from the cluster peak of linkage in *A)* Fat, *B)* Kidney, *C)* Adrenal and *D)* Left Ventricle tissues.

As predicted, the covariate effect of distance on correlation with the cluster disappears when unlinked genes are tested for correlation rather than *cis*-eQTLs (R^2: 0.00-0.01).

4.3 Discussion

The application of systematic genome-wide co-expression analysis, making use of the dichotomous SDPs across the RI strain panel, has enabled the correlation structure of all *cis*- and *trans*-eQTLs in a large dataset to be presented.

It is not practical to test regulatory hypotheses directly at a genome-wide level in the laboratory, but the application of *in silico* methods can provide valuable insight into underlying regulatory structure. These include gaining an improved understanding of the nature of the individual linkages observed, the relationships between expression traits, and the specific associations between the genetic inheritance at the eQTL loci and the expression signals of the linked transcripts.

The availability of four tissues from the same panel of RI strains; specifically adipose tissue, kidney, adrenal and left ventricle; enabled the systematic analysis effectively to be replicated four times. Although the same probesets (with extra probesets in one tissue (**2.1.2.1**)) and genetic markers were used in each genome-wide analysis, because eQTLs are so tissue-specific a different set of *cis*- and *trans*-eQTLs was produced and analysed in each of the four tissues. The levels of significant co-expression observed in *cis*-eQTLs was found to be very similar across the four tissues (0.6-0.9% significant correlation), but in *trans*-eQTLs it was found to differ substantially across the tissues (2.9-14.9%) (**Table 4.2**)

Correlations of expression profiles were considered in tandem with those of the SDPs at the genetic loci, in view of the prevailing understanding of the biological basis of eQTLs as outlined in (Yvert et al., 2003), specifically that *cis*-eQTLs indicate local and *trans*-eQTLs distant regulation of the linked transcripts (**1.8.1-2)**. These assumptions underlie most published analyses of eQTL data, and while they appear to make sense, there has been little study of their validity so it is worthwhile to consider initially the consistency of the conventional explanation with actual data.

This was achieved by assessing the degree of significant co-expression in each of the four tissue datasets of *cis*- and *trans*-eQTLs, using an empirical permutation-based method described in **2.5.3**. This was combined with assessment of the similarity of genetic inheritance at the peak of linkage; the method used for doing this is described in **2.5.4**. The results of these analyses were then interpreted with regard to the genetic location of the eQTL, and in *trans*-eQTLs also the location to which the transcript maps, culminating in the production of the scatterplots in (**Figures 4.1, 4.3 & 4.4**).

These scatterplots make clear the degree to which the sets of *cis*-eQTLs differ from *trans*-eQTLs in terms of overall levels of significant correlation. They enable patterns of correlation to be visualised in a way that makes clear the extent to which significant correlation of *cis*-eQTL transcript expression levels can potentially be explained by non-biological factors, specifically genetic linkage and long-range allelic association. In all four tissues, less than 1% of all pairs of *cis*-eQTLs were significantly correlated, and of these 80-90% were found to have significantly correlated SDPs at the peak of linkage (**Table 4.1**). In contrast, percentage significant correlation of *trans*-eQTLs ranged from 3% to 15% across the tissues, and 48-64% of these were found to have significantly correlated SDPs at the peak of linkage (**Table 4.2**).

The interpretation of correlation of *cis*-eQTLs is made challenging by the confounding effect of linkage disequilibrium among closely linked eQTLs (Doss et al., 2005). The potential for finding correlation of non-linked *cis*-eQTLs is minimal by definition, as the transcript displaying variation in expression must be located at the same locus it maps to. Only 459 such pairs were found in the richest tissue (left ventricle), from a possible 926,841 (**Table 4.1**).

Because of the confounding effect of linkage disequilibrium (Drake et al., 2006), the application of uninformed co-expression analysis to *cis*-eQTLs' expression profiles (e.g. in the construction of co-expression networks) is unlikely to be directly informative about biological relationships. It is possible to test for causal relationships in pairs of transcripts that are both linked to the same locus (e.g. (Chen et al., 2007)), however this was not attempted here, as the investigation of *trans*-eQTLs, in which a greater degree, and more interesting patterns, of correlation was observed took priority. A set of pairwise correlated *cis*-eQTLs is not suitable for testing for functional enrichment, because the co-expressed pairs are discrete. A simple functional analysis of these pairs was attempted using Chilibot (Chen and Sharp, 2004), which mines MEDLINE abstracts for genetic relationships, however the attempt to find significant functional relationships between co-expressed genes using this method was unsuccessful. The biological explanation of the derivation of *cis*-eQTLs suggests that the differential transcript abundance underlying the linkage is caused by a direct gene effect, perhaps a polymorphism in a promoter or 3'UTR region of the gene whose expression is measured by the probeset (Hubner et al., 2005).

Since the definition of a *trans*-eQTL does not require that the transcript be located in the same region of the genome that the eQTL maps to, hence there is a much larger search-space for significantly correlated transcripts than in *cis*-eQTLs, it was expected that significant correlation of *trans*-eQTLs would be much commoner than in *cis*-eQTLs. This was clearly shown to be the case. Since *trans*-eQTLs are thought to arise as a result of remote regulation producing differential expression - highly co-expressed *trans*-eQTLs have been predicted to represent key components in transcriptional networks (Lum et al., 2006) - it was anticipated that relationships between them identified by co-expression would be more biologically significant than in the *cis*-eQTL datasets. This possibility was investigated by studying *trans*-eQTL clusters.

The presence of *trans*-eQTL clusters in this and other eQTL datasets was already well-characterised when this study began, and others e.g. (Lan et al., 2006) had observed that significantly co-expressed genes often map in *trans* to the same genomic locations. These findings were replicated in this study – 81 *trans*-eQTL clusters, defined as 10 or more *trans*-eQTLs mapping to the same region of the genome, were identified (listed in ***Appendix III***) – accounting for a large proportion of the significant correlation observed across the *trans*-eQTL dataset. Across the four tissues, 47.5-63.7% of significantly correlated *trans*-eQTLs were found to have significantly correlated genotypes at their peaks of linkage (**Table 4.2**). Of these, 23.1-39.1% have peaks of linkage that are ≤1cM apart ((Grieve et al., 2008), Table S2). Furthermore, when the other significantly correlated pairs were investigated, many of these were found to consist of the same eQTLs as in the *trans*-eQTL clusters. This can be observed on the scatterplots plotted by eQTL map location (**Figure 4.4**) as 'bands' of significant correlations which coincide with the loci of clusters' peaks of linkage, such as D17Rat1 – the marker at the locus of the largest *trans*-eQTL cluster in the fat tissue. The band corresponding to this marker's genetic map location can be seen just to the left of the 2500cM mark on **Figure 4.4A**. The genes forming the *trans*-eQTL cluster are correlated not only with each other, but in many cases with other *trans*-eQTL genes that are not linked to the region.

The presence of tissue-specific cluster-forming groups of highly correlated transcripts of differing sizes goes some way to explaining the difference in levels of significant co-expression between the tissues. While it was evident from the data in **Table 4.2** that *trans*-eQTL clusters explained some of the difference between the percentage significant correlation in fat and that in the other tissues, it remained unclear why many more pairs without significantly correlated genotypes at the peak of linkage (i.e. not in the same *trans*-eQTL cluster) were found to be co-expressed in fat than in the other tissues. The observation that so many cluster-forming *trans*-eQTLs are also co-expressed with genes in other clusters (whose peaks of linkage do not have correlated genotypes with that of the first cluster) suggests that tissues in which there are large *trans*-eQTL clusters are likely to have more of both types of significant co-expression.

However, the biological basis of this co-expression, if indeed there is any – it has been speculated that the co-expressed groups of genes that form *trans*-eQTL clusters may be artifacts of non-biological variation (Breitling et al., 2008) – is still to be fully elucidated. Addtionally, the statistical power and precision of both the detection of *trans*-eQTLs (median expected FDR for *trans*-eQTLs across the four tissues = 26-35% (Petretto et al., 2006a)) and the co-expression analysis itself (**Table 4.3**) are low. This is mainly due to a combination of the generally small effect sizes of *trans*-eQTLs (Hubner et al., 2005) and the size of the panel, which is much smaller than would be ideal for this type of analysis (Carlborg et al., 2005).

The *trans*-eQTL clusters themselves were studied in more depth in order to investigate the possibility of biologically significant groups of genes within the *trans*-eQTL dataset, and perhaps generate more specific hypotheses regarding their regulation. Partial correlation analysis was performed to investigate the effect on the levels of correlation of expression profiles observed in *trans*-eQTL clusters of conditioning on the SDP of the marker at the peak of linkage of the cluster, and analysis of GO annotations was performed to test the clusters for functional enrichment.

Partial correlation and functional analysis were performed on larger *trans*-eQTL clusters (those consisting of 30 or more *trans*-eQTLs), as those below this size were considered to provide insufficient statistical power for informative conclusions to be drawn from the output.

The partial correlation analysis found the levels of significant correlation of expression levels within the 15 large *trans*-eQTL clusters to be very variable. The percentage significant pairwise partial correlation across these clusters, conditioned on their respective marker genotypes, was found to range from 18.2-93.1% (**Table 4.6**).

The tests of functional enrichment in large *trans*-eQTL clusters, performed using DAVID, found there to be significantly over-represented GO terms at $p<0.01$ in 80% of the clusters (**Table 4.7**, ***Appendix IV***). The findings of functional enrichment concord with the observation made in (Ghazalpour et al., 2005) and elsewhere that *trans*-eQTL clusters correspond with hotspots of genes enriched for a specific biological function or forming part of a common pathway. The observation that groups of highly correlated genes are frequently functionally enriched (Emilsson et al., 2008) is reinforced through carrying out the reverse analysis, i.e. first identifying the groups of genes linked to common genetic regions, then carrying out the co-expression and functional analysis.

The findings of the functional enrichment analysis, that some of the *trans*-eQTL cluster-forming groups of transcripts are significantly functionally enriched, can be considered in view of the outcome of the partial correlation analysis. The two groups of co-expressed genes that were found to be most significantly correlated once the effect of the genotypes at the peak of linkage had been taken into account (**Table 4.6**) were those with peak of linkage at D15Rat29 in LV (93.1% significant partial correlation) and that with peak of linkage at D17Rat1 in fat (83.6% significant partial correlation). These two groups were also found to be the most functionally enriched, at $p < 0.01$ and $p < 0.05$, of the 15 clusters tested by gene ontology analysis using DAVID (**Table 4.7**). 27 of 130 genes (20.8%, 16 of the probesets were not identified by DAVID) in the D17Rat1 cluster were annotated with metal ion binding molecular function. In the D15Rat29 cluster, 13 of 47 genes (27.7%, 7 of the probesets were not identified by DAVID) were annotated with response to biotic stimulus biological process, and 9 (19.2%) with DNA binding molecular function (see ***Appendix IV*** for full details of all functional enrichment found).

The application of partial correlation analysis to the *trans*-eQTL cluster-forming groups of transcripts in this dataset found that most, but not all, become much less significantly co-expressed when the effect of the genotype at the peak of linkage is taken into account. Those clusters that retain significance were found also to be more significantly enriched for biological function, suggesting that the application of this partial correlation method may be informative in differentiating those co-expressed groups that arise as a consequence of real co-expression from those that may instead be due to technical factors.

Finally, having identified groups of genes that are highly co-expressed and functionally enriched, the next stage in the investigation was to try and elucidate the biological basis of their co-regulation. The 'master regulator' hypothesis (Yvert et al., 2003) suggests that the *cis*-eQTLs colocalising with the peak of linkage of the *trans*-eQTL cluster are the best candidates.

In previous investigations of *trans*-eQTL cluster regulation, correlation has been carried out between the *cis*-eQTLs in the same region of the genome as the *trans*-eQTL cluster, however - as shown in **Figures 4.6 & 4.7**, the confounding effect of linkage disequilbrium means that it is not so straightforward as this to prioritise candidate regulators. When co-expression analysis of *trans*-eQTLs and candidate regulator *cis*-eQTLs was performed, a strong negative relationship between distance of the candidate from the peak of linkage and correlation with the *trans*-eQTL cluster transcripts was identified. This systematic effect is due to linkage effects and does not occur when unlinked transcripts are correlated with the transcripts forming the cluster. This indicates that carrying out uninformed co-expression analysis of *trans*-eQTL cluster transcripts and candidate regulators in the cluster linkage region is not a suitable method for prioritising candidate ‘master regulators’ of highly correlated groups of cluster-forming transcripts.

However, the ‘master regulator’ hypothesis and in particular the way it was investigated in **4.2.5** relise on the cluster-forming groups of *trans*-eQTLs actually representing co-ordinately regulated transcripts by a gene at or near the peak of linkage. As discussed, the work of others (e.g. (Kang et al., 2008)) as well as the findings of the partial correlation analysis described in **4.2.3** suggest that may not actually be the case in many instances of *trans*-eQTL clusters identified by linkage analysis and/or uninformed co-expression analysis.

Chapter 5: Quantitative Trait Transcript Analysis of Physiological and Expression Traits

5.1 Introduction

The analysis of microarray data together with conventional linkage and physiological trait measurements has facilitated the identification of genes underlying individual differences in several complex traits in a variety of species (Glazier et al., 2002). Before the innovation of genome wide eQTL analysis, most of the quantitative trait genes that were identified in this way were found by conventional genetic mapping of the quantitative trait by linkage or association analysis to a genetic interval, followed by fine-mapping of the interval (Flaherty et al., 2005).

The first way in which the data produced by eQTL analysis can be of use in the identification of candidate genes underlying physiological QTLs (referred to henceforth in this chapter as pQTLs in order to avoid confusion) is by identifying drivers of localised regulation of gene expression in the form of *cis*-eQTLs (**1.8.1**). Since *cis*-eQTLs are indicative of genes that are differentially expressed as a consequence of genetic variation in or close to the transcribed gene, they may suggest biological roles for specific genes. Depending on the experimental design, it may be possible to then relate this differental expression to previously mapped phenotypic variation, aiding prioritisation of candidate genes for the pQTL in question (Drake et al., 2006).

The eQTL dataset upon which the analysis here described was generated in RI strains generated from an intercross between a spontaneously hypertensive and a normotensive inbred strain of rat (SHR x BN) as described in (**2.1.1**). The SHR is an established model of the metabolic syndrome, obesity, and cardiovascular disease. Over a hundred physiological traits have been mapped in this cross (Hubner et al., 2005) and in the BXH/HXB panel itself (Pravenec et al., 2002). This means that *cis*-eQTLs colocalising with several mapped traits of substantial medical research interest can be identified and the genes underlying them prioritised for further investigation as a potential contributor to the phenotypic variation between the BN and SHR.

This outlined method of moving towards the identification of quantitative trait genes can be augmented by making use of strain-specific continuous physiological measurements (e.g. blood pressure) and expression data, in an approach called 'Quantitative Trait Transcript (QTT) analysis'. The underlying hypothesis of QTT analysis is that transcripts whose expression profiles are significantly correlated with the trait are promising candidate genes for the trait (Passador-Gurgel et al., 2007).

Measurements were obtained for a wide variety of metabolic, biochemical and physiological traits relating to the metabolic syndrome, type 2 diabetes, hypertension and left ventricular hypertrophy (LVH) and linked to the four tissues of study (**1.5**). The QTT analysis of traits relating to LVH is described in **5.2.1**, and the analysis of traits measured in the adipose tissue, kidney and adrenal, and other metabolic traits, is described in **5.2.2**.

5.2 Results

5.2.1 Quantitative Trait Transcript Analysis of Left Ventricular Mass and Related Traits

Quantitative Trait Transcript analysis was performed as outlined in **2.7**, using data from the left ventricle tissue. The expression profiles of 1,444 LV *cis*-eQTLs were correlated with profiles consisting of indexed measurements of Left Ventricular Mass (LVM) of four to six biological replicates in each RI strain (Petretto et al., 2008). Additionally, in order that QTLs potentially underlying blood-pressure dependent hypertrophy could be distinguished from genetic determinants of LVM acting independently of blood pressure, QTT analysis was carried out on measurements of Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) in the RI strains.

Those *cis*-eQTLs significantly correlated with LVM ($p < 0.01$) are shown in **Table 5.1**, and a graphical representation of the pattern of correlation across the genome is provided by the scatterplot **Figure 5.1**.

Cis-eQTL	Marker at *cis*-eQTL Peak of Linkage	Probeset Chromo-some	Probeset Map Position (Mb)	Pearson Correlation Coefficient	Empirical *p*-value
1380127_at	D17Ucsf2	17	11.011	0.66702	0.0001
1390960_at	D8Mit11	8	122.986	0.6094	0.0006
1373378_at	D17Ucsf2	17	11.156	0.60066	0.0006
1382468_at	D17Ucsf2	17	12.220	0.59515	0.0009
1397727_at	D17Ucsf2	17	10.821	-0.56226	0.0019
1388603_a_at	D8Cebr16s5	8	126.911	0.55541	0.0018
1385456_at	D8Cebr16s5	8	127.910	-0.5479	0.0024
1386128_at	D17Ucsf2	17	12.044	-0.53985	0.002
1391776_at	D8Mit11	8	123.017	0.53803	0.0024
1383759_at	Drd1a	17	11.106	0.53044	0.0026
1384301_at	D17Mit7	17	12.620	0.52931	0.0031
1383413_at	D8Cebr16s5	8	127.153	0.5253	0.0033
1377452_at	D10Rat73	10	31.282	-0.51764	0.005
1388613_at	D8Cebr16s5	8	127.910	0.51128	0.0051
1369415_at	D17Ucsf2	17	10.815	0.50104	0.0058
1389151_at	D4Rat240	4	144.142	0.50102	0.0042
1389376_at	D4Rat240	4	151.764	0.48886	0.0075
1369932_a_at	D10Rat30	10	66.609	0.48481	0.0071
1370004_at	Cacna1s	4	154.891	-0.48354	0.008
1388812_at	D17Ucsf2	17	14.489	-0.48012	0.0074
1376749_at	**D17Utr1**	17	20.987	**-0.46367**	**0.011**
1380383_at	Ppy	10	90.702	-0.46329	0.0112
1385899_at	D10Rat31	10	49.009	0.46224	0.0104
1392077_at	D3Cebr9s1	3	92.371	0.45966	0.0122
1395073_at	D2Mit18	2	114.886	-0.45901	0.0104
1383263_at	**D17Utr1**	17	20.992	**-0.45549**	**0.0121**
1370044_at	Rbp2	8	103.843	0.44906	0.0124
1378802_at	D1Rat235	1	248.162	0.44266	0.0151
1387438_at	D10Cebr1016	10	35.788	-0.44258	0.0173
1383070_at	D17Ucsf2	17	16.012	0.66702	0.0001

Table 5.1: Quantitative Trait Transcript left ventricle *cis*-eQTLs highly correlated with Left Ventricular Mass across 29 RI strains

Probesets shown in **bold** represent *Ogn*, the gene that was prioritised and subsequently shown to be a key regulator of LVM.

