BUILDING A MODEL FOR MAPPING GENETIC VARIATION AFFECTING GENE EXPRESSION

by

Peter Daniel Lee

Department of Human Genetics, McGill University, Montreal

February 2005

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Ph.D.

©Peter Daniel Lee, 2005

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-21668-2
Our file Notre référence
ISBN: 978-0-494-21668-2

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

PREFACE

At the start of this degree, DNA microarrays were still emerging on the scientific stage. Since then much has changed. Initial microarray studies inspired the imagination – we imagined that we were viewing the global machinery of the cell for the first time. Would it be possible to understand the entire gene network? As the technology evolved, this enthusiasm has been followed by the realization that statistical evaluations were needed. This has not replaced the questions raised during the initial phase of inspiration but rather made us aware of how challenging it will be to answer such questions convincingly, even at all. Over the course of this thesis I observed the evolution of a technology, a process whereby multiple disciplines interacted to develop a common vocabulary. Many questions remain. I have addressed a small fraction of the field in my thesis; the scope of this experiment is already vast beyond the imagination. Along the way, a great number of analytical tools were developed and applied. The following chapters cover my own process of evolution in the understanding and application of these techniques to address biological questions.

Along the way I have received help from many sources. I present a partial list of names to acknowledge the individuals who, without their support, this degree would not have been possible: my supervisors Thomas Hudson, Michael Hallett, members of my advisory committee, Marcel Behr, Patricia Tonin; at the McGill University and Genome Quebec Centre for Innovation, Rob Sladek, Jaroslav Novak, Celia Greenwood, Tomi Pastinen, Bing Ge, Bob Nadon, Pierre Lepage, Jamie Engert, Chon Loredo-Osti, Jenny Koulis,

Andrei Verner, Tibor van Rooij, Donna Sinnett, Sebastien Brunet, Vincenzo Forgetta, Janet Faith, Sebastien de Grandpre; members of the McGill Centre for Bioinformatics, Marina Takane, Greg Finak, Francois Pepin, Michelle Scott, Ernesto Iacucci, Trevor Bruen, Rachel Bevan, Kaleigh Smith; members of the MUHC Emil Skamene, Serge Mostowy; the Human Genetics Student Society, in particular, Emily Manderson, Judith Caron, Caroline Gallant and Vanessa Sancho, co-organizers of the Bioinformatics Research Day; the laboratory of Alan Peterson; the Animal Facility at the Montreal General Hospital Research Institute; Laura Benner, Fran Langton at the Dept of Human Genetics; Leon Glass, and Michael Mackey at the Centre for Nonlinear Dynamics; Pablo Tamayo, Todd Golub, Bing Ren, David DeGraffe at the Whitehead/MIT Center for Genome Research and Michael Rebhan at Astra-Zeneca who provided stimulating discussions at the start of this degree; Spyro Mousses at TGEN and Hilmi Ozcelik at the University of Toronto for continued encouragement along the way; Prof Charles J. Lumsden of the University of Toronto who first made me aware of bioinformatics and the importance of imagination in science; Jerome Holmes, Johanne O'Malley and the staff of Thomson House for unwavering their support, the Post Graduate Student Society, an example of graduate student governance beyond compare - Thomson House and the PGSS represent environments for interdisciplinary interaction, applied learning, and are among the highest benefits of pursuing graduate studies at McGill. Thanks to those who read and edited this thesis, Tom Hudson, Rob Sladek, Jamie Engert, Anny Fortin, Celia Greenwood and Pierre Lepage, who provided the French translation of the abstract. Special thanks to Monica Herman for her help in times of need. There are many others, Cathy Neighbor, Paul Shoniker, Julia Shiu, Martha Shiu. I thank you all.

For my sister, Mary-Esther Lee and in memory of my mom, Esther Kuo Wah Lee, my dad, Rev. David Yiu Shan Lee, and my brother, John David Lee.

TABLE OF CONTENTS

CONTRIBUTIONS OF AUTHORS

Mouse Work

A/J and C57BL/6J mice as well as the Recombinant Congenic Panel were provided by Dr Emil Skamene and Dr Anny Fortin. Genotyping data for the RCS panel were provided following a Material Transfer Agreement with Emerillon Therapeutics. Housing, sacrificing and tissue dissections were performed by staff at the MUHC Mouse Facility managed by Jean-Marie Chavannes. RNA extractions were performed by technicians Scott Gurd and Yannick Fortin, Dr Rob Sladek, graduate student Sarav Sundararajan and myself.

Microarray Hybridizations

Microarray hybridizations were performed by Dr Rob Sladek, Yannick Fortin, Daniel Vincent and Arek Siwoski in the Chip Facility at the McGill University and Genome Quebec Innovation Centre.

Bioinformatics and Infrastructure

I handled all programming for analysis of microarray data including parsing of data files using Perl, statistical analysis in SAS and R, incorporation of information from other biological databases, constructing a database for the RCS project and design of web interfaces for analyzing data. Bing Ge provided the positions of Affymetrix oligonucleotides and microsatellites from the RCS project, aligned to the UCSC mouse genome assembly. Tibor van Rooji assisted me in setting up databases and with Perl

programming. Peter Nowacki set up the lab's first server and infrastructure of critical importance to my work from 1999 to 2002. Marina Takane, M.Sc. graduate in the laboratory of Dr Mike Hallett, was involved in the early part of the RCS project in formulating proposals to search for trans-regulatory relationships, and was involved in designing programs to assign DSO to genes. Greg Finak, Ph.D. candidate in the laboratory of Drs Morag Park and Mike Hallett, assisted me with statistics programming in R. Ernesto Iacucci, M.Sc. candidate in the laboratory of Drs Hans Zingg and Mike Hallett, provided advice on the use of the Gene Ontology database. The McGill Centre for Bioinformatics provided critical infrastructure in the form of servers, databases, and computational power. All computationally intensive analyses were performed on this infrastructure from 2002-2005. Francois Pepin, M.Sc. candidate in the laboratory of Dr Mike Hallett, provide support maintaining this information system. All statistical and computationally intense analysis was done on these servers.