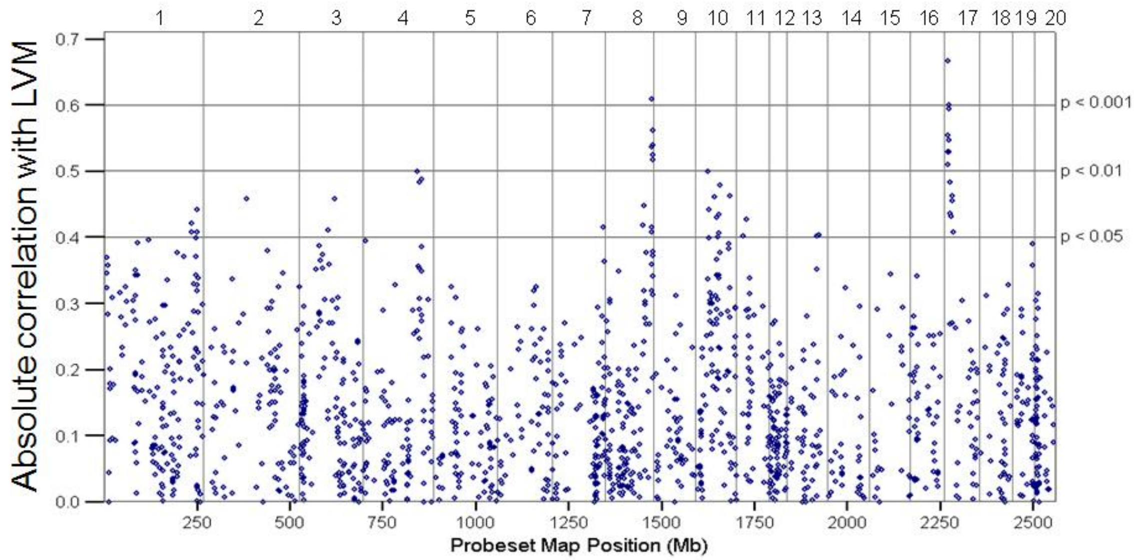


Figure 5.1: Quantitative Trait Transcript (QTT) analysis of left ventricle *cis*-eQTLs with indexed Left Ventricular Mass measured in BXH/HXB RI strains

Each point on the plot represents one of 1,444 *cis*-eQTL transcripts. Approximate empirical *p*-value thresholds are given.

Two groups of *cis*-regulated transcripts were identified as significantly correlated with LVM, on chromosome 17 (most significant p=0.0001) and chromosome 8 (most significant p=0.0006). Correlation was not observed at either of this loci with systolic (**Figure 5.2**) or diastolic (**Figure 5.3**) blood pressure, suggesting the presence of candidate determinants of LVM in one or both of these regions that may be independent of blood pressure.

The results of a separate statistical analysis not forming part of this thesis but discussed in (Petretto et al., 2008) suggest that the effect of blood pressure on LVM is small, and previous genetic analysis (Cerutti et al., 2006) of the relationship between blood pressure and gene expression in the rat LV has found the effect of blood pressure to be minimal – suggesting that blood pressure independent regulation of LVM plays an important role in genetic determination of variation.

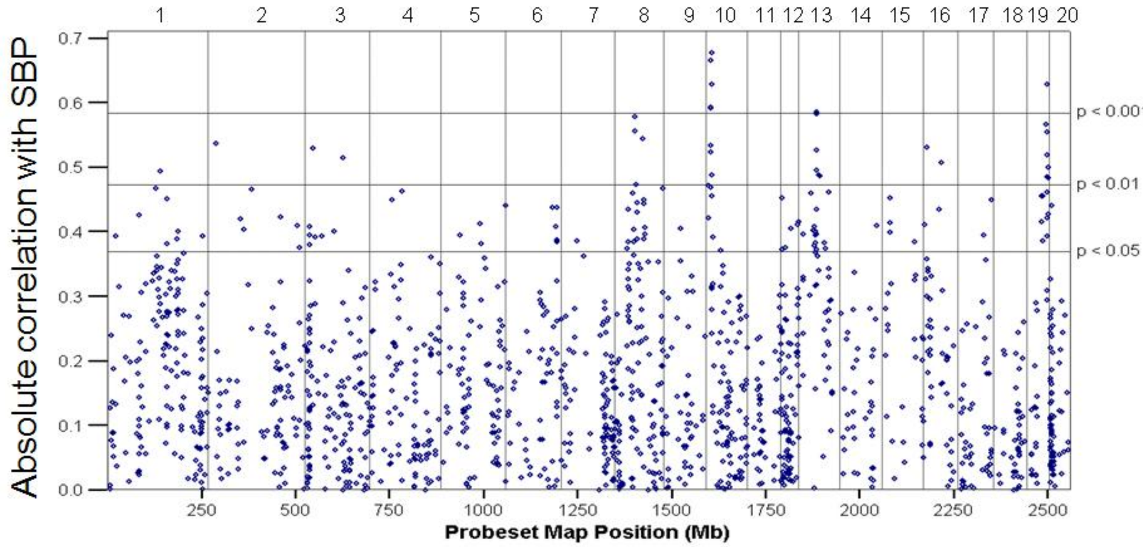


Figure 5.2: QTT analysis of left ventricle *cis*-eQTLs with Systolic Blood Pressure measured in BXH/HXB RI strains.

Symbols as Figure 5.1

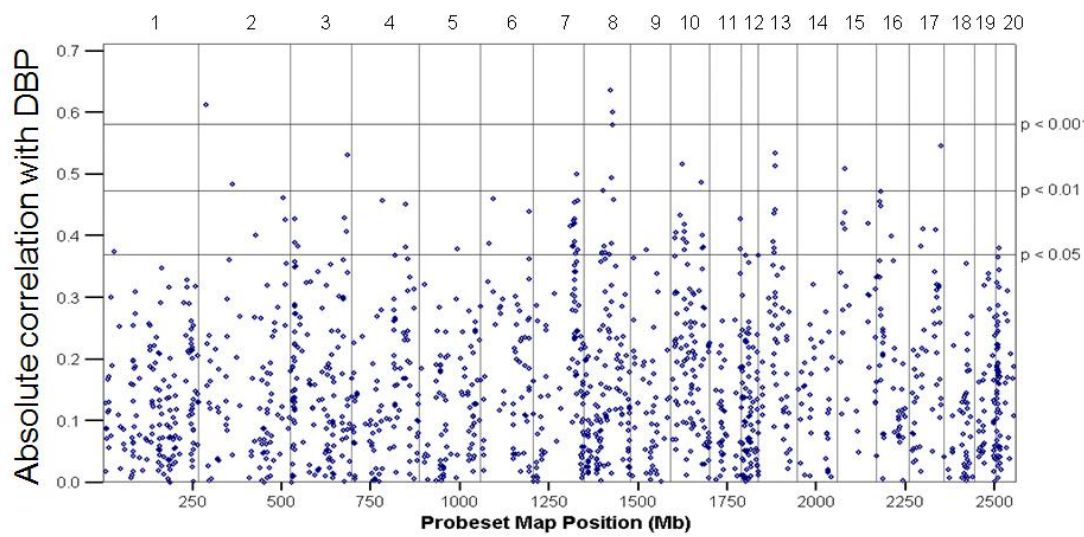


Figure 5.3: QTT analysis of left ventricle *cis*-eQTLs with Diastolic Blood Pressure measured in BXH/HXB RI strains.

Symbols as Figure 5.1

Further statistical analysis was performed on one of the regions identified by this process as candidates for the genetic determination of LVM; chromosome 17p14, which contains a pQTL for LVM previously identified in the BXH/HXB panel (Pravenec et al., 1995). Candidate genes were prioritised on the basis of the significance of their *cis* linkage and their degree of differential expression between SHR and BN.

Of the candidate genes in the QTL region 17p14 showing significant QTT with LVM, *Ogn* (Osteoglycin) was prioritised following fine-mapping using an additional 46 SNP markers within the region of interest, carried out by Enrico Petretto (Petretto et al., 2008). *Ogn* was uncorrelated with blood pressure, and showed the strongest fold change of the candidates in the region. Another candidate – *Hbld2* – was excluded, as discussed in (Petretto et al., 2008), due to its mapping outside the QTL region. *Ogn*, a gene previously known to researchers as an extracellular matrix (ECM) protein with a role in bone morphogenesis (Bentz et al., 1991), was prioritised, sequenced, and - following subsequent *in vitro* and *in vivo* analysis - implicated as a key regulator of LVM in rodents.

To begin to establish whether the rodent findings of *Ogn* regulation of LVM translate to humans, a further QTT analysis was performed using microarray data generated from biopsy samples from 26 human subjects. Expression data was correlated with LVM data obtained through echocardiography. The genes found to be best-correlated with LVM by straightforward QTT analysis (illustrated in **Figure 5.4**) in these 26 subjects are listed in **Table 5.2**. *OGN*, the human orthologue of *Ogn*, is included in this list with an empirical *p*-value of correation of *p*=0.0012.

Probeset	Chromosome	Gene Symbol	Correlation Coefficient	Nominal *p*-value	Empirical *p*-value
205846_at	12	PTPRB	0.70871	0.00005	0.0003
210848_at	N/A	N/A	-0.69006	0.00010	0.0003
219029_at	5	C5orf28	0.69001	0.00010	0.0002
220719_at	12	FLJ13769	-0.68556	0.00011	0.0003
206390_x_at	4	PF4	-0.68221	0.00012	0.0001
208169_s_at	1	PTGER3	-0.68218	0.00012	0.0001
214369_s_at	11	RASGRP2	-0.6772	0.00014	0.0002
201995_at	8	EXT1	0.67261	0.00017	0.0002
220698_at	20	MGC4294	-0.65907	0.00025	0.0004
210772_at	19	FPRL1	-0.65757	0.00026	0.0006
218184_at	6	TULP4	0.64803	0.00034	0.0001
213096_at	1	TMCC2	-0.63705	0.00047	0.0004
208476_s_at	10	FRMD4A	-0.62746	0.00060	0.0009
205453_at	17	HOXB2	-0.62711	0.00061	0.0004
221912_s_at	1	CCDC28B	-0.62697	0.00061	0.0004
216112_at	1	PKN2	-0.62493	0.00064	0.0006
219126_at	6	PHF10	0.62293	0.00068	0.0004
214814_at	4	YTHDC1	0.62019	0.00073	0.0007
204422_s_at	4	FGF2	0.61849	0.00076	0.0012
206003_at	4	CEP135	0.61829	0.00076	0.0008
210145_at	1	PLA2G4A	0.6179	0.00077	0.0011
203944_x_at	6	BTN2A1	0.61786	0.00077	0.001
216818_s_at	N/A	N/A	-0.61553	0.00082	0.0009
218730_s_at	**9**	**OGN**	**0.61546**	**0.00082**	**0.0012**
214981_at	13	POSTN	-0.61393	0.00085	0.001
208370_s_at	21	DSCR1	0.61386	0.00085	0.0009
209672_s_at	7	FLJ20323	0.60973	0.00094	0.0014
213150_at	7	HOXA10	-0.60954	0.00095	0.0015
216664_at	N/A	N/A	-0.60903	0.00096	0.0009
220979_s_at	1	ST6GALNAC5	-0.60848	0.00097	0.0014

Table 5.2: Transcripts whose expression was found to be most highly correlated with LVM in 26 human individuals by Pearson correlation.

OGN, which was subsequently implicated in the regulation of Left Ventricular Mass in rats and humans, is highlighted in bold.

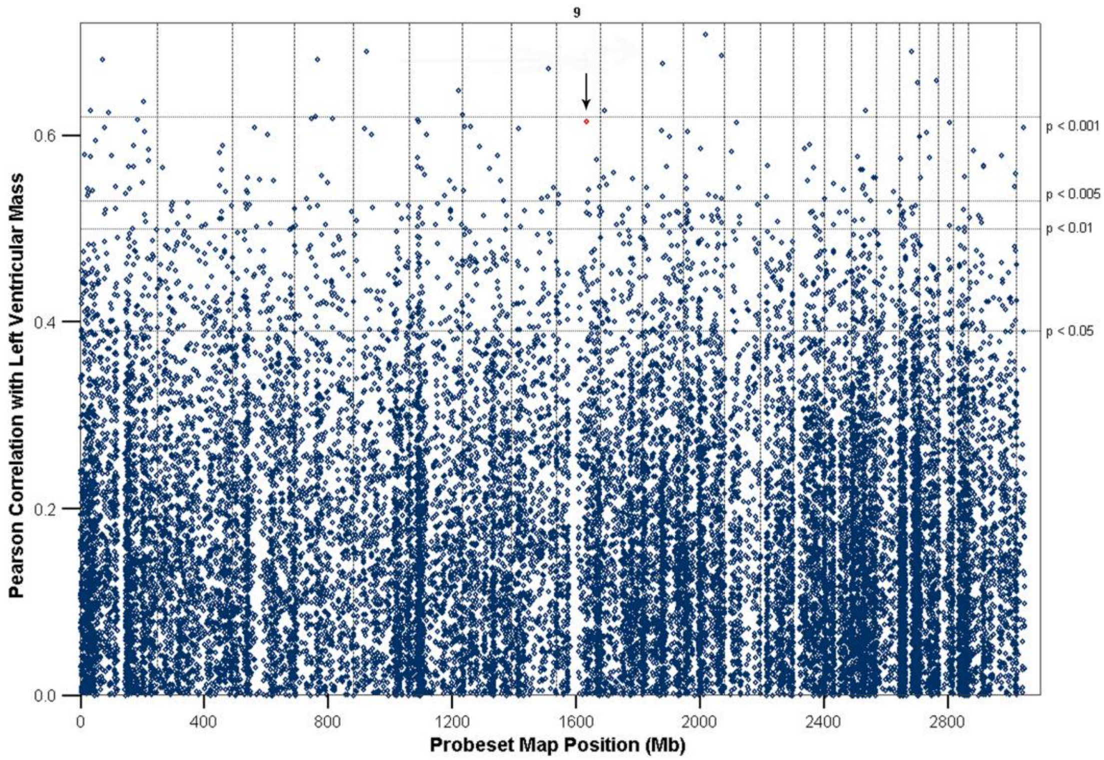


Figure 5.4: Genome-wide correlation of expression data with LVM measurements in 26 human subjects

Each point on the plot represents one of 1,444 human gene transcripts. Approximate empirical *p*-value thresholds are given. OGN transcript is shown in red and is indicated by the arrow

Figure 5.5 shows the Pearson correlation of normalised *OGN* expression with LVM in each of the 26 subjects. Note that *OGN* was found to be positively correlated with LVM in humans, whereas in rats the correlation was negative. This inconsistency was investigated by RT-PCR, as discussed in (Petretto et al., 2008), and found to be a consequence of alternative splicing in the 3'UTR of rat *Ogn*.

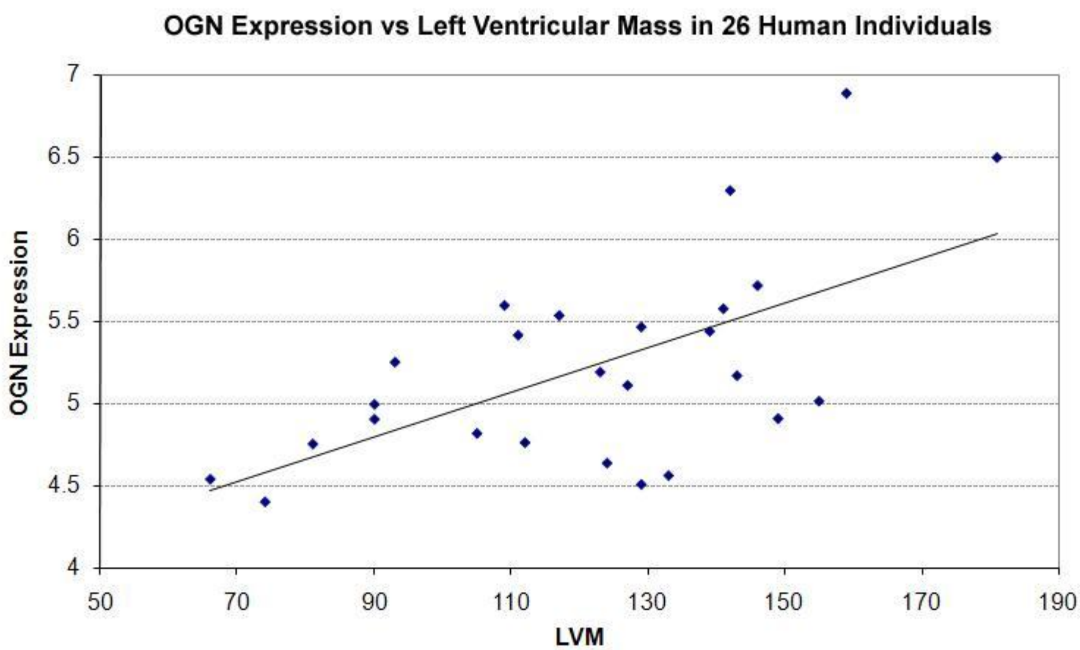


Figure 5.5: Correlation of *OGN* expression measured in 26 human individuals with Left Ventricular Mass measurements

5.2.2 Quantitative Trait Transcript Analysis of non LVM-related Traits

As well as the blood pressure and LVM data discussed in **5.2.1**, QTT analysis was carried out on many other quantitative traits that have been measured in the RI strains. Much of this analysis is preliminary, in that the findings described here have not been progressed into the laboratory, and as such no candidate genes have been prioritised for the traits discussed here. All QTT analysis was performed following assessment of data quality. Data cleansing of physiological traits was carried out by other members of the Aitman laboratory

QTT analysis was typically carried out on traits with relevance both to the metabolic syndrome or associated disorders modelled by the SHR (e.g. genetic hypertension), and to the tissues of study. Although some of the physiological traits are known to be affected by variation in more than one tissue and in others not included in the study, they have been divided here for ease of explanation into three categories: Adrenal traits, kidney traits, and fat (adipose tissue) traits.

5.2.2.1 Adrenal-related Physiological Traits

QTT analysis was performed on 14 separate adrenal traits measured by Daniel O'Connor (UCSD) in 3-5 individuals from each of the RI strains.

The factors secreted by the adrenal gland have a substantial influence on blood pressure through their influence on renal sodium reabsorption (Binder, 2007). The catecholamines (the most significant of which are dopamine, epinephrine and norepinephrine) have also been associated with insulin resistance, dyslipidaemia and glucose intolerance (Landsberg, 1999; Reynisdottir et al., 1994).

Levels of each of the three catecholamines listed above were measured in the RI strains. Additionally, catecholamine biosynthetic enzymes (Ciaranello et al., 1975) phenylethanolamine-N-methyltransferase (PNMT) and dopamine-β-hydroxylase (DBH) were quantified; as were three of the chromogranins – proteins that perform a variety of roles including regulation of catecholamine secretion (Zhang et al., 2006).

As well as the catecholamines, other hormones produced by the adrenal gland are thought to play a role in the development of the metabolic syndrome. Corticosterone is a glucocorticoid hormone, the rodent equivalent of cortisol, which is known to contribute to the hypertensive response to stress (Andrews and Walker, 1999). The level of serum corticosterone was measured.

Finally, the concentrations of sodium, potassium and chloride in the blood were quantified. High levels of sodium and chloride (Schmidlin et al., 2005) are associated with hypertension, as is deficiency in potassium (Channa et al., 2004). The balance of sodium and potassium ions is regulated by the renin-angiotensin-aldosterone system (Delles et al., 2008), in which the kidney and adrenal gland both play central roles.

As an example, the outcome of the QTT analysis of serum chloride concentration against *cis*-eQTLs in the adrenal gland is shown in **Figure 5.6**, and the 15 adrenal *cis*-eQTLs found to be correlated with serum chloride at empirical $p < 0.01$ are listed in **Table 5.3**. The same analysis was carried out for each of the quantitative traits provided.

The serum chloride QTT region on chromosome 1 coincides with two physiological QTLs for changes in body weight, Bw18 and Bw20, mapped in an SHR cross.

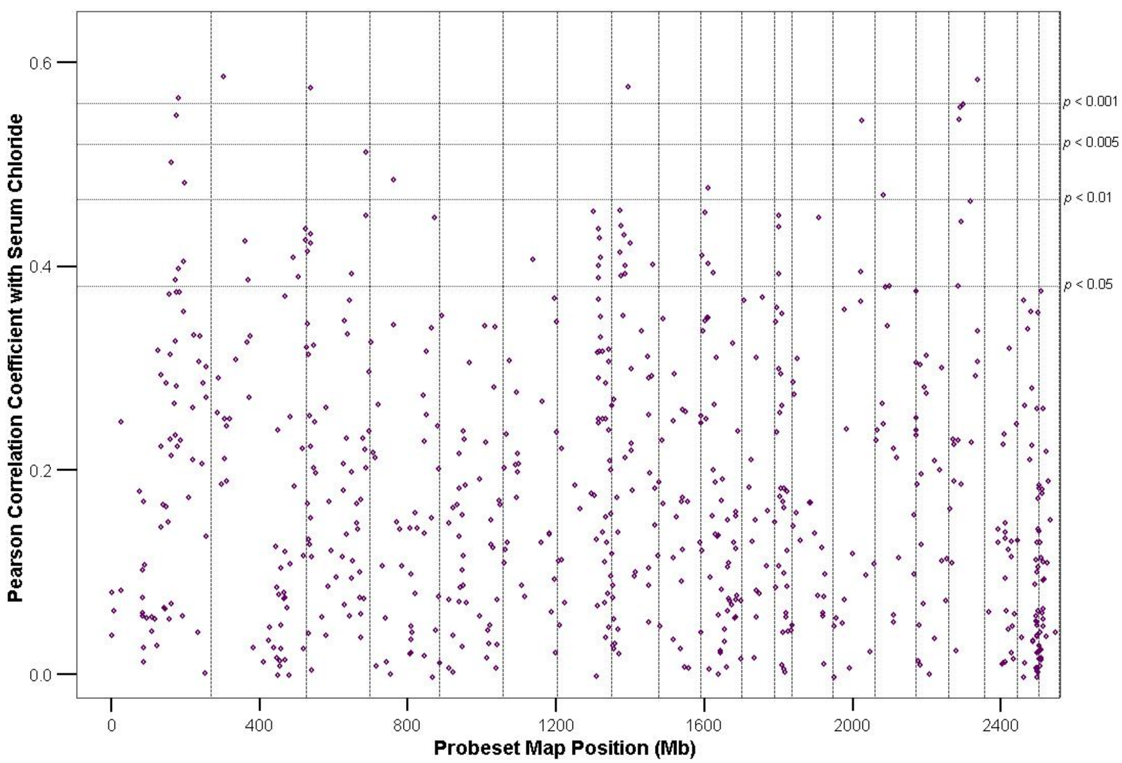


Figure 5.6: QTT analysis of adrenal gland *cis*-eQTLs with serum chloride levels measured in BXH/HXB RI strains.

Symbols as Figure 5.1

Cis-eQTL Probeset	Chr	Probeset Map Position (Mb)	Gene Symbol	Absolute Pearson Correlation Coefficient	Empirical *p*-value
1368585_at	2	30.890	Cart	0.59	0.0006
1368267_at	3	11.366	Pomt1	0.58	0.0006
1372068_at	17	71.418	BS69 protein	0.58	0.0007
1370665_at	8	47.356	Hyou1	0.58	0.0009
1373672_at	17	37.799	Similar to SPI6	0.56	0.0011
1377290_at	1	182.032	Tnrc6a	0.57	0.0013
1388617_at	17	37.144	Bphl	0.55	0.0013
1370929_at	14	81.306	Lrpap1	0.54	0.0017
1367788_at	1	186.870	Phkg2	0.55	0.0023
1368225_at	17	40.124	Rsec5	0.56	0.0027
1393856_x_at	N/A	N/A	--	0.50	0.0054
1373316_at	3	165.811	--	0.51	0.0059
1390625_at	4	76.093	RGD1304879	0.49	0.0077
1368311_at	1	196.867	Mgmt	0.48	0.0078
1374467_at	10	11.762	Trap1	0.48	0.009

Table 5.3: Quantitative Trait Transcript candidates identified as significantly correlated to an exemplar adrenal gland-related trait (serum chloride level)

15 adrenal *cis*-eQTLs were found to be correlated to this trait with empirical $p<0.01$. Numbers of *cis*-eQTLs significantly correlated to 19 other adrenal gland traits ranged from 3 to 23 (median: 9.5).

5.2.2.2 Kidney-related Physiological Traits

The kidney is known to play a key role in the development of genetic hypertension (Rettig and Grisk, 2005). The renin-angiotensin-aldosterone system involves both the kidneys and adrenal gland – renin is a hormone secreted by the kidneys – and regulated blood pressure by inducing blood vessel contraction and by modulating the balance of sodium and potassium ions (Binder, 2007). Additionally, mutations in several genes governing renal salt handling have been found to produce disorders of blood pressure regulation (Ji et al., 2008).

Measurements of systolic (SBP) and diastolic blood pressure (DBP), pulse pressure (PP) and heart rate (HR) were tested against *cis*-eQTLs in the kidney by QTT analysis. The outcome of the QTT analysis of pulse pressure is shown in **Figure 5.7**, and the kidney *cis*-eQTLs found to be correlated with each of the four traits tested at empirical $p < 0.01$ are listed in **Table 5.4**.

A pulse pressure QTT identified on chromosome 19 at empirical $p < 0.001$ coincides with a blood pressure QTL, Bp17, that had previously been mapped in the BXH/HXB cross, and a heart rate QTT region identified on chromosome 7 at empirical $p < 0.005$ is colocalised with a pQTL for albuminuria (renal function), Uae23, that was mapped in an SHR/Fub backcross.

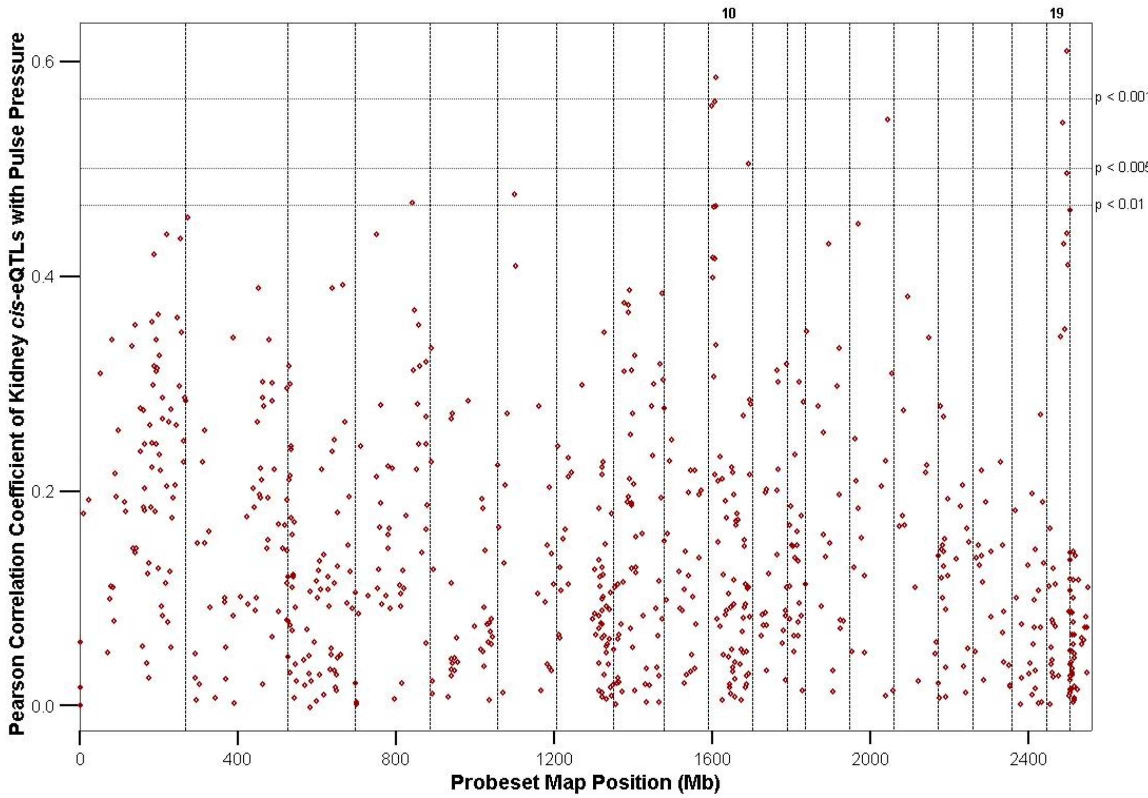


Figure 5.7: QTT analysis of Kidney *cis*-eQTLs with pulse pressure measured in BXH/HXB RI strains.

Symbols as Figure 5.1. Potential QTT regions of interest were identified on chromosomes 10 and 19.

Cis-eQTL Probeset	Chr	Probeset Map Position (Mb)	Gene Symbol	Absolute Pearson Correlation Coefficient	Empirical *p*-value
1368891_at	19	55.047	Gnpat	0.61	0.0007
1388151_at	10	11.067	Coro7	0.58	0.0008
1371785_at	10	12.941	Tnfrsf12a	0.56	0.0014
1370705_at	10	11.919	Znf597	0.56	0.0016
1370830_at	14	97.703	Egfr	0.55	0.0016
1387157_at	19	39.411	Pmfbp1	0.54	0.0027
1371870_at	19	53.732	Similar to FKSG26	0.50	0.0048
1373438_at	10	106.647	Ube2op	0.51	0.0054
1375532_at	6	42.781	Id2	0.48	0.0079
1376136_at	10	15.110	Rhot2	0.47	0.01

Table 5.4 a): Kidney *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to Pulse Pressure by QTT analysis

Cis-eQTL Probeset	Chr	Probeset Map Position (Mb)	Gene Symbol	Absolute Pearson Correlation Coefficient	Empirical *p*-value
1386212_at	13	90.028	Spna1	0.72	0
1387146_a_at	15	87.893	Ednrb	0.60	0.0006
1374379_at	13	88.637	Dusp23_ Predicted	0.48	0.0076

Table 5.4 b): Kidney *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to Heart Rate by QTT analysis

Cis-eQTL Probeset	Chr	Probeset Map Position (Mb)	Gene Symbol	Absolute Pearson Correlation Coefficient	Empirical *p*-value
1371689_at	N/A	N/A	--	0.54	0.0019
1375966_at	16	10.052	Mmrn2	0.52	0.0028
1371749_at	7	123.813	Celsr1_ Predicted	0.53	0.0034
1374782_at	7	113.676	Rhpn1_ Predicted	0.51	0.0041
1376501_at	7	122.794	Arhgap8	0.50	0.0057
1374608_at	4	149.617	--	0.50	0.0058
1389800_at	1	199.042	Pwwp2	0.48	0.0072
1379783_at	7	113.394	--	0.47	0.008

Table 5.4 *c)*: Kidney *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to Diastolic Blood Pressure by QTT analysis

Cis-eQTL Probeset	Chr	Probeset Map Position (Mb)	Gene Symbol	Absolute Pearson Correlation Coefficient	Empirical *p*-value
1370705_at	10	11.919	Znf597	0.57	0.0008
1387157_at	19	39.411	Pmfbp1	0.58	0.001
1371785_at	10	12.941	Tnfrsf12a	0.56	0.0014
1377303_at	10	14.137	--	0.55	0.0025
1370269_at	8	61.468	Cyp1a1	0.50	0.0054
1370830_at	14	97.703	Egfr	0.49	0.0058
1388151_at	10	11.067	Coro7	0.50	0.0059
1374782_at	7	113.676	Rhpn1_ Predicted	0.48	0.0086
1377342_s_at	3	10.091	Fnbp1	0.47	0.0092

Table 5.4 *d)*: Kidney *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to Systolic Blood Pressure by QTT analysis

5.2.2.3 Fat-related and Miscellaneous Physiological Traits

The adipose tissue, along with muscle and liver, is central to the pathogenesis of the metabolic syndrome, insulin resistance and type II diabetes (Biddinger and Kahn, 2006).

QTT analysis was performed systematically on 75 metabolic traits, using *cis*-eQTLs from adipose tissue and also from the other three tissues. The measurements that were carried out to explain this physiological data are detailed in ***Appendix V***. Full data on all QTT *cis*-eQTLs at $p<0.01$ by permutation in each of the four tissues are provided on the **Data CD**. Totals for each of the four tissues tested are given in **Table 5.5 *a)*** (fat), ***b)*** (kidney), ***c)*** (adrenal gland) and ***d)*** (left ventricle).