Data Analysis and Statistics

I performed all statistical programming and data parsing for all studies in this degree. Methods for trans regulation and statistical analysis were conceptualized in discussions with Dr Mike Hallett and Dr Celia M.T. Greenwood. Dr Bob Nadon provided advice on interpretation of analysis results as well as double-checking the results. Dr Jaroslav Novak provided assistance, guidance and support throughout my degree. Dr Jamie Engert was an invaluable source of inspiration and perspective with respect to systems approaches in biology. Dr Leon Glass and Dr Michael Mackey showed me where it all leads.

Allelic Imbalance and Sequencing

Dr Tomi Pastinen developed the method to detect allelic imbalance and bioinformatician Bing Ge designed the software (PeakPicker1.0) to process the results from these studies. All sequencing was performed in the Sequencing Facility at the McGill University and Genome Quebec Innovation Centre by Donna Sinnett and Sebastien Brunet, research technicians under the supervision of Dr Pierre Lepage. Dr Robert Sladek and I designed the primers for AI analysis. Analysis of sequences and AI results for the pilot project (first 20 genes) was done by Dr Robert Sladek and Donna Sinnett. I completed subsequent analysis of sequence and AI data.

ABBREVIATIONS

AcB: RCS having ~87.5% genome from A/J and 12.5% from C57BL/6J

AffyID: Affymetrix probe set identifier

AI: allelic imbalance

ANOVA: analysis of variance

BcA: RCS having ~87.5% genome from C57BL/6J and 12.5% from A/J

BG: background

bp: base pair

BLAST: Basic local alignment search tool

BLAT: BLAST-like alignment tool

chr: chromosome

cM: centimorgan (~2000kb in mice)

CSS: chromosome substitution strain

CV: coefficient of variation

cDNA: complementary deoxyribonucleic acid

DB: database

DNA: deoxyribonucleic acid

DSO: donor strain of origin

EST: expressed tag sequence

gDNA: genomic deoxyribonucleic acid

GEO: Gene Expression Omnibus

GLM: general linear model

GO: gene ontology

kb: kilobase pair (thousand base pair)

LOD: logarithmic odds ratio

MAS: Affymetrix Microarray Suite

Mb: megabase pair (million base pair)

MGED: microarray gene expression data society (http://www.mged.org)

MFC: maximum fold change

MM: mismatch oligonucleotide (Affymetrix)

mRNA: messenger ribonucleic acid

MVA plot: minus versus average plot

NCBI: National Center for Biotechnology Information

NIH: National Institutes of Health

PCR: polymerase chain reaction

PM: perfect match oligonucleotide (Affymetrix)

QQ-plot: quantile-quantile plot

QTL: quantitative trait locus/loci

RCS: recombinant congenic strains

RMA: robust multichip averaging

RNA: ribonucleic acid

SD: standard deviation

SNP: single nucleotide polymorphism

SSLP: simple sequence length polymorphism

UTR: untranslated region

ABSTRACT (ENGLISH)

The majority of genetic traits including most common diseases are believed to be multigenic and arise both from variations in coding sequences as well as from regulatory polymorphisms. Genome-wide approaches are needed to develop models for understanding this complexity. This thesis develops approaches for studying genetic variation affecting gene expression on a genome-wide scale. This included development of experimental design principles and analytical methods for microarray data. These principles were then applied to characterize differences between commonly-used A/J and C57BL/6J inbred mouse strains at the molecular level identifying over 2000 genes differentially expressed between these strains across 4 tissues. To further investigate the role of genetic variation in genome-wide expression changes, we analyzed expression profiles of lung tissue obtained from a panel of recombinant congenic strains (RCS) derived from the same inbred strains. An ANOVA was applied using a model to test the association of expression profiles with donor-strain of origin (DSO, inferred from RCS genotyping data), and with genetic background. This model identified over 1500 genes whose expression levels were associated with DSO status ($P<0.05$) having adjusted for the variability due to predominant strain of background, suggestive of cis-regulatory variation in these genes. We randomly selected 50 positive genes displaying association between DSO and 80 negative genes for validation using allelic imbalance (AI), a method that uses intragenic SNPs for detecting genes with cis-regulatory variation that measures allele-specific transcript levels in cDNA of heterozygous individuals. Of the genes chosen, 54% of positive versus

27% of negative genes contained at least one SNP within ≥ 1kbp of 3' UTR sequenced ($P<0.05$ Fisher exact test). AI was found in 63% of positive genes versus 23% of negative genes ($P<0.01$ Fisher exact test) representing a greater than 10-fold enrichment over random screening for the detection of genes with potential cis-acting regulation. The study conservatively estimates 34% potentially cis-regulated genes, similar to other studies in mammalian systems. This study furthermore demonstrates a multidisciplinary approach capable of genome-wide cataloguing of genes subject to cis-acting regulatory variation, an initial step towards mapping complex traits and developing models of gene regulation throughout the mammalian genome.

RÉSUMÉ (FRANÇAIS)

La plupart des traits génétiques dont ceux causant les maladies communes semblent être de nature multigénique. Ces traits découlent de variations dans la séquence codante ou dans les régions régulatrices des gènes. Des approches à l'échelle du génome devront être développées pour comprendre la complexité de ces traits multigéniques. Cette thèse décrit de telles approches pour l'étude des variations de séquences qui peuvent affecter l'expression génique. L'une de ces approches visait à définir des principes de base pour le développement de protocoles expérimentaux reliés aux études faisant usage de puces à ADN et pour l'analyse des résultats qui en découlent. Ces principes de base ont ensuite été appliqués dans une étude visant à identifier les différences d'expression génique entre deux souches pures de souris couramment utilisées, A/J et C57BL/6J. Cette étude a permis d'identifier plus de 2000 gènes dont l'expression varie entre ces deux souches dans 4 tissus différents. Afin d'étudier plus à fond l'influence que les variations génétiques peuvent avoir sur les changements de niveaux d'expression génique à l'échelle du génome, les profils d'expression d'extraits de tissus pulmonaires provenant d'un panel de lignées de souris congéniques dérivées des deux mêmes souches pures ont été analysés. Une ANOVA a été appliquée selon un modèle pour évaluer s'il existe une association entre l'expression des gènes et leur souche donatrice d'origine (SDO, déterminée par génotypage des lignées congéniques). Après un ajustement pour compenser la variabilité due à la souche dominante dans chaque lignée congénique, ce modèle a permis d'identifier plus de 1500 gènes dont le niveau d'expression est associé avec leur SDO ($P < 0,05$), ce qui suggère la présence de variation en cis pour la