Co-localisation analysis of the QTTs identified in this analysis and those described earlier in this thesis with physiological QTLs has not been carried out to date, but work on the development of testable functional hypotheses from these data is ongoing in the host laboratory, including the development of congenic strains (Morrissey, C., 2010, personal communication).

ins=0/cell	8	MAP	11	totchol0	8
absnefa/cell	9	RLVHWT	13	totphos0	4
ins=50k-0/cell	17	kidneydif	13	totchol4	18
fold	6	kidney5wk	29	totphos4	4
ins50k/cell	23	kidney8wk	14	ins0/vol	5
absnefa	11	livchol	4	ins50/vol	19
ratwt	8	liver(g)	4	delins/vol	11
cell vol	23	livpool	4	hematocrit 90	8
ins0	7	offspring	3	platelet aggrega	4
ins15	4	IDLch0	8	CORT-20	14
glu0	1	IDLch4	12	CORT-40	5
glu15	6	IDLph0	3	CORT-80	4
ins/glu0	7	IDLph4	4	CORT-140	7
ins/glu15	5	LDLch0	14	ACTH-20	21
IPGTT0	5	LDLch4	9	NA/h/kg	11
IPGTT30	1	LDLph0	9	A/h/kg	8
IPGTT60	12	LDLph4	10	DOP/h/kg	7
IPGTT120	6	HDL2ch0	9	CTA	14
ch0	7	HDL2ch4	6	CUPOOL	6
ch7	6	HDL2ph0	11	CUDW	5
ch15	0	HDL2ph4	10	CUWW	2
tg0	5	HDL3ch0	26	NO	7
tg7	9	HDL3ch4	6	Padwt/BW	30
tg15	1	HDL3ph0	2	AUC	7
HDL15	4	HDL3ph4	5	log fold	6

Table 5.5 *a)*: Total numbers of fat *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to 75 metabolic traits by QTT analysis

10,000 permutations were used to find empirical *p*-values for each trait (definitions provided in ***Appendix V***). A cumulative total of 665 QTT candidates were identified. Full data from this analysis is provided on the **Data CD**.

ins=0/cell	18	MAP	6	totchol0	7
absnefa/cell	12	RLVHWT	18	totphos0	5
ins=50k-0/cell	27	kidneydif	3	totchol4	10
fold	8	kidney5wk	21	totphos4	4
ins50k/cell	29	kidney8wk	16	ins0/vol	2
absnefa	7	livchol	7	ins50/vol	10
ratwt	14	liver(g)	8	delins/vol	6
cell vol	12	livpool	5	hematocrit 90	9
ins0	3	offspring	8	platelet aggrega	7
ins15	7	IDLch0	4	CORT-20	16
glu0	6	IDLch4	8	CORT-40	7
glu15	11	IDLph0	7	CORT-80	9
ins/glu0	5	IDLph4	3	CORT-140	7
ins/glu15	6	LDLch0	13	ACTH-20	15
IPGTT0	13	LDLch4	9	NA/h/kg	15
IPGTT30	3	LDLph0	9	A/h/kg	13
IPGTT60	18	LDLph4	7	DOP/h/kg	9
IPGTT120	7	HDL2ch0	11	CTA	16
ch0	10	HDL2ch4	8	CUPOOL	12
ch7	2	HDL2ph0	12	CUDW	12
ch15	3	HDL2ph4	9	CUWW	16
tg0	2	HDL3ch0	25	NO	8
tg7	10	HDL3ch4	4	Padwt/BW	18
tg15	6	HDL3ph0	9	AUC	15
HDL15	12	HDL3ph4	12	log fold	11

Table 5.5 *b)*: Total numbers of kidney *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to 75 metabolic traits by QTT analysis

1,000 permutations were used to find empirical *p*-values for each trait (definitions provided in ***Appendix V***). A cumulative total of 752 QTT candidates were identified. Full data from this analysis is provided on the **Data CD**.

ins=0/cell	6	MAP	11	totchol0	4
absnefa/cell	6	RLVHWT	7	totphos0	8
ins=50k-0/cell	10	kidneydif	1	totchol4	5
fold	9	kidney5wk	8	totphos4	4
ins50k/cell	11	kidney8wk	8	ins0/vol	2
absnefa	12	livchol	8	ins50/vol	6
ratwt	7	liver(g)	7	delins/vol	6
cell vol	9	livpool	7	hematocrit 90	7
ins0	4	offspring	6	platelet aggrega	3
ins15	8	IDLch0	2	CORT-20	6
glu0	4	IDLch4	6	CORT-40	2
glu15	3	IDLph0	9	CORT-80	8
ins/glu0	3	IDLph4	11	CORT-140	5
ins/glu15	8	LDLch0	7	ACTH-20	15
IPGTT0	4	LDLch4	5	NA/h/kg	5
IPGTT30	3	LDLph0	9	A/h/kg	12
IPGTT60	3	LDLph4	17	DOP/h/kg	5
IPGTT120	8	HDL2ch0	13	CTA	15
ch0	4	HDL2ch4	3	CUPOOL	7
ch7	2	HDL2ph0	12	CUDW	12
ch15	2	HDL2ph4	5	CUWW	4
tg0	5	HDL3ch0	14	NO	9
tg7	10	HDL3ch4	6	Padwt/BW	12
tg15	3	HDL3ph0	17	AUC	4
HDL15	6	HDL3ph4	4	log fold	11

Table 5.5 *c)*: Total numbers of adrenal *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to 75 metabolic traits by QTT analysis

1,000 permutations were used to find empirical *p*-values for each trait (definitions provided in ***Appendix V***). A cumulative total of 530 QTT candidates were identified. Full data from this analysis is provided on the **Data CD**.

ins=0/cell	26	MAP	30	totchol0	10
absnefa/cell	34	RLVHWT	19	totphos0	18
ins=50k-0/cell	27	kidneydif	12	totchol4	13
fold	13	kidney5wk	45	totphos4	7
ins50k/cell	33	kidney8wk	37	ins0/vol	13
absnefa	27	livchol	21	ins50/vol	24
ratwt	33	liver(g)	8	delins/vol	17
cell vol	24	livpool	24	hematocrit 90	23
ins0	13	offspring	16	platelet aggrega	26
ins15	21	IDLch0	3	CORT-20	18
glu0	12	IDLch4	22	CORT-40	8
glu15	11	IDLph0	14	CORT-80	19
ins/glu0	17	IDLph4	17	CORT-140	10
ins/glu15	17	LDLch0	21	ACTH-20	13
IPGTT0	25	LDLch4	16	NA/h/kg	29
IPGTT30	8	LDLph0	11	A/h/kg	21
IPGTT60	20	LDLph4	13	DOP/h/kg	9
IPGTT120	11	HDL2ch0	22	CTA	25
ch0	11	HDL2ch4	11	CUPOOL	7
ch7	7	HDL2ph0	35	CUDW	22
ch15	9	HDL2ph4	12	CUWW	5
tg0	12	HDL3ch0	22	NO	19
tg7	27	HDL3ch4	2	Padwt/BW	20
tg15	8	HDL3ph0	12	AUC	17
HDL15	18	HDL3ph4	13	log fold	20

Table 5.5 *d)*: Total numbers of left ventricle *cis*-eQTLs identified as significantly correlated (empirical $p<0.01$) to 75 metabolic traits by QTT analysis

1,000 permutations were used to find empirical *p*-values for each trait (definitions provided in ***Appendix V***). A cumulative total of 1335 QTT candidates were identified. Full data from this analysis is provided on the **Data CD**.

5.3 Discussion

By improving our understanding of the relationships between genes underlying *cis*- and *trans*-eQTLs, it becomes possible to construct co-expression networks and make hypotheses regarding genetic modules. However, in order to achieve real insight into the genetics of complex traits such as the metabolic syndrome and left ventricular hypertrophy (LVH), it is necessary to integrate physiological trait data into this type of statistical genetic and computational analysis.

The use of rodent models in research into the pathophysiology and genetics of the metabolic syndrome (Lusis et al., 2008) and LVH (Sarwar and Cook, 2009) has formed the basis of much of our understanding of the mechanisms underlying these pathologies. Although there are of course substantial differences in terms of biology between humans and rodents, homologues of genes first identified in rodents have in several cases been hypothesised to underlie aspects of human disease. *Cd36* (Aitman et al., 1999), for instance, was the first complex trait gene identified through the use of DNA microarrays, in the SHR. Human *CD36* protein is 86% identical to that found in the rat (Lepretre et al., 2004), and *CD36* was subsequently implicated in human insulin resistance (Miyaoka et al., 2001), and has been associated with components of the metabolic syndrome (Love-Gregory et al., 2008) as well as type 2 diabetes (Lepretre et al., 2004).

Rodents have several inherent advantages over humans as sources of data for gene discovery (Guryev and Cuppen, 2009). The QTT analyses described here have two main components – the identification of *cis*-eQTLs in biologically relevant tissues, and the correlation of the expression of these with the quantitative traits measured in the same strains . Rodent RI strains are ideal for this purpose because they are renewable, meaning that the animals in which the traits are measured are genetically identical to those from which the tissue samples for microarray analysis were taken (**1.4**) (Pravenec, 2008). In addition, the range and depth of clinical research that can be carried out in humans, particularly healthy 'control' subjects, is limited for ethical and practical reasons. This often means that studies in humans are less specific than would be optimal for gene discovery (Delles et al., 2008). As discussed in Chapter 4 and in (Petretto et al., 2006a), gene expression is highly tissue-specific.

In rodents, obtaining tissues is not a problem and it was possible in the studies described here to perform full gene expression and QTT analysis on *cis*-eQTLs in four tissues that are highly relevant to the metabolic syndrome and/or LVH: Adipose tissue, kidney, adrenal gland and left ventricle.

In **5.2.1**, a QTT analysis that led to the implication of *Ogn* in blood pressure-independent regulation of LVM in the RI strains, and subsequently of *OGN* in human LVM (Petretto et al., 2008), was described. It had previously been shown that blood pressure effects do not substantially correlate with LVM in both humans and rodents (Sarwar and Cook, 2009), even though hypertension and LVH are both risk factors for heart failure (Schocken et al., 2008). The *Ogn* transcripts were not found by the QTT analysis to be correlated with either systolic or diastolic blood pressure. The cardiac role of *Ogn*, along with that of *Ephx2* (Monti et al., 2008), was discovered following genome-wide expression profiling in a rat model – in the latter case, an F2 cross derived from SHHF, the stroke-prone spontaneously hypertensive rat. *Ephx2* was subsequently found to be differentially regulated in human heart failure.

In **5.2.2**, analyses involving 103 measurements of a variety of quantitative traits in the other three tissues were summarised. The QTT data produced through correlation of *cis*-eQTL gene expression with these measurements across the RI strain panel potentially informs a wide range of hypotheses on the relationships between genome and phenome with regard to the components of the metabolic syndrome and related traits.

Similar relationships between gene expression and trait measurements as were identified on chromosome 17 in the investigation of LVM regulation in the left ventricle (**Figure 5.1**) were found in other traits. These included serum chloride, for which potential adrenal QTTs were found on chromosomes 1 and 17 (**Figure 5.6**), and pulse pressure, for which potential kidney QTTs were found on chromosomes 10 and 19 (**Figures 5.7**). It was not possible for these analyses to be taken forward within the time-frame of this thesis. However, data is provided on the **Data CD** for the purpose of mining for other potential relationships between *cis*-eQTL genes and phenotypes, and investigations using these data are ongoing in the host laboratory.

The worth of QTT analysis in this dataset in precipitating gene discovery has been demonstrated here (**5.2.1**, (Petretto et al., 2008)) and elsewhere (Pravenec et al., 2008). In the latter, *Cd36* was identified as a significant QTT for diastolic blood pressure. A blood pressure pQTL had previous been identified in the vicinity of the *Cd36* locus, and *Cd36* transcript expression was found to be inversely correlated with radiotelemetric measurements of systolic and diastolic blood pressure at $p < 0.01$.

Although the 'genetical genomics' (Jansen and Nap, 2001) methodology has been widely adopted by geneticists working in a variety of species, the integration of physiological trait data to such genome-wide expression analyses only began relatively recently. The term 'quantitative trait transcript' was coined in (Passador-Gurgel et al., 2007), in which QTT analysis in *Drosophila melanogaster* was used to identify transcripts whose expression correlated with survival on exposure to nicotine, although the potential of approach of correlating clinical trait values with transcript levels had been discussed previously (Drake et al., 2006). An approach that with hindsight could be seen as a forerunner of QTT analysis was successfully applied to prioritise *Insig2* as a gene influencing plasma cholesterol levels in mice (Cervino et al., 2005).

Recent QTT analyses carried out in humans (Emilsson et al., 2008) and in mice (Chen et al., 2008) have identified genes expressed in adipose tissue and in liver that influence obesity-associated traits, such as plasma HDL cholesterol – many of which have also been measured and analysed in the rat RI strains (**5.2.2**). In addition to the identification of three previously unknown obesity-related genes (*Lpl*, *Lactb* and *Ppm1l*), Chen *et al* described the use of their eQTL data to identify functional units in metabolic networks that are hypothesised to regulate the physiological traits found by QTT to be correlated with transcripts in the network. Some of the component genes of one of these networks were subsequently validated, in that genetic perturbation resulted in significant alterations to the metabolic traits under consideration, such as fat/muscle ratios (Yang et al., 2009). The same metabolic network, the macrophage-enriched metabolic network (MEMN) was identified in humans (Emilsson et al., 2008), showing that rodent research can lead to improved understanding of human systems at more than a single-gene level.

There are some caveats associated with the use of QTT analysis to identify candidate genes for the regulation of physiological traits. As a method entailing direct correlation of transcript expression with quantitative trait measurements, it is not able to identify genetic regulation of traits by means other than transcriptional control, such as alternative splicing (Drake et al., 2006) or DNA methylation (Hansen et al., 2008), nor is QTT suitable for identification of coding sequence variation which may impact upon phenotype, except where these may alter the amount of transcript produced (as in the case of *Ogn*). Additionally, QTT analysis cannot in and of itself provide information on complex traits, only on their quantifiable components (Wentzell et al., 2007).

More prosaically, it is required for gene identification that probes representing the transcript concerned be present on the microarray chip. The recent introduction of next generation sequencing (NGS) approaches to expression profiling, by direct RNA sequencing (RNA-Seq) may obviate some of these difficulties associated with microarrays by allowing transcript abundance to be determined without the need for pre-designed probes (Wang et al., 2009). Either way, the integration of the methods such as this with other approaches, such as the network construction methods applied by (Chen et al., 2008), and the analysis of more diverse data types (Schadt, 2009) will be required in order to gain real insight into the complex genetics of traits like the metabolic syndrome.

Chapter 6: General Discussion

The overall aims of this study were to make use of the combined linkage and gene expression data available in four tissues from the BXH/HXB panel to improve our understanding of the relationships between gene expression traits, eQTLs and physiological traits modelled by the SHR. This entailed the undertaking of a systematic analysis of the correlation structures of *cis*- and *trans*-eQTLs and the genetic basis of *trans*-eQTL clusters across all four tissues. Additionally, the control of transcript levels potentially underlying one physiological trait, LVM, was studied in depth, and QTT analysis was carried out on a further 103 traits using *cis*-eQTL data.

6.1 Summary of Findings

In **Chapter 3**, a full-scale comparative validation of the output of the linkage analysis software used to generate the eQTL dataset is described. The demonstration of the irreproducible nature of linkage data produced by QTL Reaper where there was missing genotype data at the peak of linkage motivated the repair of the software bug and the remapping of the eQTLs, excluding many spurious *trans*-eQTLs and clusters. Separately, an investigation of the possibility of false *cis*-eQTLs as a consequence of polymorphisms in the microarray probes' binding regions is discussed. An exploratory study of a statistical model for predicting potential false *cis*-eQTLs, in which newly available SHR sequence data was used to test the model's predictions, is described.

In **Chapter 4**, the application of an FDR-based method for determining statistical significance of pairwise correlations of gene expression across the 29 strains is described. This enabled the correlation structure of the sets of *cis*- and *trans*-eQTLs in all four tissues to be visualised and contrasted. Statistically significant correlation of *cis*-eQTLs is rare, and genetic linkage or long-range allelic association are potential explanations for it in almost all cases. Statistically significant correlation of *trans*-eQTLs is much commoner; however there is much more potential, given the definitions of *cis*- and *trans*-eQTLs, for correlated transcripts to be identified in the *trans*-eQTL dataset. Groups of co-expressed transcripts forming *trans*-eQTL clusters were identified in all four tissues; in most cases these were tissue-specific. Functional analysis of these *trans*-eQTL cluster-forming groups of co-expressed transcripts, carried out using DAVID (a Gene Ontology based tool) found several of the larger *trans*-eQTL clusters to be functionally enriched. Partial correlation analysis of these clusters found the two most functionally enriched also to be the most significantly co-expressed once genotype at the peak of linkage was taken into account. The findings described in this chapter have several implications for the use of gene expression and eQTL data to predict functional relationships between genes and in the construction of networks.

In **Chapter 5**, a QTT analysis carried out using *cis*-eQTLs in the left ventricle and physiological trait measurements obtained in the RI strains is described. This QTT analysis led to the prioritisation of a region on chromosome 17 and in particular the gene *Ogn* as a candidate regulator of left ventricular mass in the rat, a function that was subsequently confirmed in rodents. An associated QTT analysis that was carried out using data from 26 human individuals, showing a clear positive correlation between *OGN* expression and left ventricular mass, is described. QTT analyses that were performed systematically using *cis*-eQTL data from all four tissues and 103 physiological measurements (**Table 5.5**, ***Appendix V***) are summarised – data produced by these analyses is provided on the **Data CD**.

6.2 General Discussion

The analysis described in this thesis was carried out on eQTL datasets derived from a panel of 29 rat RI strains. It has been possible to validate and generate testable hypothesis by applying statistical and computational methods to this dataset, and in the case of *Ogn* make use of it to precipitate the discovery of a complex trait gene. However, as discussed in **1.8.3**, the statistical power to detect eQTL accorded by such a panel is suboptimal. This means that much significant genetic variation, particularly *trans*-acting variation, in transcript levels in these rats is likely to have been missed (Carlborg et al., 2005). In addition, the expected FDR at genome-wide significance $p<0.05$ for *trans*-eQTLs – whose effect sizes are typically small – is high (Petretto et al., 2006a), and most *trans*-eQTL clusters are no longer present in the dataset when a more stringent threshold of p<0.01 is used.

RI strains do provide a boost in statistical power relative to F_2 populations (de Koning et al., 2005) because they offer a simpler genetic model (no heterozygotes); however the inbreeding process is laborious and takes a long time. A project is underway to produce a mouse RI panel consisting of 400 lines, the Collaborative Cross (Mott and Flint, 2008) that will, if successful, offer much greater statistical power to detect QTLs with small effects. In both rats and mice, the development of heterogeneous stock panels, in which QTLs can be directly mapped to a high resolution, by interbreeding inbred strains is ongoing (Aitman et al., 2008). By applying novel analytical methods to a rat heterogeneous stock panel derived from eight inbred strains, an anxiety trait was mapped to within a 95% confidence interval of 13Mb (Johannesson et al., 2009). The BXH/HXB RI panel does however offer benefits to researchers into the metabolic syndrome and associated traits that are not shared by the heterogeneous stock, thanks to its derivation from the SHR (**1.3.3**). The sequencing of the SHR genome (Atanur et al., 2010), which in combination with the BN sequence enabled the study described in **3.3** to be performed, also offers the potential to generate a marker map for use in genetic analysis of the RI strains of much greater resolution than is currently available, through the identification of novel SNPs, indels and copy number variants.

The work described in chapters **4** and **5** involves the application of correlation-based methods to genomic, transcriptomic and quantitative phenotypic trait data. Such methods are widely used in statistical genetics as they provide a straightforward and easily understood means of quantifying relationships within large datasets, and because correlation methods are amenable for use in the construction of models such as co-expression networks (Emilsson et al., 2008). Pairwise correlation methods such as Pearson correlation (**2.4.1.1**) are suitable for testing for linear relationships, which biological relationships – particularly across levels e.g. between transcript and phenotypic trait – may not be (Jansen et al., 2009). These methods do not directly provide information about causality (Wu et al., 2009). Measurements of pairwise correlation is influenced by confounding factors such as, in the case of co-expression analysis, linkage disequilibrium (Lusis et al., 2008) which typically means that an eQTL transcript is correlated with a wide region rather than a single marker. The interpretation of correlation of closely linked genes is further complicated by the fact that closely linked but separate genes may influence the same trait, and their effects may interact (Rockman and Kruglyak, 2006).

Some such confounding factors can be taken into account through the use of partial correlation, which enables the correlation between two variables to be found taking into account the effect of others (**2.4.5**). This method was applied to groups of 30 or more transcripts linked in *trans* to the same genetic locus (large *trans*-eQTL clusters) in **4.2.3**. It has been suggested that *trans*-eQTL clusters are mostly artefacts, resulting from the influence of unknown confounding factors – e.g. the method used to standardise RNA abundance across samples (Kliebenstein, 2009) – and that they occur in eQTL datasets generated in relatively small populations, such as the BXH/HXB RI panel, because if one of a group of highly correlated transcripts is linked to a region of the genome, the others necessarily (as an effect of that correlation) will be also. It is therefore possible that a large cluster may by chance be linked to a specific locus, and that therefore no gene at this locus is involved in regulating the expression of cluster's constituent transcripts.

A model-based method of accounting for the inter-sample correlation and removing spurious *trans*-eQTL clusters has been proposed, and in a sample dataset applying this removed all such groups (Kang et al., 2008). Some studies, e.g. (Ghazalpour et al., 2008), have however found clusters that remain significant after correcting for such effects. The outcome of the investigations using partial correlation and functional enrichment described in **4.2.3** and **4.2.4** suggest that at least two clusters, one of which consists of 146 *trans*-eQTL genes, in the BXH/HXB dataset may be genuinely co-expressed, biologically interesting and worthy of further investigation, for instance through the application of pathway analysis methods (Kliebenstein, 2009). However, the utility of such methods, as well as the GO-based functional enrichment analysis described in **4.2.4**, is limited by the quality of the annotation available (Auffray et al., 2009).

Another limitation associated with the eQTL study design has to do with the biological source of the RNA samples used in expression profiling. As shown by this

(**4.2.2**) and other (e.g. (Yang et al., 2006)) studies, gene expression is highly tissue-specific. Furthermore, it has been found to be highly cell-type dependent (Gerrits et al., 2009). Tissue samples used in this and many other eQTL studies (e.g. (Chesler et al., 2005)) are heterogeneous in terms of cell-type – for instance, the adrenal gland samples used in this study includes both cortex and medulla. This means that inter-sample variation can occur as a consequence of differing sample compositions (Gerrits et al., 2009), although this will be true across the four biological replicates as well as strains, so in many but not necessarily all cases will be taken into account.

The linkage mapping of quantitative physiological traits such as blood pressure to the genome as physiological QTLs (pQTLs) is a well established methodology, and was carried out in the inbred strains used to generate the RI strain dataset prior to the commencement of the work described in this thesis (Hubner et al., 2005). These pQTLs are broad chromosomal regions as a consequence of linkage disequilibrium and contain many potential candidate genes. The information on expression variation provided by eQTLs can help to narrow down the list of candidate causative genes (Haley and de Koning, 2007).

One way of doing this is to test for correlation between the phenotypic and expression traits (Drake et al., 2006) by carrying out QTT analysis, as was used to prioritise *Ogn* as a candidate gene underlying a cardiac mass pQTL in **5.2.1**. Here, *cis*-eQTLs across the genome were tested for correlation with physiological measurements across the RI panel. It should be borne in mind that, in a GeneChip-generated eQTL dataset such as this, some *cis*-eQTLs have been found (Alberts et al., 2007b) to arise as a consequence of differential hybridisation due to SNPs in probe binding regions, rather than genuine differential expression of the transcript(s) (**3.2 & 3.3**). For those *cis*-eQTLs that do represent differentially expressed transcripts, it is true that the *cis*-eQTLs' linkage to the QTL region implies a potential relationship between their expression and the physiological trait without the need for QTT analysis, and also that linkage disequilibrium means that fine mapping is required to identify the candidate gene from the significant QTTs. Although the presence of the QTL was taken into account, it was mapped in the parental strains not the RI panel. Additionally, the use of QTT in this instance enabled the relationship between the expression traits, LVM and blood pressure to be visualised and compared easily (**Figures 5.1-5.3**).

The QTT method is however not limited to *cis*-eQTLs and has been used in other settings on transcripts not previously known to be linked to a region of interest (Emilsson et al., 2008). It has been suggested that the QTT approach could potentially be used in personalised medicine, combining data on the effect of a drug with expression analysis in a human clinical population (Ruden, 2007), although difficulties with such a study include the aforementioned tissue/cell-type specificity of variation in expression, the variation inherent in a human population, making achieving a suitable statistical power to achieve significance problematic. The latter two are also issues which arise where GWAS studies are considered (**1.2.3**), and difficulties in replication that have been reported in association with GWAS (Frayling et al., 2007) would probably arise with this method also. Model organisms, and RI panels in particular, allow much more highly controlled and more reproducible experiments to be carried out than do humans (Guryev and Cuppen, 2009), and access to relevant tissues/cell types is much less of an issue.

Other limitations associated with QTT analysis include the fact that not all regulation of biological processes takes place through control of gene expression (Drake et al., 2006), that unaccounted-for confounding factors can produce uninformative findings of relationships (Yang et al., 2009), and that the effects of any one locus on a physiological phenotype may be too small to be detected by QTT analysis (Schliekelman, 2008).

Like conventional QTL analysis, the QTT is a method of generating testable hypothesis regarding genes underlying trait variation – it does not prove anything in and of itself. The testing of specific causative hypotheses, as ever, takes place in the laboratory using methods, e.g. fine-mapping, *in vitro* models, RT-PCR and the generation of transgenics, such as were used to identify *Ogn* as a regulator of cardiac mass in (Petretto et al., 2008). Other QTT candidates identified as described in **5.2.2** are currently being investigated in the host laboratory by methods including the generation of congenic strains (Morrissey C., 2010, personal communication). This method was used, following QTT analysis in the RI strains, along with transgenesis to identify *Cd36* as a quantitative trait gene for blood pressure (Pravenec et al., 2008), and also in the prioritisation of *Acaa2* as a candidate for blood pressure regulation (Johnson et al., 2009).

Computational approaches, including the integration of expression data into molecular networks (Chen et al., 2008), pathway analysis methods such as Ingenuity Pathways Analysis (Wang et al., 2007), and GSEA (functional analysis using curated gene sets (Subramanian et al., 2005))(Yang et al., 2009) have been used to assess the functional relevance of genes found to be correlated to a phenotypic trait.

Because many genes contribute to phenotypic variation in complex traits, the consideration of expression variation within an integrative framework that takes into account this complexity may be more useful in understanding disease states (Schadt, 2009). Data on the correlation between phenotypic and expression variation may contribute to the effort to construct models that incorporate interactions between different levels (e.g. genome, transcriptome, metabolome) within the cell and the organism as a whole (Hansen et al., 2008). The application of statistical procedures, such as the likelihood-based causal model selection (LCMS) probabilistic model described by (Schadt et al., 2005), can help differentiate between correlations that are causal to a clinical phenotype and those that are reactive to it (Yang et al., 2009), and also assess the joint probabilities of two closely linked loci influencing the phenotype independently of one another (Schadt, 2009). This method was used in the identification of the MEMN as a co-expression network having a causal relationship with obesity traits (Chen et al., 2008).

The methods of collecting biological data and of analysing it are continually developing. In the study of gene expression, RNA-Seq has been described as *"a revolutionary tool for transcriptomics"* (Sarwar and Cook, 2009) that some have speculated may represent the beginning of the end for microarrays (Auffray et al., 2009). This recently-developed method is based on next-generation sequencing (NGS) technologies which enable tens of millions of DNA molecules to be sequenced in parallel (Simon et al., 2009). RNA is sequenced after being converted to cDNAs, or – in a still newer method – directly (Ozsolak et al., 2009). RNA-Seq has been shown by comparison with qRT-PCR able to quantify transcript abundance accurately and has been found to be reproducible across replicates (Wang et al., 2009). Additionally, RNA-Seq has high dynamic range; it is able to detect transcripts that are expressed at both very low and very high levels (Mortazavi et al., 2008; Turner et al., 2009). It has also been shown to be well-suited to translational studies (Sugarbaker et al., 2008) and investigation of splice variants, chromatin modification sites and protein and transcription factor binding sites (Guryev and Cuppen, 2009).

With RNA-Seq, the quantification of transcript abundance is not limited to those transcripts that are represented on a chip, and cross-hybridisation and spurious signals such as that discussed in **3.2** and **3.3** (Alberts et al., 2007b; Simon et al., 2009) do not impact data quality as with microarrays. Relative to microarrays, RNA-Seq is however very expensive at present, and poses computational and analytical challenges because of the massive quantities of data generated, the short reads, and the identification of reads that span exon junctions or contain large polymorphisms (Wang et al., 2009). Because microarrays are relatively cheap and easy to use and analytical methodologies are now well-established, they are unlikely to be supplanted in the near future (Simon et al., 2009). Hence, improvements in our understanding of microarrays and their limitations remain of great research value.

The extent of the complexity of the biological systems underlying traits such as the metabolic syndrome is just beginning to be appreciated (Schadt, 2009). The generation of ever-larger datasets using improving technologies with greater range and sensitivity is anticipated to provide the materials to investigate this complexity and build our understanding of the biology of complex traits. In order for these data to be exploited to the full and meaningful hypotheses to be generated from them, computational techniques and methods of statistical analysis must keep up with the pace.

6.3 Concluding Remarks

The use of combined gene expression and linkage analysis to generate regulatory and functional hypotheses has been proven here and elsewhere to be a powerful technique. In this thesis, the utility of correlation analysis on such a dataset to prioritise *cis*-eQTLs as candidate genes for physiological traits has been demonstrated through the application of QTT analysis with left ventricular mass and blood pressure data to precipitate the identification of *Ogn* as a candidate regulator of left ventricular mass. Many other candidate regulatory genes underlying physiological traits have been generated through the systematic application of this technique, and data resulting from this is provided. The study of the correlation structure of groups of transcripts forming *trans*-eQTL clusters, combined with functional enrichment analysis, has suggested that some such groups may indeed be biologically interesting.

Overall, the results described in this thesis demonstrate the strength of systematic, statistically sound approaches to the analysis of a large, consistent eQTL dataset. The importance of having a thorough understanding of both the biological and computational basis of such a dataset has been emphasised throughout.

References

Aitman, T.J. (2001). CD36, insulin resistance, and coronary heart disease. Lancet *357*, 651-652.

Aitman, T.J., Critser, J.K., Cuppen, E., Dominiczak, A., Fernandez-Suarez, X.M., Flint, J., Gauguier, D., Geurts, A.M., Gould, M., Harris, P.C., *et al.* (2008). Progress and prospects in rat genetics: a community view. Nat Genet *40*, 516-522.

Aitman, T.J., Gotoda, T., Evans, A.L., Imrie, H., Heath, K.E., Trembling, P.M., Truman, H., Wallace, C.A., Rahman, A., Dore, C., *et al.* (1997). Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. Nat Genet *16*, 197-201.

Alberts, R., Terpstra, P., Bystrykh, L.V., de Haan, G., and Jansen, R.C. (2005). A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. Genetics *171*, 1437-1439.

Alberts, R., Terpstra, P., Hardonk, M., Bystrykh, L.V., de Haan, G., Breitling, R., Nap, J.P., and Jansen, R.C. (2007a). A verification protocol for the probe sequences of Affymetrix genome arrays reveals high probe accuracy for studies in mouse, human and rat. BMC Bioinformatics *8*, 132.

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.P., and Jansen, R.C. (2007b). Sequence Polymorphisms Cause Many False cis eQTLs. PLoS ONE *2*, e622.

Alberts, R., Vera, G., and Jansen, R.C. (2008). affyGG: computational protocols for genetical genomics with Affymetrix arrays. Bioinformatics *24*, 433.

Andrews, R.C., and Walker, B.R. (1999). Glucocorticoids and insulin resistance: old hormones, new targets. Clinical Science *96*, 513-523.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Atanur, S., Birol, I., Guryev, V., Hirst, M., Hummel, O., Behmoaras, J., Fernandez, X., Johnson, M.D., McLaren, W., Morrissey, C., *et al.* (2010). The Genome Sequence of the Spontaneously Hypertensive Rat: Analysis and Functional Significance. Genome Research *(submitted)*.

Auffray, C., Chen, Z., and Hood, L. (2009). Systems medicine: the future of medical genomics and healthcare. Genome Medicine *1*, 2.

Bailey, D.W. (1971). Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation *11*, 325-327.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Benjamini, Y., and Yekutieli, D. (2005). Quantitative trait Loci analysis using the false discovery rate. Genetics *171*, 783-790.

Bentley, D.S., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Bentz, H., Thompson, A.Y., Armstrong, R., Chang, R.J., Piez, K.A., and Rosen, D.M. (1991). Transforming growth factor-beta 2 enhances the osteoinductive activity of a bovine bone-derived fraction containing bone morphogenetic protein-2 and 3. Matrix (Stuttgart, Germany) *11*, 269.

Biddinger, S.B., and Kahn, C.R. (2006). From mice to men: insights into the insulin resistance syndromes. Annu Rev Physiol *68*, 123-158.

Binder, A. (2007). A review of the genetics of essential hypertension. Curr Opin Cardiol *22*, 176-184.

Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L.V., de Haan, G., and Su, A.I. (2008). Genetical Genomics: Spotlight on QTL Hotspots. PLoS Genetics *4*, e1000232.

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature *436*, 701-703.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752-755.

Broman, K.W. (2005). Mapping expression in randomized rodent genomes. Nat Genet *37*, 209-210.

Carlborg, O., De Koning, D.J., Manly, K.F., Chesler, E., Williams, R.W., and Haley, C.S. (2005). Methodological aspects of the genetic dissection of gene expression. Bioinformatics *21*, 2383-2393.

Cerutti, C., Kurdi, M., Bricca, G., Hodroj, W., Paultre, C., Randon, J., and Gustin, M.P. (2006). Transcriptional alterations in the left ventricle of three hypertensive rat models. Physiol Genomics *27*, 295-308.

Cervino, A.C., Li, G., Edwards, S., Zhu, J., Laurie, C., Tokiwa, G., Lum, P.Y., Wang, S., Castellini, L.W., and Lusis, A.J. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics *86*, 505-517.

Channa, M.L., Somova, L., and Nadar, A. (2004). Facets of the metabolic syndrome in Dahl hypertensive rats. Cardiovasc J S Afr *15*, 61-63.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., and Fodor, S.P.A. (1996). Accessing Genetic Information with High-Density DNA Arrays. Science *274*, 610-614.

Chen, H., and Sharp, B. (2004). Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics *5*, 147.

Chen, L., Emmert-Streib, F., and Storey, J. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology *8*, R219.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., *et al.* (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429-435.

Chesler, E., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H., Mountz, J., Baldwin, N., and Langston, M. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature Genetics *37*, 233-242.

Chuaqui, R.F., Bonner, R.F., Best, C.J., Gillespie, J.W., Flaig, M.J., Hewitt, S.M., Phillips, J.L., Krizman, D.B., Tangrea, M.A., Ahram, M., *et al.* (2002). Post-analysis follow-up and validation of microarray experiments. Nat Genet *32 Suppl*, 509-514.

Churchill, G.A., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., Beatty, J., Beavis, W.D., Belknap, J.K., Bennett, B., Berrettini, W., *et al.* (2004). The Collaborative

Cross, a community resource for the genetic analysis of complex traits. Nat Genet *36*, 1133-1137.

Churchill, G.A., and Doerge, R.W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. Genetics *138*, 963-971.

Ciaranello, R.D., Wooten, G.F., and Axwlrod, J. (1975). Regulation of dopamine beta-hydroxylase in rat adrenal glands. Journal of Biological Chemistry *250*, 3204-3211.

Clee, S.M. (2005). Expressing the nature of quantitative traits. Clinical Genetics *68*, 1-5.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning *20*, 273-297.

Cowley, M.J., Cotsapas, C.J., Williams, R.B.H., Chan, E.K.F., Pulvers, J.N., Liu, M.Y., Luo, O.J., Nott, D.J., and Little, P.F.R. (2008). Intra–and inter–individual genetic differences in gene expression. Nature Precedings.

Dabney, A., and Storey, J.D. (2008). QVALUE.

de Koning, D.-J., Carlborg, Ö., and Haley, C.S. (2005). The genetic dissection of immune response using gene-expression studies and genome mapping. Veterinary Immunology and Immunopathology *105*, 343-352.

de Koning, D.J., and Haley, C.S. (2005). Genetical genomics in humans and model organisms. Trends in Genetics *21*, 377-381.

de la Cruz, N., Bromberg, S., Pasko, D., Shimoyama, M., Twigger, S., Chen, J., Chen, C.F., Fan, C., Foote, C., Gopinath, G.R., *et al.* (2005). The Rat Genome Database (RGD): developments towards a phenome database. Nucleic Acids Res *33*, D485-491.

De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics *20*, 3565-3574.

Delles, C., McBride, M.W., Padmanabhan, S., and Dominiczak, A.F. (2008). The genetics of cardiovascular disease. Trends in Endocrinology & Metabolism *19*, 309-316.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., and Lempicki, R. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology *4*, R60.

Dickens, N. (2006). Systematic Categorisation of Affymetrix Microarray Probe Set Annotations (SCAMPA).

Doss, S., Schadt, E.E., Drake, T.A., and Lusis, A.J. (2005). Cis-acting expression quantitative trait loci in mice. Genome Res *15*, 681-691.

Drake, T.A., Schadt, E.E., and Lusis, A.J. (2006). Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. Mamm Genome *17*, 466-479.

Eckel, R.H., Grundy, S.M., and Zimmet, P.Z. (2005). The metabolic syndrome. Lancet *365*, 1415-1428.