régulation de l'expression de ces gènes. Nous avons sélectionné au hasard 50 gènes avec une association positive (positifs) avec leur SDO et 80 gènes non associées (négatifs) pour une validation de cette association à l'aide d'une méthode basée sur le déséquilibre allélique. Cette méthode utilise des variations de séquences intragéniques pour déterminer la différence du niveau d'expression des deux allèles d'un même gène chez un individu et ainsi détecter des gènes dont l'expression est affectée par la présence de variations en cis. Sur les gènes sélectionnés, 54% des gènes positifs renfermaient au moins une variation de séquence dans une région de 1000 nucléotides du 3' non traduit contre seulement 27% pour les gènes négatifs ($P < 0,05$, test exact de Fisher). Ces gènes ont été utilisés pour détecter la présence d'un déséquilibre allélique. Chez 63% des positifs contre 23% des négatifs ($P < 0,01$, test exact de Fisher), on a constaté la présence d'un déséquilibre allélique. Cette proportion représente un enrichissement de la détection de régions régulatrices potentielles de plus de 10 fois par rapport à une recherche aléatoire. Cette étude permet d'estimer de façon conservatrice que 34% des gènes renferment une région de régulation potentielle, une estimation très semblable à celles rapportées dans d'autres études effectuées chez les mammifères. De plus, cette étude démontre qu'une approche multidisciplinaire peut permettre de cataloguer les gènes susceptibles de contenir des variations en cis pouvant affecter leur régulation, une étape primordiale pour la cartographie des traits génétiques complexes et le développement de modèles de régulation génique des génomes de mammifères.

INTRODUCTION

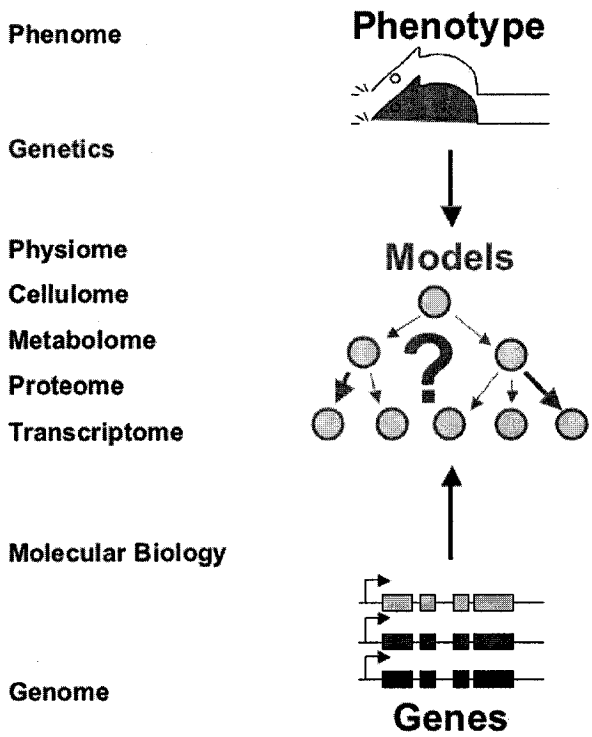
PART 1. BIOLOGICAL MODELS

Biology is a science where general principles for the most part remain to be defined. A few have been elucidated. Perhaps the equivalent to the Copernican revolution in biology is evolution[1], Darwin's statement that species evolve from each other under selective pressure from the environment. Around the same time in 1866, Mendel demonstrated that traits are transmitted from generation to generation. This has formed the foundation of quantitative genetics, the search for genetic causes of phenotypes[2]. Nearly 100 years later, the molecular basis for heritability was discovered[3]. The so-called "Central Dogma"[4], states that DNA is the material used to transmit information from generation to generation and that genes code for proteins via a 3 base code preserved across all life forms. This concept has formed the foundation of molecular biology for the last 50 years and has enabled the study of genes by manipulation of DNA. However, much remains to be answered. While the genetic code accounts for the process by which a protein is synthesized using a specific segment of DNA as its template, the system does not behave in ways predictable by these principles alone. Knowledge of the DNA alone has not revealed to us how phenotypes arise from genes. The whole appears far more complex than its parts.

Perhaps the major conceptual leap offered by completion of the genome projects and large-scale genomics is that we now realize how complicated the picture really is. Awareness of this complexity began to emerge over 40 years ago when the logic of gene regulation was first observed[5]. Proteins interact with other proteins, DNA, and RNA in

precise logical patterns regulating the activity of genes, akin to a molecular circuitry[6]. When expanded to the genome-wide scale, the concept of a molecular circuitry reaches levels of complexity that defy explanation using current models of molecular biology. The concept of a genome-wide regulatory network was initially proposed over 30 years ago[7]. Only in recent years through advances in genomic technology has it become possible to gather data on the scale necessary to test these concepts. Further layers of complexity exist; Transcripts are furthermore known to interact with other transcripts, as well as acting enzymatically. DNA is intricately woven into chromatin, the rules by which it behaves yet to be resolved. Cells interact with each other via finely regulated communication mechanisms. It appears that each level of organization represents a giant leap in the complexity of the system and in its ability to respond to intrinsic and extrinsic factors. This has led some to describe a model of biology as one consisting of layers of organization, each with its own complexity from the genome to proteome to metabolome to cellulome to physiome, all sublayers of the phenome[8]. Our ability to translate our knowledge from one level to the next currently remains limited. In the face of such complexity, how do we begin to construct the scaffolding upon which to assemble our understanding of the biological system as a whole? How do all these factors (and perhaps others we have yet to imagine) interact to bring about a functioning cell, a tissue, and an organism from the quaternary genetic code? Information transmission appears to occur throughout biological systems, from gene to protein, from generation to generation. Systematic cataloguing of each layer may expose the missing links by which information passes between these layers of organization. The goal of genomics research and systems biology is to find these links.

Figure 1. Biological models.



PART 2. GENETICS OF GENE REGULATION

The field of genetics has demonstrated a great ability to isolate genes responsible for monogenic traits, those caused by a single alteration of a single gene and displaying inheritance in ratios predicted by Mendelian inheritance. However, monogenic traits are generally rare. The majority of phenotypes including most common diseases are multigenic or complex. A trait is said to be complex when multiple factors (both genetic and environmental) contribute to the phenotype. Complex traits are much more difficult to dissect. While it is assumed that the individual genetic components of a multigenic phenotype remain transmitted in proportions according to Mendel, merely estimating the number of these factors involved in a given trait is a difficult task. The coordinated interactions between these factors only further obscure the relationship between genes and phenotypes. The biologist is faced with a potentially immense degree of complexity that increases exponentially with the number of genes involved. To date, while the study of simple Mendelian or monogenic traits has resulted in the identification of numerous genes, relatively few examples exist where genes contributing to complex phenotypes have been identified [9].

The search for genetic determinants of complex traits typically proceeds via an approach of measuring the association between phenotypes and genotypes across a given population. The use of inbred mice greatly facilitates this task by offering a population where genetic and environmental sources of variability may be controlled relatively well, and where phenotypes and segregating alleles may be observed over generations of individuals. Controlled breeding strategies may be used to further isolate strains or

individuals with a genetic composition more amenable to genetic analysis. The search culminates in the mapping of one or more quantitative trait loci (QTL). These may range in size from regions containing tens of genes to entire chromosomes. Fine mapping to narrow down these intervals is a laborious task, demanding much time and often involves breeding of animals with successively smaller portions of the segment of interest. The task may further be complicated by the presence of interactions within the locus in question. Genomic approaches offer a genome-wide perspective that provides a more immediate link between phenotypes and the level of individual genes.

In recent years, much attention has focused on the development of methods for determining complex traits where multiple loci (and, presumably, genes) interact to bring about a complex phenotype. This interaction effect is known as epistasis, or gene dependence. The importance of background genetic effects [10] and modifier genes have been recognized for some time. In certain cases, QTLs were discovered from the analysis of modifier genes [11, 12]. One example providing evidence for the substantial effect of modifier genes as well as the complexity of interactions, is that of bristle number in Drosophila [13, 14]. Examples where epistatic loci have been mapped in mice leading to candidate gene interactions include Mom1 mouse intestinal cancer [15, 16] and Moth1 for hearing loss[17]. The challenge of determining such interactions has led to the development of special experimental systems designed to dissect the nature of complex interactions. These include recombinant congenic strains (RCS) [18] and chromosome substitution strains (CSS) [19]. A major advantage offered by these strains over and above the F2 inbred panels traditionally used in mapping studies is the segregation of interacting alleles.

Two major theories about the allelic distribution of complex traits exist, one proposing a minority of alleles of major effect, the other proposing a multitude of genes all having a minor effect. In all likelihood both exist in varying degrees on a continuous scale as would be consistent with a hypothesis of genes interacting in complex regulatory networks. However, determining the extent to which each exists, in what conditions, or what proportions, remains to be characterized. Surveying this landscape of complexity is a necessary step in defining the framework for further discussion.

Our concepts of gene regulation date back over 40 years ago to the characterization of the first genetic regulatory system, the lac operon controlling the metabolism of lactose and B-galactosides in *E. coli* [5]. This work provided the theoretical groundwork upon which subsequent studies of gene regulation have been based – namely, that the control of transcription is governed by sequence elements upstream of the transcription start site, within the 5' UTR as well as within introns. Since then, numerous gene regulatory mechanisms have been elucidated. While the model of transcription controlled by activation factors and sequence elements upstream of a gene is conceptually simple, regulation of transcript levels are now believed to be much more complicated. The definitions of activation sequences have proven difficult to generalize, mainly due to the great variety in genetic promoter and enhancer elements. Context dependence of these elements, and the large distances (>100kb) over which these elements may be dispersed[20] further complicate the task of finding and predicting gene regulatory mechanisms. The methodology for assaying gene regulation currently lies at the scale of individual genes