Editorial (2008). Year of the Rat. Nat Genet *40*, 513-513.

Efron, B., and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol *23*, 70-86.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association *96*, 1151-1160.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., *et al.* (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423-428.

Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., and Blake, J.A. (2005). The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Research *33*, D471.

Ferrannini, E. (2007). Metabolic Syndrome: A Solution in Search of a Problem. J Clin Endocrinol Metab *92*, 396-398.

Flaherty, L., Herron, B., and Symula, D. (2005). Genomics of the future: Identification of quantitative trait loci in the mouse. Genome Res *15*, 1741-1745.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.* (2008). Ensembl 2008. Nucl Acids Res *36*, D707-714.

Flint, J., Valdar, W., Shifman, S., and Mott, R. (2005). Strategies for mapping and cloning quantitative trait genes in rodents. Nat Rev Genet *6*, 271-286.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., and Rayner, N.W. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889.

Fuller, T., Ghazalpour, A., Aten, J., Drake, T., Lusis, A., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. Mammalian Genome *18*, 463-472.

Gerrits, A., Li, Y., Tesson, B., Bystrykh, L., Weersing, E., and Gibson, G. (2009). Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. PLoS Genet *5*, e1000692.

Ghazalpour, A., Doss, S., Kang, H., Farber, C., Wen, P.Z., Brozell, A., Castellanos, R., Eskin, E., Smith, D.J., and Drake, T.A. (2008). High-Resolution Mapping of Gene Expression Using Association in an Outbred Mouse Stock. PLoS Genetics *4*, e1000149.

Ghazalpour, A., Doss, S., Sheth, S.S., Ingram-Drake, L.A., Schadt, E.E., Lusis, A.J., and Drake, T.A. (2005). Genomic analysis of metabolic pathway gene expression in mice. Genome Biol *6*, R59.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J., *et al.* (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet *2*, e130.

Gilad, Y., Rifkin, S.A., Bertone, P., Gerstein, M., and White, K.P. (2005). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Research *15*, 674-680.

Glazier, A.M., Nadeau, J.H., and Aitman, T.J. (2002). Finding genes that underlie complex traits. Science *298*, 2345-2349.

Goring, H.H.H., Curran, J.E., Johnson, M.P., Dyer, T.D., Charlesworth, J., Cole, S.A., Jowett, J.B.M., Abraham, L.J., Rainwater, D.L., Comuzzie, A.G., *et al.* (2007).

Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature *39*, 1208-1216.

Grieve, I.C., Dickens, N.J., Pravenec, M., Kren, V., Hubner, N., Cook, S.A., Aitman, T.J., Petretto, E., and Mangion, J. (2008). Genome-Wide Co-Expression Analysis in Multiple Tissues. PLoS ONE *3*, e4033.

Grundy, S.M., Brewer, H.B., Jr., Cleeman, J.I., Smith, S.C., Jr., Lenfant, C., and for the Conference, P. (2004). Definition of Metabolic Syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on Scientific Issues Related to Definition. Circulation *109*, 433-438.

Guryev, V., and Cuppen, E. (2009). Next-generation sequencing approaches in genetic rodent model systems to study functional effects of human genetic variation. FEBS letters *583*, 1668-1673.

Haley, C.S., and de Koning, D.J. (2007). Towards in vitro genetics. Trends in Genetics *23*, 382-386.

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

Hall, J.E. (2004). The pioneering use of systems analysis to study cardiac output regulation. Am J Physiol Regul Integr Comp Physiol *287*, R1009-1011.

Hansen, B.G., Halkier, B.A., and Kliebenstein, D.J. (2008). Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends Plant Sci *13*, 72-77.

Harbig, J., Sprinkle, R., and Enkemann, S.A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucl Acids Res *33*, e31.

Hill, J.A., and Olson, E.N. (2008). Cardiac Plasticity. New England Journal of Medicine *358*, 1370.

Hosack, D.A., Dennis Jr, G., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol *4*, R70.

Huang, D., Sherman, B., Tan, Q., Collins, J., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M., Lane, H.C., and Lempicki, R. (2007). The DAVID Gene Functional

Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology *8*, R183.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al.* (2002). The Ensembl genome database project. Nucl Acids Res *30*, 38-41.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., *et al.* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet *37*, 243-253.

Hubner, N., Yagil, C., and Yagil, Y. (2006). Novel integrative approaches to the identification of candidate genes in hypertension. Hypertension *47*, 1-5.

Inomata, H., Watanabe, T., Iizuka, Y., Liang, Y.Q., Mashimo, T., Nabika, T., Ikeda, K., Yanai, K., Gotoda, T., and Yamori, Y. (2005). Identification of quantitative trait loci for cardiac hypertrophy in two different strains of the spontaneously hypertensive rat. Hypertens Res *28*, 273-281.

Jacob, H.J. (1999). Functional Genomics and Rat Models. Genome Res *9*, 1013-1016.

Jain, M., Shrager, J., Harris, E.H., Halbrook, R., Grossman, A.R., Hauser, C., and Vallon, O. (2007). EST assembly supported by a draft genome sequence: an analysis of the Chlamydomonas reinhardtii transcriptome. Nucleic Acids Research *35*, 2074.

Jansen, R., Tesson, B., Fu, J., Yang, Y., and McIntyre, L. (2009). Defining gene and QTL networks. Current Opinion in Plant Biology *12*, 241-246.

Jansen, R.C., and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. Trends in Genetics *17*, 388-391.

Jaworski, R.L., Jirout, M., Closson, S., Breen, L., Flodman, P.L., Spence, M.A., Kren, V., Krenova, D., Pravenec, M., and Printz, M.P. (2002). Heart Rate and Blood Pressure Quantitative Trait Loci for the Airpuff Startle Reaction (Am Heart Assoc), pp. 348-352.

Ji, W., Foo, J.N., Brian, J.O.R., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., Matthew, W., Levy, D., and Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nature Genetics *40*, 592-599.

Jirout, M., Krenova, D., Kren, V., Breen, L., Pravenec, M., Schork, N.J., and Printz, M.P. (2003). A new framework marker-based linkage map and SDPs for the rat HXB/BXH strain set. Mamm Genome *14*, 537-546.

Johannesson, M., Lopez-Aumatell, R., Stridh, P., Diez, M., Tuncel, J., Blázquez, G., Martinez-Membrives, E., Cañete, T., Vicens-Costa, E., and Graham, D. (2009). A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. Genome Research *19*, 150.

Johnson, M., He, L., Herman, D., Wakimoto, H., Wallace, C., Zidek, V., Mlejnek, P., Musilova, A., Simakova, M., and Vorlicek, J. (2009). Dissection of Chromosome 18 Blood Pressure and Salt-Sensitivity Quantitative Trait Loci in the Spontaneously Hypertensive Rat. Hypertension *54*, 639-645.

Kang, H.M., Ye, C., and Eskin, E. (2008). Accurate Discovery of Expression Quantitative Trait Loci Under Confounding from Spurious and Genuine Regulatory Hotspots. Genetics *180*, 1909-1925.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support vector machines in R. Journal of Statistical Software *15*, 1–28.

Kendziorski, C., and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. Mammalian Genome *17*, 509-517.

Keurentjes, J.J., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci U S A *104*, 1708-1713.

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics *21*, 3587-3595.

Kirst, M., Basten, C.J., Myburg, A.A., Zeng, Z.-B., and Sederoff, R.R. (2005). Genetic Architecture of Transcript-Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics *169*, 2295-2303.

Kliebenstein, D. (2009). Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. Annu Rev Plant Biol *60*, 93-114.

Kline, R.B. (2004). Principles and practice of structural equation modeling (The Guilford Press).

Kren, V., Pravenec, M., Lu, S., Krenova, D., Wang, J.M., Wang, N., Merriouns, T., Wong, A., St Lezin, E., Lau, D., *et al.* (1997). Genetic isolation of a region of chromosome 8 that exerts major effects on blood pressure and cardiac mass in the spontaneously hypertensive rat. J Clin Invest *99*, 577-581.

Krinke, G.J. (2000). Handbook of Experimental Animals: The Laboratory Rat (New York, Academic Press).

Kruglyak, L. (2008). The road to genome-wide association studies. Nat Rev Genet *9*, 314-318.

Lakka, H.M., Laaksonen, D.E., Lakka, T.A., Niskanen, L.K., Kumpusalo, E., Tuomilehto, J., and Salonen, J.T. (2002). The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men. JAMA *288*, 2709-2716.

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F., *et al.* (2006). Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet *2*, e6.

Landsberg, L. (1999). Role of the sympathetic adrenal system in the pathogenesis of the insulin resistance syndrome. Annals of the New York Academy of Sciences *892*, 84-90.

Lazar, J., Moreno, C., Jacob, H.J., and Kwitek, A.E. (2005). Impact of genomics on research in the rat. Genome Res *15*, 1717-1728.

Lepretre, F., Linton, K.J., Lacquemant, C., Vatin, V., Samson, C., Dina, C., Chikri, M., Ali, S., Scherer, P., and Séron, K. (2004). Genetic study of the CD36 gene In a French diabetic population. Diabetes and Metabolism *30*, 459-463.

Lewin, A., and Grieve, I.C. (2006). Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. BMC Bioinformatics *7*, 426.

Li, H., Lu, L., Manly, K.F., Chesler, E.J., Bao, L., Wang, J., Zhou, M., Williams, R.W., and Cui, Y. (2005). Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet *14*, 1119-1125.

Li, Y., Alvarez, O.A., Gutteling, E.W., Tijsterman, M., Fu, J., Riksen, J.A., Hazendonk, E., Prins, P., Plasterk, R.H., Jansen, R.C., *et al.* (2006). Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet *2*, e222.

Liew, C.-C., and Dzau, V.J. (2004). Molecular genetics and genomics of heart failure. Nat Rev Genet *5*, 811-825.

Lincoln, S.E., Daly, M., and Lander, E. (1993). Constructing Genetic Linkage Maps With MAPMAKER/EXP.

Lorell, B.H., and Carabello, B.A. (2000). Left ventricular hypertrophy: pathogenesis, detection, and prognosis. Circulation *102*, 470-479.

Love-Gregory, L., Sherva, R., Sun, L., Wasson, J., Schappe, T., Doria, A., Rao, D.C., Hunt, S.C., Klein, S., and Neuman, R.J. (2008). Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol. Human Molecular Genetics *17*, 1695.

Lum, P.Y., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T.A., Lusis, A.J., and Schadt, E.E. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. Journal of Neurochemistry *97*, 50-62.

Lusis, A., Attie, A., and Reue, K. (2008). Metabolic syndrome: from epidemiology to systems biology. Nature Reviews Genetics *9*, 819-830.

Manly, K., Nettleton, D., and Hwang, J. (2004). Genomics, prior probability, and statistical tests of multiple hypotheses. Genome Research *14*, 997.

Manly, K.F. (2006). QTL Reaper.

Manly, K.F., Cudmore, R.H., Jr., and Meer, J.M. (2001). Map Manager QTX, cross-platform software for genetic mapping. Mamm Genome *12*, 930-932.

Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J Clin Invest *118*, 1590-1605.

McLaren, D.S. (1997). Is insulin resistance becoming a global epidemic? Nutrition *13*, 64-66.

Mehrabian, M., Allayee, H., Stockton, J., Lum, P.Y., Drake, T.A., Castellani, L.W., Suh, M., Armour, C., Edwards, S., Lamb, J., *et al.* (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet *37*, 1224-1233.

Meng, H., Vera, I., Che, N., Wang, X., Wang, S.S., Ingram-Drake, L., Schadt, E.E., Drake, T.A., and Lusis, A.J. (2007). Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. Proceedings of the National Academy of Sciences *104*, 4530-4535.

Miyaoka, K., Kuwasako, T., Hirano, K., Nozaki, S., Yamashita, S., and Matsuzawa, Y. (2001). CD36 deficiency associated with insulin resistance. Lancet *357*, 686-687.

Monnier, L., Colette, C., and Owens, D.R. (2008). Type 2 diabetes: A well-characterised but suboptimally controlled disease. Can we bridge the divide? Diabetes & Metabolism *34*, 207-216.

Monti, J., Fischer, J., Paskas, S., Heinig, M., Schulz, H., Gösele, C., Heuser, A., Fischer, R., Schmidt, C., and Schirdewan, A. (2008). Soluble epoxide hydrolase is a susceptibility factor for heart failure in a rat model of human disease. Nature Genetics *40*, 529-537.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743-747.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Morton, N.E. (2008). Into the post-HapMap era. Adv Genet *60*, 727-742.

Mott, R. (2006). Finding the molecular basis of complex genetic variation in humans and mice. Philos Trans R Soc Lond B Biol Sci *361*, 393-401.

Mott, R., and Flint, J. (2008). Prospects for complex trait analysis in the mouse. Mamm Genome *19*, 306-208.

Mozhui, K., Ciobanu, D., Schikorski, T., Wang, X., Lu, L., and Williams, R. (2008). Dissection of a QTL hotspot on mouse distal chromosome 1 that modulates neurobehavioral phenotypes and gene expression. PLoS Genetics *4*, e1000260.

Mueller, M., Goel, A., Thimma, M., Dickens, N.J., Aitman, T.J., and Mangion, J. (2006). eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. Bioinformatics *22*, 509-511.

Nathan, D.G., and Orkin, S.H. (2009). Musings on genome medicine: genome wide association studies. Genome Medicine *1*, 3.

Nathan, D.M., Meigs, J., and Singer, D.E. (1997). The epidemiology of cardiovascular disease in type 2 diabetes mellitus: How sweet it is... or is it? The Lancet *350*, 4-9.

Okamoto, K., and Aoki, K. (1963). Development of a strain of spontaneously hypertensive rats. Jpn Circ J *27*, 282-293.

Opgen-Rhein, R., and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology *1*, 37.

Ortlepp, J.R., Breithardt, O., Ohme, F., Hanrath, P., and Hoffmann, R. (2001). Lack of association among five genetic polymorphisms of the renin-angiotensin system and cardiac hypertrophy in patients with aortic stenosis. American Heart Journal *141*, 671.

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., and Milos, P.M. (2009). Direct RNA sequencing. Nature.

Pabo, C.O., and Sauer, R.T. (1992). Transcription factors: structural families and principles of DNA recognition. Ann Rev Biochem *61*, 1053-1095.

Passador-Gurgel, G., Hsieh, W.P., Hunt, P., Deighton, N., and Gibson, G. (2007). Quantitative trait transcripts for nicotine resistance in Drosophila melanogaster. Nat Genet *39*, 264-268.

Pastinen, T., Ge, B., and Hudson, T.J. (2006). Influence of human genome polymorphism on gene expression. Hum Mol Genet *15 Spec No 1*, R9-16.

Peirce, J.L., Lu, L., Gu, J., Silver, L.M., and Williams, R.W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet *5*, 7.

Pérez-Enciso, M., Toro, M.A., Tenenhaus, M., and Gianola, D. (2003). Combining Gene Expression and Molecular Marker Information for Mapping Complex Trait Genes A Simulation Study. Genetics *164*, 1597-1606.

Petretto, E., Mangion, J.M., Dickens, N., Cook, S., Mandek, K.K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., *et al.* (2006a). Heritability and Tissue-Specificity of Expression QTLs. PLoS Genetics *2*, e172.

Petretto, E., Mangion, J.M., Pravenec, M., Hubner, N., and Aitman, T.J. (2006b). Integrated gene expression profiling and linkage analysis in the rat. Mammalian Genome *17*, 480-489.

Petretto, E., Sarwar, R., Grieve, I., Lu, H., Kumaran, M., Muckett, P., Mangion, J., Schroen, B., Benson, M., and Punjabi, P. (2008). Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass. Nature Genetics *40*, 546.

Pounds, S. (2006). Estimation and control of multiple testing error rates for microarray studies. Briefings in bioinformatics *7*, 25.

Pravenec, M. (1996). A genetic linkage map of the rat derived from recombinant inbred strains. Mammalian Genome *7*, 117-127.

Pravenec, M. (2008). Insight into the genetics of hypertension, a core component of the metabolic syndrome. Curr Opin Clin Nutr Metab Care *11*, 393.

Pravenec, M., Churchill, P., Churchill, M., Viklicky, O., Kazdova, L., Aitman, T., Petretto, E., Hubner, N., Wallace, C., and Zimdahl, H. (2008). Identification of renal Cd36 as a determinant of blood pressure and risk for hypertension. Nature Genetics *40*, 952-954.

Pravenec, M., Gauguier, D., Schott, J.J., Buard, J., Kren, V., Bila, V., Szpirer, C., Szpirer, J., Wang, J.M., Huang, H., *et al.* (1995). Mapping of quantitative trait loci for blood pressure and cardiac mass in the rat by genome scanning of recombinant inbred strains. J Clin Invest *96*, 1973-1978.

Pravenec, M., Klir, P., Kren, V., Zicha, J., and Kunes, J. (1989). An analysis of spontaneous hypertension in spontaneously hypertensive rats by means of new recombinant inbred strains. J Hypertens *7*, 217-221.

Pravenec, M., Zídek, V., Musilová, A., Simáková, M., Kostka, V., Mlejnek, P., Kren, V., Krenova, D., Bílá, V., Míková, B., *et al.* (2002). Genetic analysis of metabolic defects in the spontaneously hypertensive rat. Mammalian Genome *13*, 253-258.

Reaven, G.M. (1988). Banting lecture 1988. Role of insulin resistance in human disease. Diabetes *37*, 1595-1607.

Rettig, R., and Grisk, O. (2005). The Kidney as a Determinant of Genetic Hypertension. Hypertension *46*, 463-468.

Reynisdottir, S., Ellerfeldt, K., Wahrenberg, H., Lithell, H., and Arner, P. (1994). Multiple lipolysis defects in the insulin resistance (metabolic) syndrome. Journal of Clinical Investigation *93*, 2590.

Risch, N., and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. Science *273*, 1516-1517.

Rockman, M., and Kruglyak, L. (2006). Genetics of global gene expression. Nature Reviews Genetics *7*, 862-872.

Rockman, M.V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. Nature *456*, 738-744.