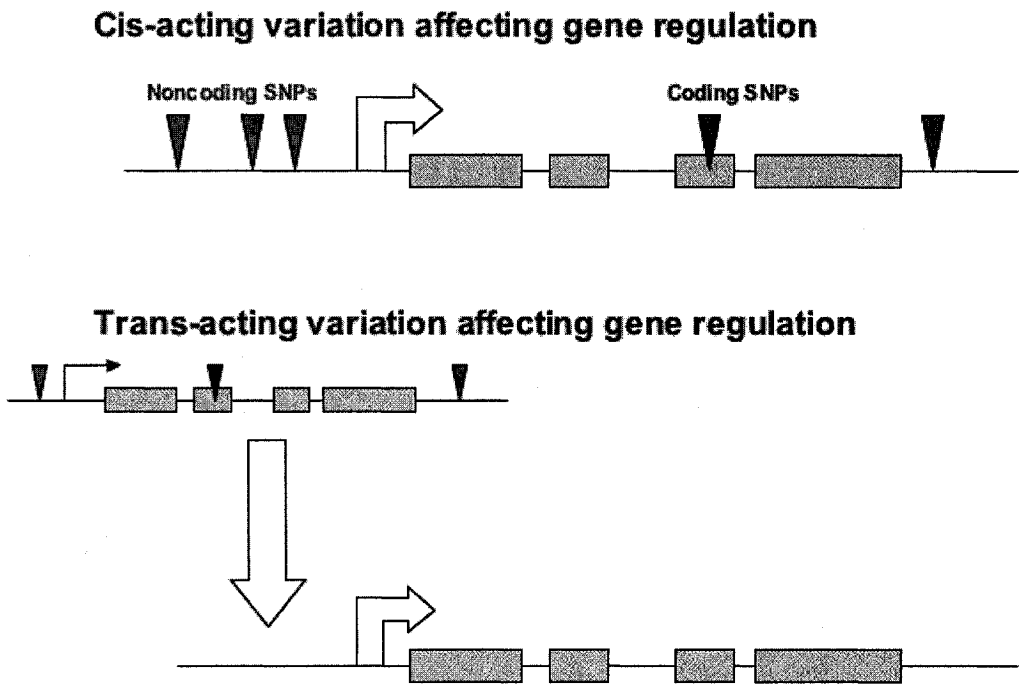
through transient transfection and mutagenesis assays *in vitro*. While these enable targeted analysis of an affected gene, such studies remove the gene from its natural *in vivo* context, making it difficult to relate results on a broader scale to observations in other experimental systems. In spite of our knowledge that many proteins bind to DNA, relatively few regulatory sequences have been characterized. At the genome-wide level, the picture of gene regulation remains sparse. New approaches are needed to find these mechanisms, and new vocabulary is required to expand our models of gene regulation encompass the tens of thousands of genes thought to exist in mammalian genomes.

Genes regulated by multiple factors often exhibit interactions between factors in a complex set of rules and context dependencies. One example is the cis-regulatory logic elucidated for the sea urchin gene, *endo16* [21]. Here, 12 binding species and 13 cis-regulatory elements located within a 2300bp region interact to control the expression of *endo16*, which is involved in early development. The precise mechanisms by which these species interact have been described by a series of conditional logic statements that accurately predicts operation of the system. Should such regulatory complexity exist for the majority of genes, the task of elucidating these dependencies on a large scale shall be challenging indeed.

The terms, cis and trans, used to described gene regulation, find their origins in the complementation test in *Drosophila* [22] where the terms were used to describe the configuration of mutant alleles. The terms now refer to the configuration of regulatory factors with respect to the affected gene and whether or not they reside on the same DNA

molecule or chromosome [23] (Figure 2). Darnell defines "Cis-acting: Referring to a regulatory sequence in DNA (e.g., enhancer, promoter) that can control a gene only on the same chromosome. In bacteria, cis-acting elements are adjacent or proximal to the gene(s) they control, whereas in eukaryotes they may also be far away," and "Trans-acting: Referring to DNA sequences encoding diffusible proteins (e.g., transcription activators and repressors) that control genes on the same or different chromosomes." [24]. Naturally, gene regulation is a concept that defies such a simplistic classification scheme. The sheer complexity of gene interactions shall require a new vocabulary to accurately describe and classify genetic regulatory networks on the scale of hundreds to thousands of genes. However, until further categories of gene regulation are described, the distinction, no matter how simplistic or even erroneous, remains useful to geneticists and biologists alike trying to make inroads into the morass of gene interactions and regulatory networks.

Figure 2. Models of gene regulation - cis versus trans acting mechanisms.



For the purpose of our discussion of genetic variation affecting gene regulation, gene expression shall be said to be cis-regulated when the factor affecting gene expression level is located proximal to it on the same chromosome. Conversely, gene expression variation shall be said to be trans-regulated when it is affected by other genes or elements located distally on other chromosomes. For genetic mapping studies attempting to identify genetic variants involved in complex traits, knowledge of whether markers and disease alleles reside on the same or different chromosomes is an important step in simplifying the search for these genes. Recent studies estimate the proportion of genes affected by cis-regulatory mechanisms to be between 30% and 50% in humans[25, 26]. Trans-acting regulatory interactions are believed to affect the majority of genes stemming mostly from observations in yeast including surveys for DNA binding proteins[20], gene network inference from gene expression profiling in gene knockout panels[27], and QTL

mapping of expression traits[28]. In mammals, estimating the proportion of trans-regulatory variation has been more difficult. A recent study in humans estimated 77% of gene expression variation to be trans-regulated[29, 30]. Evidence from the evolutionary perspective may shed further light on the importance of regulatory variation between populations. A comparison of gene expression in primates and mice subspecies has shown that species-specific expression profiles may be used to reconstruct the phylogenetic relationship between species [31]. Similar differential expression has been shown between related populations in a number model systems including yeast[32], *Fundulus*[33 34], and *Drosophila*[35]. The observations point to the importance of variation in gene expression with respect to phenotypic differences and imply an extensive role of variants affecting gene regulation throughout the genome.

While traditional molecular biology enables dissection of regulatory mechanisms for individual genes or small sets of genes, the work is time-intensive and generally must proceed based on prior hypotheses of the mechanism in question. This piece-wise progression has led to a gradual increase in knowledge about regulatory mechanisms biased towards a minority of genes for which prior knowledge exists. From this type of research, pathway diagrams of small sets of genes have been constructed with little knowledge of the contextual dependencies of such models with respect to the system as a whole. With information becoming available for entire genomes, a picture is emerging that most genes interact in some way with other genes. If we are to understand the action of one gene, or even one pathway, we must understand the simultaneous action of many others in the system. In short, we cannot fully understand one gene or pathway outside

the context of the entire network in which it exists. The advantage of system-wide experimental techniques is that we may be able to identify new mechanisms of which we had no prior knowledge, and to begin creating models that capture context dependence over the entire system.

In all likelihood, genes function in a complex network of dependencies, as predicted over 30 years ago[7]. During the course of this thesis, evidence has emerged in support of this prediction. Analysis of single and double gene deletion yeast strains has revealed a varying degree of connectivity between genes[27, 36, 37]. Studies of DNA binding proteins using chromatin immunoprecipitation have arrived at similar conclusions[20]. Efforts to understand complex phenotypes will require a broader map of this network in mammals. If the genome functions in a highly coordinated fashion, as many lines of evidence suggest, then a system-wide survey of gene regulatory mechanisms is needed to gain the necessary perspective for constructing the map of gene regulation on a genome-wide scale. What kinds of regulation exist in the genome? What are the categories that will be useful for genetic studies linking genes to phenotypes? How many genes will there be in each category? Are there general principles regarding gene regulation applicable to all biological systems? If so then how can we find them?

PART 3. MICROARRAY TECHNOLOGY

During my graduate studies, I witnessed the evolution of DNA microarray technology. Upon their inception, the ability to monitor expression levels for all genes simultaneously was both exciting and daunting at the same time. Excitement stemmed from the

knowledge that these profiles were a reading from the molecular circuitry of the cell. It was hoped that with the correct interpretation of these signatures, the entire circuitry of gene interactions could be deduced. The initial experiments applied methods from artificial intelligence to the data in hopes that the patterns would lead to an understanding of the circuitry[38, 39]. However, the task has proven to be far more difficult than initially imagined. Biological processes are generally believed to give rise to patterns in data. A large number of patterns may be extracted from microarray data using a variety of methods[38, 39]. However, distinguishing biologically meaningful patterns from spurious correlations has proven difficult in practice. There are many sources of variability contained in microarray data, all of which may confound biologically significant correlations. The number of steps between the capture of biological material and observation of gene hybridization levels renders it difficult to extract biologically significant signals from experimental and technical noise (to be discussed in detail further on). Besides issues of noise pertaining to statistics and study design, the complexity of the biological system may be greater than previously anticipated. The prevalence of complex gene interactions leaves scientists with relatively few starting points upon which to base a frame of reference. The challenge is furthermore complicated by the fact that biological systems are dynamic whereas microarray measurements represent static images of the system. The eventual goal of such an approach is to construct a genome-wide model of the molecular workings of the cell. While the understanding may not arise from a single study, a cumulative approach may enable the snapshots to be assembled into a rough sketch of the genetic network.

A BRIEF REVIEW OF MICROARRAY TECHNOLOGY

DNA microarrays represent a miniaturization and scaling up of traditional hybridization assays such as dot blots where DNA fragments derived from each gene or EST are attached to a solid support. These attached fragments then act as probes for labeled cDNAs derived from sample mRNA. Two main platforms are in common use: 1) Glass slide arrays (often termed cDNA arrays), which involve depositing cDNAs onto a coated glass slide using a robotic spotting apparatus[40]. This method offers the advantages of flexibility in the design of the array and a lower cost to produce. The disadvantages of the method include the multitude of variables affecting the manufacture of the array, most of which translate in variations in spot quality. Conditions such as temperature and humidity can influence the quality of arrays produced. Variations in spot size and shape typical of capillary tube spotting, contribute further to the overall variability observed in the system. Commercially produced arrays alleviate some of these issues. Since none of the experimentation in this thesis involved use of spotted cDNA arrays, the platform shall not be discussed any further. 2) Oligonucleotide arrays produced by Affymetrix were used for all expression studies in this thesis. These arrays are generated by a process of photolithography where oligonucleotides are synthesized directly onto the array substrate[41, 42]. A collection of oligonucleotides is designed generally from the 3' UTR of each transcript and each probe is paired with an oligonucleotide containing a mismatch at a central position. The hybridization intensities from each probe in a probeset are then combined or summarized to generate a single intensity measurement for each gene. A variety of statistics exist for summarizing probesets in to a single reading and are examined throughout the following chapters.

Any method used to analyze microarray data faces the same question: How good is the data? It is a difficult question to answer. This thesis attempts to address some of these issues, in particular, those of experimental control and reproducibility. The fact remains that microarrays provide images of genomic function, unprecedented in biological research. Relating these results to existing research continues to be a challenge. One may choose to base estimates of accuracy on how well the technique identifies previously characterized relationships. However, most existing knowledge has been generated in targeted experimental systems with no guarantees of universality across broader experimental designs. Comparison of microarray analysis against independent biological assays such as RT-PCR or Northern blotting has been viewed as an acceptable method of cross validation. These techniques inevitably involve a greater investment of time and resources, in addition to requiring larger amounts of biological material than is sometimes possible. Furthermore, each technique involves a degree of variability, sensitivity and reproducibility of its own, not all of which have been subjected to as much statistical investigation to date[43]. While microarrays are more recent than traditional methods, much more is known about the platform statistically. Since their inception, microarrays have received intense statistical scrutiny aiming to determine the optimal methods for analyzing the data. By comparison, RT-PCR and blotting assays have received less attention with respect to reproducibility, replication and overall statistical characteristics. Indeed the issue of using single-gene assays to verify microarray datasets raises more questions than it answers, and the need to corroborate microarray results with traditional assays remains a contentious issue within the community.

PART 4. INBRED MICE

Inbred mice have been used for close to a century as a model system in which to study genetics[44]. Indeed there are numerous mouse models for human diseases[45-49]. The main advantage of the mouse is that it is a model organism that is relatively closely related to humans. While certain subsets of characteristics are shared across all organisms, many are only shared between mammals. Startling similarities are known to exist between mouse and human phenotypes[50]. Decades of work characterizing phenotypic differences between strains have led to a wealth of knowledge about this system at the cellular, physiologic and organismal level. Furthermore, this information has been systematically catalogued over the last 5 years culminating in the Mouse Phenome Database[51]. This kind of knowledge base is an invaluable resource for the validation of genomic results.