Ruden, D. (2007). Personalized medicine and quantitative trait transcripts. Nature Genetics *39*, 144-145.

Sarwar, R., and Cook, S. (2009). Genomic Analysis of Left Ventricular Remodeling. Circulation *120*, 437-444.

Schadt, E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218-223.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet *37*, 710-717.

Schliekelman, P. (2008). Statistical Power of Expression Quantitative Trait Loci for Mapping of Complex Trait Loci in Natural Populations. Genetics *178*, 2201.

Schmidlin, O., Tanaka, M., Bollen, A.W., Yi, S.L., and Morris, R.C. (2005). Chloride-dominant salt sensitivity in the stroke-prone spontaneously hypertensive rat (Am Heart Assoc), pp. 867-873.

Schocken, D.D., Benjamin, E.J., Fonarow, G.C., Krumholz, H.M., Levy, D., Mensah, G.A., Narula, J., Shor, E.S., Young, J.B., and Hong, Y. (2008). Prevention of Heart Failure: A Scientific Statement From the American Heart Association Councils on Epidemiology and Prevention, Clinical Cardiology, Cardiovascular Nursing, and

High Blood Pressure Research; Quality of Care and Outcomes Research Interdisciplinary Working Group; and Functional Genomics and Translational Biology Interdisciplinary Working Group. Circulation *117*, 2544-2565.

Sieberts, S.K., and Schadt, E.E. (2007). Moving toward a system genetics view of disease. Mamm Genome *18*, 389-401.

Simon, S.A., Zhai, J., Nandety, R.S., McCormick, K.P., Zeng, J., Mejia, D., and Meyers, B.C. (2009). Short-Read Sequencing Technologies for Transcriptional Analyses. Annual review of plant biology *60*, 305-333.

Sladek, R., and Hudson, T.J. (2006). Elucidating cis- and trans-regulatory variation using genetical genomics. Trends Genet *22*, 245-250.

Soler, J.M., Pereira, A.C., Torres, C.H., and Krieger, J.E. (2006). Gene by environment QTL mapping through multiple trait analyses in blood pressure salt-sensitivity: identification of a novel QTL in rat chromosome 5. BMC Med Genet *7*, 47.

STAR Consortium (2008). SNP and haplotype mapping for genetic analysis in the rat. Nat Genet *40*, 560-566.

Stephens, M., and Balding, D. (2009). Bayesian statistical methods for genetic association studies. Nature Reviews Genetics *10*, 681-690.

Storey, J.D. (2002). A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society Series B (Statistical Methodology) *64*, 479-498.

Storey, J.D. (2003a). The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. The Annals of Statistics *31*, 2013-2035.

Storey, J.D. (2003b). Q-value.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.

Stumvoll, M., Goldstein, B.J., and van Haeften, T.W. (2005). Type 2 diabetes: principles of pathogenesis and therapy. The Lancet *365*, 1333-1346.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sugarbaker, D.J., Richards, W.G., Gordon, G.J., Dong, L., De Rienzo, A., Maulik, G., Glickman, J.N., Chirieac, L.R., Hartman, M.L., and Taillon, B.E. (2008). Transcriptome sequencing of malignant pleural mesothelioma tumors. Proceedings of the National Academy of Sciences *105*, 3521.

Sun, W., Yu, T., and Li, K.-C. (2007). Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics *23*, 2290-2297.

Tu, Z., Wang, L., Arbeitman, M.N., Chen, T., and Sun, F. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics *22*, e489-496.

Turner, D., Keane, T., Sudbery, I., and Adams, D. (2009). Next-generation sequencing of vertebrate experimental organisms. Mammalian Genome *20*, 327-338.

Twigger, S.N., Pruitt, K.D., Fernandez-Suarez, X.M., Karolchik, D., Worley, K.C., Maglott, D.R., Brown, G., Weinstock, G., Gibbs, R.A., Kent, J., *et al.* (2008). What everybody should know about the rat genome and its online resources. Nat Genet *40*, 523-527.

Van Zant, G., and de Haan, G. (2004). Genetic control of lifespan: studies from animal models. Expert Reviews in Molecular Medicine *1*, 1-12.

Vendruscolo, L.F., Terenina-Rigaldie, E., Raba, F., Ramos, A., Takahashi, R.N., and Mormede, P. (2006). A QTL on rat chromosome 7 modulates prepulse inhibition, a neuro-behavioral trait of ADHD, in a Lewis x SHR intercross. Behav Brain Funct *2*, 21.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet *9*, 255-266.

Vogel, C.I.G., Greene, B., Scherag, A., Muller, T.D., Friedel, S., Grallert, H., Heid, I.M., Illig, T., Wichmann, H., and Schafer, H. (2009). Non-replication of an association of CTNNBL 1 polymorphisms and obesity in a population of Central European ancestry. BMC Medical Genetics *10*, 14.

Wagner, R.A., Tabibiazar, R., Powers, J., Bernstein, D., and Quertermous, T. (2004). Genome-wide expression profiling of a cardiac pressure overload model identifies major metabolic and signaling pathway responses. Journal of Molecular and Cellular Cardiology *37*, 1159-1170.

Wang, S.S., Schadt, E.E., Wang, H., Wang, X., Ingram-Drake, L., Shi, W., Drake, T.A., and Lusis, A.J. (2007). Identification of Pathways for Atherosclerosis in Mice: Integration of Quantitative Trait Locus Analysis and Global Gene Expression Data. Circ Res *101*, e11-30.

Wang, X., Ishimori, N., Korstanje, R., Rollins, J., and Paigen, B. (2005). Identifying novel genes for atherosclerosis through mouse-human comparative genetics. Am J Hum Genet *77*, 1-15.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

Wentzell, A.M., Rowe, H.C., Hansen, B.G., Ticconi, C., Halkier, B.A., and Kliebenstein, D.J. (2007). Linking Metabolic QTL with Network and cis-eQTL Controlling Biosynthetic Pathways. PLoS Genetics *3*, e162.

West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W., and St. Clair, D.A. (2007). Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. Genetics *175*, 1441-1450.

Whitehead, A., and Crawford, D.L. (2006). Variation within and among species in gene expression: raw material for evolution. Mol Ecol *15*, 1197-1211.

Wilkins, B.J., Dai, Y.-S., Bueno, O.F., Parsons, S.A., Xu, J., Plank, D.M., Jones, F., Kimball, T.R., and Molkentin, J.D. (2004). Calcineurin/NFAT Coupling Participates in Pathological, but not Physiological, Cardiac Hypertrophy. Circ Res *94*, 110-118.

Wu, C., Delano, D.L., Mitro, N., Su, S.V., Janes, J., McClurg, P., Batalov, S., Welch, G.L., Zhang, J., and Orth, A.P. (2008). Gene Set Enrichment in eQTL Data Identifies Novel Annotations and Pathway Regulators. PLoS Genetics *4*, e1000070.

Wu, S., Lusis, A., and Drake, T. (2009). A systems-based framework for understanding complex metabolic and cardiovascular disorders. The Journal of Lipid Research *50*, S358.

Xiao, Y. (2004). Qvalue methods may not always control false discovery rate in genomic applications. Paper presented at: Computational Systems Bioinformatics Conference, 2004 CSB 2004 Proceedings 2004 IEEE.

Yagil, Y., and Yagil, C. (2006). Integration--a key to success in the genetic dissection of complex diseases? Trends Cardiovasc Med *16*, 35-38.

Yang, X., Deignan, J., Qi, H., Zhu, J., Qian, S., Zhong, J., Torosyan, G., Majid, S., Falkard, B., and Kleinhanz, R. (2009). Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nature Genetics *41*, 415-423.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A., and Lusis, A.J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res *16*, 995-1004.

Young, I.T. (1977). Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. Journal of histochemistry and cytochemistry *25*, 935.

Young, W.F. (2007). Adrenal causes of hypertension: pheochromocytoma and primary aldosteronism. Reviews in Endocrine & Metabolic Disorders *8*, 309-320.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet *35*, 57-64.

Zhang, K., Rao, F., Wen, G., Salem, R.M., Vaingankar, S., Mahata, M., Mahapatra, N.R., Lillie, E.O., Cadman, P.E., and Friese, R.S. (2006). Catecholamine storage vesicles and the metabolic syndrome: the role of the chromogranin A fragment pancreastatin. Diabetes, Obesity & Metabolism *8*, 621.

Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B., *et al.* (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res *105*, 363-374.

Zhu, X., and Cooper, R.S. (2007). Admixture Mapping Provides Evidence of Association of the VNN1 Gene with Hypertension. PLoS ONE *2*, e1244.

Appendix I MATLAB script used to find empirical *p*-values of correlation in QTT analysis

This MatLab script was integrated into a Perl pipeline:

```
% matlab script for cis vs trans correlation analysis
clear;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%load data
[cisdata, cistrait, cisline] = tblread('DBP_physiol_trait_dat',
'\t');
[transdata, transtrait, transline] = tblread('CIS_heart_dat','\t');
ntrans = size(transdata,2); %ntrans was ntraits.
ncis = size(cisdata,2); %new line.
ntotal = ntrans*ncis;
ntraits = sqrt(ntotal); %ntraits in single-type analyses is total
trans. These are then correlated with each other for ntrans*ntrans
correlations. Here ntrans*ncis correlations are carried out.
[trans_r,trans_c] = size(transdata);     %%%%%% trans_r is no. of
trans strains, trans_c is no. of trans traits. Was simply r and c.
[cis_r,cis_c] = size(cisdata);            %%%%%% cis_r = cis
strains, cis_c = cis traits. New line based on above.

cc = zeros(trans_c,cis_c); % Initiate correlation results matrix.
p = zeros(trans_c,cis_c); % Initiate correlation p-values matrix.
p1t = zeros(trans_c,cis_c); % Initiate one-tailed correlation p-
values matrix. % p one t not p ell t!
emp = zeros(trans_c,cis_c); % Initiate empirical p-values matrix.


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% correlations

for i = [1:ntrans];
  corrtesttrans = transdata(:,i);
  corrtesttemp(:,1) = corrtesttrans;

  for j = [1:ncis];
    corrtestcis = cisdata(:,j);
    corrtesttemp(:,2) = corrtestcis;
    [cc1,pvals] = corrcoef(corrtesttemp);
    currcorrcoef = cc1(1,2);
    currp = pvals(1,2);
    cc(i,j) = currcorrcoef;
    p(i,j) = currp;
    onetailed = currp / 2;
    p1t(i,j) = onetailed;
    clear corrtestcis currcorrcoef currp cc1 pvals;
  end;
  clear corrtesttemp corrtesttrans;
end;
clear i j;
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% set up parameters for
perms.
nperms = 10000;
combinations = ntrans;
alpha = 0.05;
combination_threshold = round(nperms - ((alpha/(combinations/2)) *
nperms));

cctosort = zeros(trans_c,cis_c,nperms); % Initiatialise 3D
correlation results matrix.
for k = [1:nperms];
    for m = [1:trans_c];
     permdata(:,m) = transdata(randperm(trans_r),m);
    end;
    for i = [1:ntrans];
     nullcorrtesttrans = permdata(:,i);
     nullcorrtesttemp(:,1) = nullcorrtesttrans;

       for j = [1:ncis];
        nullcorrtestcis = cisdata(:,j);
        nullcorrtesttemp(:,2) = nullcorrtestcis;
        cc1 = corrcoef(nullcorrtesttemp);
        nullcurrcorrcoef = cc1(1,2);
        cctosort(i,j,k) = abs(nullcurrcorrcoef);
       end;
    end;
  end;

   ccsort = sort((cctosort),3);
   thresh = squeeze(ccsort(:,:,combination_threshold));
   abstest = abs(cc);
   signif = abs(cc) > thresh;

[i j] = find(signif);

ncorrcoef = size(cc,1);
nullcorrcoeftosort = zeros(nperms,1);

for i = [1:ncorrcoef];
  empi = 0;
  nullcorrcoeftosort = ccsort(i,1,:);
  nullcorrcoefsorted = sort(nullcorrcoeftosort,'descend');
  currcorrcoef = abs(cc(i,1));
  for j = [1:nperms];
    if nullcorrcoeftosort(:,:,j) >= currcorrcoef
      empi = empi + 1;
    end;

    if empi == 0;
      empi = 1;
    end;

    empp = empi / nperms;
  end;
```

```
  emp(i,:) = empp;
end;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% output results

dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', transtrait,
'delimiter', '','coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', cistrait, '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', cc, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', printp, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', thresh, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', signif, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', emp, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', p1t, '-append',
'delimiter', '    ', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out', '***', '-append',
'delimiter', '', 'coffset', 0, 'roffset', 1)
dlmwrite('DBP_physiol_trait_dat_CIS_heart.out',
combination_threshold, '-append', 'delimiter', '      ', 'coffset',
0, 'roffset', 1)

close all

clear ans tick r c k tick i j k nperms

quit
```

Contents of exemplar physiological trait data file provided as input to MatLab correlation script:

```
      DBP
1H    84.79
2H    88.47
3H    96.34
4H    83.25
5H    87.09
7H    84.23
10H   94.64
15H   84.92
17H   100.34
18H   88.36
20H   80.65
21H   88.75
22H   89.37
23H   86.19
24H   77.13
25H   81.34
27H   84.7
29H   83.54
31H   83.47
2B    83.18
3B    79.87
5B    84.86
6B    90.77
8B    86.08
9B    79.22
10B   89.81
11B   89.02
12B   84.91
13B   86.84
```

Extract from contents of cis-eQTL expression data file provided as input to MatLab correlation script:

	1367551_a_at	1367562_at	1367574_at	1367628_at
"1H"	7.67973709	10.93780422	10.90817738	10.99627399
"2H"	7.1980772	10.91172218	11.12731838	10.92982292
"3H"	7.21639681	10.63175392	11.10826874	10.93883705
"4H"	7.27095747	11.49904156	11.09344959	10.89237213
"5H"	7.3797121	11.40389347	11.3356514	11.1481514
"7H"	7.71912003	11.84435558	11.59171867	10.90641594
"10H"	7.55096245	10.90875053	11.30874825	11.26289845
"15H"	7.58139229	10.16538239	10.66611767	10.33894157
"17H"	7.28413916	10.78122997	11.26590443	11.11721992
"18H"	7.22759676	10.85401917	11.42521286	11.07767487
"20H"	7.66844988	10.96870708	11.24119377	11.06907177
"21H"	7.38165379	10.92046738	10.77867508	10.61616993
"22H"	7.77224445	11.24554157	11.04502678	11.24810982
"23H"	7.18954372	10.68624783	11.2198925	10.92363071
"24H"	7.11914206	10.77099705	11.15253925	11.17246342
"25H"	7.5004549	11.62852097	11.27778912	10.80047417
"27H"	7.62077045	11.53417587	11.18069267	11.23991013
"29H"	7.2272234	11.30589294	11.20408058	10.79101849
"31H"	7.79323292	10.67498493	11.11670303	11.01991272
"2B"	7.73587513	11.55472469	11.2713747	11.25311947
"3B"	7.57724667	10.66844463	11.2103014	10.91379261
"5B"	7.12386179	11.00831032	11.15570927	10.92574596
"6B"	7.53646517	10.90838242	11.22941494	11.33149529
"8B"	7.78968382	11.4825716	11.58284187	11.2701416
"9B"	7.88862085	10.76527596	11.06812954	10.79598141
"10B"	7.42299223	10.85657406	11.56780624	10.94619942
"11B"	7.37523937	10.92200661	11.30420399	11.31643295
"12B"	7.20876074	10.68705273	11.10902119	10.69730568
"13B"	7.63697529	11.13474655	11.63410568	11.3121376

Appendix II Empirical *p*-values corresponding to numbers of matches in correlation of marker genotypes by simple matching coefficient