From the standpoint of genetic research, mice offer numerous practical advantages to scientists. Besides being small and easily housed in controlled environments, they breed rapidly and have easily identifiable physiological phenotypes. Collectors initially established the inbred lines currently in use in the late 1800s and early 1900s. The systematic cataloguing and standardization of breeding culminated in the foundation of the Jackson Laboratory in 1930 by Castle[44]. The genealogy of the strains has been well recorded [52], the most commonly-used inbred strains deriving from the *Mus musculus* group [44]. Nomenclature, breeding and distribution of inbred mouse strains are now standardized at the Jackson lab where strains are defined as inbred after greater than 20 generations of sibling breeding, and are considered to be homozygous at every locus (>96%). This provides a population that is effectively genetically identical, dramatically

simplifying the search for genetic loci. Furthermore, the presence of multiple strains, each with different genetic compositions, allows the design of targeted breeding experiments whereby traits and genetic markers may be observed simultaneously over multiple generations. Indeed, mice represent an experimental system free from many confounding issues facing studies in humans, such as sampling bias, genetic diversity of the sampled population and environmental variables. The elimination of confounding variables together with the ability to gather large sample sizes create an experimental system with increased power to observe biological variables of interest.

In recent years, much attention has focused on characterizing the molecular differences between mouse strains. The complete genome sequence was completed for *Mus musculus* in 2002, followed by partial genome sequencing of a number of the more common inbred strains[53]. One rationale for this sequencing was to characterize in finer detail the genetic polymorphisms between these strains. Traditionally, mapping studies have used microsatellites or SSLPs, (simple sequence length polymorphisms) located on the chromosomes by genetic mapping in mouse pedigrees or by radiation hybrid mapping of the mouse genome[54]. While SSLPs still constitute a sequence-verified repository of genetic markers, SNPs are estimated to exist at a much higher frequency and thus promise to allow mapping of genetic variation at a much higher resolution[55]. SNPs exist in regions of high density (100 per 500kbp) and low density (10 per 500kbp)[56]. These regions appear to correlate with the haplotype structure of the mouse genome[57, 58]. Studies have revealed that the genome of inbred strains appears to share a proportion of blocks derived from a common ancestor with different overlapping regions shared

between different strains. Knowledge of this haplotype structure not only complements existing mapping studies but may also provide a means to map traits via comparison of inbred strains directly[59].

This study focuses on two of the most commonly used inbred strains, A/J and C57BL/6J. These have been bred for over 100 generations and represent two of the earliest established inbred strains[44]. Phenotypic comparisons between these two strains date back decades[60]. These strains have been characterized for hundreds of phenotypes (including asthma, diabetes, cancer) with over 850 registered studies in the Phenome Database (www.phenome.org). The two strains both derive from the *Mus musculus* subgroup and are known to differ for at least 3200 microsatellite markers [61, 62] and for over 120,000 SNPs [63]. Genetic studies in A/J and C57BL/6J mice have resulted in numerous QTL mapped for a range of phenotypes (for a comprehensive list see Table 6, Chapter 2). The wealth of systematically catalogued knowledge from numerous studies at all levels of function, together with the availability of genomic resources makes the use of these animals extremely attractive for genomic approaches.

PART 5. INTEGRATIVE APPROACHES

The difficulties in finding genes underlying complex traits using current methods have been well documented[64]. The number of epistatic interactions between loci together with combinatory complexity makes the definitive isolation of a QTL highly context dependent. Effects seen in one context may manifest entirely differently in a separate experimental system. Furthermore, strategies for fine mapping to identify candidate genes

are also hampered by the combinatorial complexity of gene interactions, as well as the genomic size of many QTL, which may contain hundreds of genes. Because of this, the molecular basis of most QTLs remains a mystery. More recent studies have demonstrated the effectiveness of incorporating gene expression profiling into the process of studying complex traits in mice[28, 65-67]. By this approach, gene expression levels are considered molecular phenotypes, which can then be associated with recombination patterns detected by genetic marker analysis. Such an approach marries a system-wide view of molecular biology with genetic marker analysis and offers the advantage of bringing the study of complex traits from the level of genetic regions to individual genes[68]. However, the complexity of gene regulation and interactions suggests that expression profiling per se will not suffice[69].

As shall be evident from the following chapters, one aspect of utilizing genomic technologies like DNA microarrays is the difficulty in drawing conclusions similar to those of molecular biology where the causality of a given gene-gene interaction, for example, may be demonstrated conclusively albeit within a restricted experimental system. In a genome-wide experiment there are degrees of variation, the measurement of which comes with estimates of error. While it is possible to measure degrees of association, assignment of causality is far more difficult. This is where a diversity of approaches and information sources may excel over a single-faceted study. The combined approach of using multiple information databases, multiple genomic technologies across multiple experimental systems has been shown to be an effective means to answering complex biological questions[70]. Integrative approaches provide a multifaceted view of the

biological system that surpasses the capacity of each technique when applied separately. This thesis represents such an integration of approaches.

PART 6. THESIS OBJECTIVES

The focus of my thesis was to investigate the genetic basis of gene expression and regulation on a genome-wide scale focusing on understanding gene expression profiles in inbred mice. The thesis is divided into three chapters. Chapter 1 represents an effort to understand the technology of genome-wide expression profiling, examining the variability inherent in the data in order to determine how to best design and control microarray experiments. These studies showed that microarray data contains different sources of variability, which must be taken into account in the design of experiments. Chapter 2 deals with the application of the technology to a model system of importance to biology and genetics, inbred mice. This study analyzed expression profiles of two inbred strains of mice (A/J and C57BL/6J), demonstrating extensive baseline gene expression variability between inbred strains for multiple tissues and addressing issues of reproducibility in microarray experiments. The study further identified genes within previously mapped QTL indicating the potential for the approach to prioritize the search for candidate genes within QTL. Chapter 3 further investigates the dependencies of gene expression differences upon specific genetic differences via an integrated approach combining expression profiling obtained over a panel of RCS mice, and validating the results using allelic imbalance (AI). The increased incidence of allelic imbalance in genes identified by expression profiling demonstrates the effectiveness of an integrated approach for enriched detection of potentially cis-regulated genes. This approach may facilitate large-scale cataloguing of cis-regulated genes.