Matches	**29**	**28**	**27**	**26**	**25**	**24**	**23**	**22**	**21**	**20**	**19**	**18**	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**10**
0	1.9E-09	3.7E-09	7.5E-09	1.5E-08	3.0E-08	6.0E-08	1.2E-07	2.4E-07	4.8E-07	9.5E-07	1.9E-06	3.8E-06	7.6E-06	1.5E-05	3.1E-05	6.1E-05	1.2E-04	2.4E-04	4.9E-04	9.8E-04
1	5.4E-08	1.0E-07	2.0E-07	3.9E-07	7.5E-07	1.4E-06	2.7E-06	5.2E-06	1.0E-05	1.9E-05	3.6E-05	6.9E-05	1.3E-04	2.4E-04	4.6E-04	8.5E-04	0.002	0.003	0.005	0.010
2	7.6E-07	1.4E-06	2.6E-06	4.8E-06	8.9E-06	1.6E-05	3.0E-05	5.5E-05	1.0E-04	1.8E-04	3.3E-04	5.8E-04	0.001	0.002	0.003	0.006	0.010	0.016	0.027	0.044
3	6.8E-06	1.2E-05	2.2E-05	3.9E-05	6.9E-05	1.2E-04	2.1E-04	3.7E-04	6.3E-04	0.001	0.002	0.003	0.005	0.009	0.014	0.022	0.035	0.054	0.081	0.117
4	4.4E-05	7.6E-05	1.3E-04	2.2E-04	3.8E-04	6.3E-04	0.001	0.002	0.003	0.005	0.007	0.012	0.018	0.028	0.042	0.061	0.087	0.121	0.161	0.205
5	2.2E-04	3.7E-04	6.0E-04	9.8E-04	0.002	0.003	0.004	0.006	0.010	0.015	0.022	0.033	0.047	0.067	0.092	0.122	0.157	0.193	0.226	0.246
6	8.8E-04	0.001	0.002	0.003	0.005	0.008	0.012	0.018	0.026	0.037	0.052	0.071	0.094	0.122	0.153	0.183	0.209	0.226	0.226	0.205
7	0.003	0.004	0.007	0.010	0.014	0.021	0.029	0.041	0.055	0.074	0.096	0.121	0.148	0.175	0.196	0.209	0.209	0.193	0.161	0.117
8	0.008	0.012	0.017	0.023	0.032	0.044	0.058	0.076	0.097	0.120	0.144	0.167	0.185	0.196	0.196	0.183	0.157	0.121	0.081	0.044
9	0.019	0.026	0.035	0.047	0.061	0.078	0.097	0.119	0.140	0.160	0.176	0.185	0.185	0.175	0.153	0.122	0.087	0.054	0.027	0.010
10	0.037	0.049	0.063	0.079	0.097	0.117	0.136	0.154	0.168	0.176	0.176	0.167	0.148	0.122	0.092	0.061	0.035	0.016	0.005	9.8E-04
11	0.064	0.080	0.097	0.115	0.133	0.149	0.161	0.168	0.168	0.160	0.144	0.121	0.094	0.067	0.042	0.022	0.010	0.003	4.9E-04	
12	0.097	0.113	0.130	0.144	0.155	0.161	0.161	0.154	0.140	0.120	0.096	0.071	0.047	0.028	0.014	0.006	0.002	2.4E-04		
13	0.126	0.139	0.149	0.155	0.155	0.149	0.136	0.119	0.097	0.074	0.052	0.033	0.018	0.009	0.003	8.5E-04	1.2E-04			
14	0.144	0.149	0.149	0.144	0.133	0.117	0.097	0.076	0.055	0.037	0.022	0.012	0.005	0.002	4.6E-04	6.1E-05				
15	0.144	0.139	0.130	0.115	0.097	0.078	0.058	0.041	0.026	0.015	0.007	0.003	0.001	2.4E-04	3.1E-05					
16	0.126	0.113	0.097	0.079	0.061	0.044	0.029	0.018	0.010	0.005	0.002	5.8E-04	1.3E-04	1.5E-05						
17	0.097	0.080	0.063	0.047	0.032	0.021	0.012	0.006	0.003	0.001	3.3E-04	6.9E-05	7.6E-06							
18	0.064	0.049	0.035	0.023	0.014	0.008	0.004	0.002	6.3E-04	1.8E-04	3.6E-05	3.8E-06								
19	0.037	0.026	0.017	0.010	0.005	0.003	0.001	3.7E-04	1.0E-04	1.9E-05	1.9E-06									
20	0.019	0.012	0.007	0.003	0.002	6.3E-04	2.1E-04	5.5E-05	1.0E-05	9.5E-07										
21	0.008	0.004	0.002	9.8E-04	3.8E-04	1.2E-04	3.0E-05	5.2E-06	4.8E-07											
22	0.003	0.001	6.0E-04	2.2E-04	6.9E-05	1.6E-05	2.7E-06	2.4E-07												
23	8.8E-04	3.7E-04	1.3E-04	3.9E-05	8.9E-06	1.4E-06	1.2E-07													
24	2.2E-04	7.6E-05	2.2E-05	4.8E-06	7.5E-07	6.0E-08														
25	4.4E-05	1.2E-05	2.6E-06	3.9E-07	3.0E-08															
26	6.8E-06	1.4E-06	2.0E-07	1.5E-08																
27	7.6E-07	1.0E-07	7.5E-09																	
28	5.4E-08	3.7E-09																		
29	1.9E-09																			

Appendix III Sizes of *trans*-eQTL clusters at $p < 0.05$ and $p < 0.01$, and percentage significant co-expression levels in all *trans*-eQTL clusters

Tissue	**Marker at *trans*-eQTL cluster peak of linkage**	**No. Transcripts in Cluster ($p < 0.05$)**	**No. *trans*-eQTLs in Cluster ($p < 0.01$)**	**% Significant *trans-trans* expression correlation**
Fat	Cacna1s	33	5	96.8
Fat	D11Rat7	11	0	85.5
Fat	D12Ntr2	19	2	94.7
Fat	D14Rat52	14	2	100
Fat	D16Cebr204s40	20	5	87.4
Fat	D16Mit1	12	5	87.9
Fat	**D17Rat1**	**146**	**26**	**96.8**
Fat	D1Rat27	19	5	87.7
Fat	D1Rat7	13	2	97.4
Fat	D4Rat240	31	8	84.7
Fat	D8Utr5	21	9	73.8
Kidney	Abpa	12	7	81.8
Kidney	Crabp1	18	4	96.1
Kidney	Cyp45c	23	3	83.6
Kidney	D10Cebrp1016s2	15	4	85.7
Kidney	D10Cebrp207s1	22	1	81.8
Kidney	D15Rat29	11	2	72.7
Kidney	**D15Rat69**	**57**	**11**	**91.2**
Kidney	D16Cebr204s40	16	7	67.5
Kidney	D16Mit2	14	2	79.1
Kidney	D2Cebr11s4	13	6	79.5
Kidney	D3Cebrp1038s1	14	5	74.7
Kidney	D4Mit11	13	2	88.5
Kidney	D4Rat35	24	7	87
Kidney	D4Utr4	12	4	81.8
Kidney	D5Rat174	27	3	90.6
Kidney	D5Rat38	10	3	95.6
Kidney	D8Rat_42	17	4	81.6

Kidney	D8Rat150	11	1	89.1
Kidney	D8Rat21	12	6	90.9
Kidney	D8Utr5	16	6	72.5
Kidney	Igk@	49	8	89.6
Kidney	Scnb2	20	2	90.5
Kidney	Slc12a1	22	4	86.1
Adrenal	Abpa	10	5	75.6
Adrenal	D11Rat16	31	9	75.5
Adrenal	D15Rat29	11	0	100
Adrenal	**D17Rat144**	**47**	**16**	**88.6**
Adrenal	D1Utr6	11	1	74.5
Adrenal	D20Mit1	11	4	78.2
Adrenal	D20Rat55	16	6	80.8
Adrenal	D8Rat56	14	3	83.5
Adrenal	D8Utr3	20	4	85.3
LV	Abpa	15	7	73.3
LV	Ckb	43	4	79.2
LV	**Crabp1**	**165**	**47**	**77.1**
LV	Cyp45c	35	8	87.9
LV	D11Cebr11s6	11	0	67.3
LV	D11Mit4	10	2	80
LV	D13Cebr9s2	27	7	86.9
LV	D13Cebr9s3	12	6	72.7
LV	D13Utr6	13	1	66.7
LV	D15Rat123	11	3	74.5
LV	D15Rat29	54	9	99
LV	D15Rat98	30	3	88
LV	D15Ucsf1	46	8	92.9
LV	D15Utr2	51	9	85.5
LV	D16Cebr204s40	15	2	74.3
LV	D16Mit3	13	2	82.1
LV	D16Rat46	28	7	92.1
LV	D16Rat67	26	4	72.9
LV	D17Mit6	20	8	78.9

LV	D17Rat17	10	3	66.7
LV	D1Cebrp37s18	10	4	60
LV	D20Arb249	11	0	54.5
LV	D3Cebr37s29	10	3	80
LV	D3Mit16	14	2	67
LV	D4Rat140	10	2	82.2
LV	D6Mit10	14	3	93.4
LV	D6Rat79	20	3	77.9
LV	D6Rat80	15	1	84.8
LV	D8Cebr81s4	14	3	64.8
LV	D8Mgh4	10	2	68.9
LV	**D8Mit12**	**77**	**27**	**68.8**
LV	D8Rat219	21	3	83.3
LV	D8Rat49	11	4	63.6
LV	D8Utr2	10	1	82.2
LV	**D8Utr5**	**22**	**11**	**72.3**
LV	Edn3	10	0	68.9
LV	Kcnj1	18	2	82.4
LV	Tpm1	23	9	79.1

Appendix IV Distinct DAVID significant ($p < 0.01$) Biological Process and Molecular Function GO terms, and numbers of genes annotated with each term, in large *trans*-eQTL clusters

***Trans*-eQTL cluster peak of linkage**	**Tissue**	**Significantly enriched GO terms ($p < 0.01$)**	**No. (%) genes enriched for GO term**
D11Rat16	Adrenal	Carboxylic acid metabolism	6 (20.7%)
		Organic acid metabolism	6 (20.7%)
D17Rat144	Adrenal	Metabolism	20 (45.5%)
Cacna1s	Fat	Cellular metabolism	16 (50%)
D4Rat240	Fat	Extracellular matrix structural constituent	5 (17.9%)
		Skeletal development	5 (17.9%)
		Organ development	7 (25%)
		Structural molecule activity	5 (17.9%)
D17Rat1	Fat	Calcium ion binding	15 (11.5%)
		Alpha-type channel activity	11 (8.5%)
		Channel or pore class transporter activity	11 (8.5%)
		Alkali metal ion binding	7 (5.4%)
		Morphogenesis	17 (13.1%)
		Ion binding	27 (20.8%)
		Metal ion binding	27 (20.8%)
		Metal ion transport	9 (6.9%)
		Sexual reproduction	8 (6.2%)
		Reproduction	9 (6.9%)
		Synaptic transmission	9 (6.9%)
		Cation transport	10 (7.7%)
		Transporter activity	22 (16.9%)

		Ion transport	12 (9.2%)
		Gametogenesis	7 (5.4%)
		Enzyme inhibitor activity	7 (5.4%)
		Cell-cell signalling	12 (9.2%)
		Transport	28 (21.5%)
		Establishment of localization	31 (23.9%)
		Transmission of nerve impulse	9 (6.9%)
D15Rat69	Kidney	Establishment of localization	22 (39.3%)
		Localization	22 (39.3%)
		Transport	20 (35.7%)
		Intracellular transport	10 (17.9%)
		Establishment of cellular localization	10 (17.9%)
		Cellular localization	10 (17.9%)
		Protein transport	8 (14.3%)
		Establishment of protein localization	8 (14.3%)
		Intracellular protein transport	7 (12.5%)
		Protein localization	8 (14.3%)
		Vesicle-mediated transport	7 (12.5%)
		Secretion	6 (10.7%)
Igk@	Kidney	Synaptic transmission	6 (14%)
		Transmission of nerve impulse	6 (14%)
		Cell-cell signalling	7 (16.3%)
Ckb	Left Ventricle	Establishment of localization	12 (34.3%)
		Localization	12 (34.3%)
		Cell organization & biogenesis	9 (25.7%)
		Transport	9 (25.7%)

Crabp1	Left Ventricle	Transporter activity	19 (14.2%)
		Establishment of localization	27 (20.2%)
		Localization	27 (20.2%)
		Transport	24 (17.9%)
Cyp45c	Left Ventricle	*None*	*None*
D15Rat29	Left Ventricle	Response to biotic stimulus	13 (27.7%)
		Defense response	12 (25.5%)
		Immune response	11 (23.4%)
		Modification-dependent protein catabolism	5 (10.6%)
		Proteolysis during cellular protein catabolism	5 (10.6%)
		Cellular protein catabolism	5 (10.6%)
		Protein catabolism	5 (10.6%)
		Biopolymer catabolism	5 (10.6%)
		Cellular macromolecule catabolism	5 (10.6%)
		Protein metabolism	13 (27.7%)
		Macromolecule catabolism	5 (10.6%)
		Proteolysis	6 (12.8%)
		Cellular protein metabolism	12 (25.5%)
		Cellular macromolecule catabolism	12 (25.5%)
		DNA binding	9 (19.2%)
D15Rat98	Left Ventricle	*None*	*None*
D15Ucsf1	Left Ventricle	Nucleotide binding	9 (22%)
		Regulation of biological process	12 (29.3%)
D15Utr2	Left Ventricle	Protein targeting	5 (12.8%)
		Cell organization and biogenesis	10 (25.6%)

		Protein transport	6 (15.4%)
		Establishment of protein localization	6 (15.4%)
		Protein localization	6 (15.4%)
		Intracellular transport	6 (15.4%)
		Establishment of cellular localization	6 (15.4%)
		Cellular localization	6 (15.4%)
		Intracellular protein transport	5 (12.8%)
		Protein import into nucleus	3 (7.7%)
		Nuclear import	3 (7.7%)
		Protein import	3 (7.7%)
		Organelle organization and biogenesis	6 (15.4%)
D8Mit12	Left Ventricle	*None*	

Appendix V Description of metabolic traits on which systematic QTT analysis was carried out

Physiological Trait	Details
ins=0/cell	Basal glucose uptake by adipocytes
absnefa/cell	Non-essential fatty acids after isoproterenol induced lipolysis, corrected to cell number
ins=50k-0/cell	Difference of stimulated and basal glucose uptake
fold	Increase of glucose uptake by adipocytes
ins50k/cell	Insulin stimulated glucose uptake by adipocytes
absnefa	Non-essential fatty acids after isoproterenol induced lipolysis
ratwt	Rat weight
cell vol	Volume of adipocytes
ins0	Serum insulin concentrations at the age of 8 weeks in male rats fed a normal diet
ins15	Serum insulin concentrations at the age of 10 weeks in male rats fed a diet with 60% fructose for 15 days
glu0	Serum glucose concentrations at the age of 8 weeks in male rats fed a normal diet
glu15	Serum glucose concentrations at the age of 10 weeks in male rats fed a diet with 60% fructose for 15 days
ins/glu0	Insulin/glucose ratio at the age of 8 weeks in male rats fed a normal diet
ins/glu15	Insulin/glucose ratio at the age of 10 weeks in male rats fed a diet with 60% fructose for 15 days
IPGTT0	Serum glucose concentrations during intraperitoneal glucose tolerance test at 0 min after glucose loading performed in male rats fed a diet with 60% fructose for 13 days
IPGTT30	Serum glucose concentrations during intraperitoneal glucose tolerance test at 30 mins after glucose loading performed in male rats fed a diet with 60% fructose for 13 days
IPGTT60	Serum glucose concentrations during intraperitoneal glucose tolerance test at 60 mins after glucose loading performed in male rats fed a diet with 60% fructose for 13 days
IPGTT120	Serum glucose concentrations during intraperitoneal glucose tolerance test at 120 min after glucose loading performed in male rats fed a diet with 60% fructose for 13 days
ch0	Serum cholesterol concentrations at the age of 8 weeks in male rats fed a normal diet
ch7	Serum cholesterol concentrations in male rats fed a diet with 60% fructose for 7 days
ch15	Serum cholesterol concentrations in male rats fed a diet with

	60% fructose for 15 days
tg0	Serum triglyceride concentrations at the age of 8 weeks in male rats fed a normal diet
tg7	Serum triglyceride concentrations in male rats fed a diet with 60% fructose for 7 days
tg15	Serum triglyceride concentrations in male rats fed a diet with 60% fructose for 15 days
HDL15	Serum HDL cholesterol concentrations in male rats fed a diet with 60% fructose for 15 days
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
PP	Pulse Pressure
MAP	Mean Arterial Pressure
RLVHWT	Relative Left Ventricular Heart Weight
kidneydif	Difference between remnant kidney at the age of 8 weeks and a kidney at the age of 5 weeks (measure of compensatory renal hypertrophy)
kidney5wk	Relative weight of a kidney removed at the age of 5 weeks
kidney8wk	Relative weight of a remnant kidney removed at the age of 8 weeks
livchol	Liver cholesterol concentrations in rats fed a high fat, high cholesterol diet from the age of 7-11 weeks
liver(g)	Liver weight
livpool	Cholesterol content of the liver
offspring	Number of young per litter
IDLch0	IDL cholesterol in rats at the age of 7 weeks in rats fed a normal diet
IDLch4	IDL cholesterol in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
IDLph0	IDL phospholipids in rats at the age of 7 weeks in rats fed a normal diet
IDLph4	IDL phospholipids in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
LDLch0	LDL cholesterol in rats at the age of 7 weeks in rats fed a normal diet
LDLch4	LDL cholesterol in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
LDLph0	LDL phospholipids in rats at the age of 7 weeks in rats fed a normal diet
LDLph4	LDL phospholipids in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks

HDL2ch0	HDL2 cholesterol in rats at the age of 7 weeks in rats fed a normal diet
HDL2ch4	HDL2 cholesterol in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
HDL2ph0	HDL2 phospholipids in rats at the age of 7 weeks in rats fed a normal diet
HDL2ph4	HDL2 phospholipids in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
HDL3ch0	HDL3 cholesterol in rats at the age of 7 weeks in rats fed a normal diet
HDL3ch4	HDL3 cholesterol in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
HDL3ph0	HDL3 phospholipids in rats at the age of 7 weeks in rats fed a normal diet
HDL3ph4	HDL3 phospholipids in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
totchol0	Total serum cholesterol in rats at the age of 7 weeks in rats fed a normal diet
totphos0	Total serum phospholipids in rats at the age of 7 weeks in rats fed a normal diet
totchol4	Total serum cholesterol in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
totphos4	Total serum phospholipids in rats at the age of 11 weeks in rats fed a high fat, high cholesterol diet for 4 weeks
ins0/vol	Basal glucose uptake by isolated adipocytes corrected to cell volume
ins50/vol	Insulin stimulated uptake of glucose by isolated adipocytes corrected to cell volume
delins/vol	Difference between ins50/vol and ins0/vol
absnefa/vol	Non-essential fatty acids after isoproterenol induced lipolysis, corrected to cell number
hematocrit90	Hematocrit (proportion of blood volume occupied by red blood cells) at the age of 90 days
platelet aggrega	Platelet aggregation
CORT-20	Serum corticosteron levels after 20mins of immobilization stress
CORT-40	Serum corticosteron levels after 40mins of immobilization stress
CORT-80	Serum corticosteron levels after 80mins of immobilization stress
CORT-140	Serum corticosteron levels after 140mins of immobilization stress

ACTH20	Serum ACTH (corticotropin) levels after 20mins of immobilization stress
NA/h/kg	Urine noradrenaline, corrected for body weight
A/h/kg	Urine adrenaline, corrected for body weight
DOP/h/kg	Urine dopamine, corrected for body weight
CTA	Conditioned taste aversion
CUPOOL	Liver copper concentration after rats were fed a high fat, high cholesterol diet for 4 weeks
CUDW	Dry weight of liver copper
CUWW	Wet weight of liver copper
NO	Serum nitric oxide in male rats fed a normal diet
Padwt/BW	Weight of the fat pad relative to body weight
AUC	Area under the IPGTT curve
log fold	Log of the increase of glucose uptake by adipocytes