CHAPTER 1 – UNDERSTANDING THE SYSTEM: CONTROLS IN MICROARRAY EXPERIMENTS

The first and most pressing question facing early microarray experiments was what constituted an effective control in comparing different expression profiles. In spite of much progress in the field, the question remains an issue to this day. Much of what we know about gene expression experiments comes from experience with traditional hybridization-based assays, such as Northern blots, slot blots, and RT-PCR, where use of a control gene constitutes an adequate control for measuring small numbers of transcripts. It was not clear that similar strategies could be applied to microarray data. This study aimed to better characterize the nature of microarray data and to determine adequate methods for developing internal controls for microarray experiments. What appeared to be a simple question at first has since become a far more involved discussion, as microarray data has proven far more complex than had previously been imagined.

Upon first sight of microarray experiments, scientists were faced with a phenomenon that had never been seen before, snapshots of expression levels genes measured simultaneously for thousands of transcripts. One approach to describing a novel observation is to relate the new phenomenon to a pre-existing concept as a means of defining a framework for further discussion and exploration. Some of the earliest studies applied clustering and other pattern recognition tools derived from other disciplines to analyze the data[38, 39]. However, while these proved to be a quick and accessible way to visualize and superficially explore microarray datasets, they had limited utility in terms of defining the data in a way that could be related to existing hypotheses. There was no

way to test the accuracy or precision of the gene expression changes detected. In the face of this dilemma, we first posed the question: What constitutes an effective control in microarray experiments? How could we better estimate the association between gene expression changes and independent variables within one experiment, and how reliable would such estimates hold up between experiments and test systems?

Early microarray experiments were primarily exploratory[71]. Datasets were generated without replicates and controls, and consisted of comparisons against reference samples. Admittedly, the excitement of these early stages led to studies that explored the possibilities of the technology, rather than exposing its pitfalls. However, the question remained, how significant were these changes, and how reliable were conclusions that were based on this data?

Quantitative measurement of mRNA transcripts was established early in technological development of microarrays[40, 42]. However, establishing the reliability of transcript measurements for thousands of genes raised new questions of controls. Traditional methods for mRNA detection for single gene measurements compared transcript levels with those of a control gene, usually one thought to be ubiquitously expressed at a relatively stable level [72]. Calculations of relative transcript levels have been the standard for techniques such as Northern blots and RT-PCR. While such studies focused on questions confined to specific cellular contexts, microarray studies tended to involve comparisons over much broader experimental contexts. Previously it had been assumed that genes existed that were expressed ubiquitously across all tissues, and whose

expression did not vary. The term "housekeeping genes" were coined to describe their presumed function, one that was necessary for some maintenance role in every cell regardless of tissue. Controls for all mRNA hybridization experiments to date were based on this assumption. This study examined the validity of this assumption by analyzing three previously published datasets in addition to one generated with containing technical and biological replicates. Levels of traditional control genes and different sources of variability were examined. This study also examined the impact of these observations on normalization methods and began to evaluate statistical methods for the determination of differentially expressed genes, as opposed to methods based on fold-change calculations commonly used at the time.

Note about analysis methods: Since the time this analysis was performed, analytical methods have progressed substantially. Work on summary statistics was limited at the time. This study used MAS4.0 (an average difference calculated from both PM and MM probe signals). This method has since been deprecated. No longer are mismatch oligonucleotides included in summary calculations of expression levels for individual genes. Numerous normalization and summary methods have since emerged and were used in Chapters 2 and 3. While these methods have helped to improve the sensitivity of subsequent analyses, a comparison of the various methods led us to believe these differences would not have had appreciable effect on the conclusions of this chapter.

CONTROL GENES AND VARIABILITY: ABSENCE OF UBIQUITOUS REFERENCE TRANSCRIPTS IN DIVERSE MAMMALIAN EXPRESSION STUDIES

Lee PD, Sladek R, Greenwood CM, Hudson TJ.

Originally published in Genome Research, February 2002, 12(2):292-7.

ABSTRACT

Control genes, commonly defined as genes that are ubiquitously expressed at stable levels in different biological contexts, have been used to standardize quantitative expression studies for more than 25 years. We analyzed a group of large mammalian microarray datasets including the NCI60 cancer cell line panel, a leukemia tumor panel, and a TPA induction time course as well as human and mouse tissue panels. Twelve housekeeping genes commonly used as controls in classical expression studies (including GAPD, ACTB, B2M, TUBA, G6PD, LDHA, and HPRT) show considerable variability of expression both within and across microarray datasets. While we can identify genes with lower variability within individual datasets by heuristic filtering, such genes invariably show different expression levels when compared across other microarray datasets. We confirm these results with an analysis of variance in a controlled mouse dataset demonstrating the extent of variability in gene expression across tissues. The results demonstrate the problems inherent in the classical use of control genes in estimating gene expression levels in different mammalian cell contexts, and highlight the importance of controlled study design in the construction of microarray experiments.

INTRODUCTION

While DNA microarrays open the door to large-scale expression experiments [73, 74], a major challenge facing these studies is the design of experimental controls that will permit comparison of quantitative expression profiles obtained from diverse biological contexts. In traditional assays, standardization of mRNA levels has been achieved by comparison to the level of a control gene, commonly defined as one that is ubiquitously expressed at stable levels across many biological contexts. Methods of standardization based on control genes have furthermore been used in microarray and genomic studies [75, 76]. We re-examine the traditional concepts of controls in expression experiments with the aim of determining appropriate measures for the control of microarray experiments.

In an attempt to identify genes that are expressed at constant levels across a wide range of biological contexts, we analyzed four published datasets prepared following similar methods based on a single microarray technology (Affymetrix oligonucleotide microarrays). The NCI60 dataset [77] consists of microarray measurements of gene expression in 60 cancer cell lines originating from 9 tissue types. A dataset obtained from patients with hematologic malignancies [78] includes expression profiles for multiple homogeneous ALL and AML tumor samples. Temporal and developmental fluctuations in control gene expression were assessed using a dataset obtained from four cell lines following treatment with TPA [38]. Finally, the Huge Index dataset provides in vivo gene expression data for six human tissues [79].

METHODS

Public Microarray Datasets

Microarray datasets for the NCI60 cancer cell line panel, the ALL/AML tumors, and the TPA treatment in HL60, U937, NB4 and Jurkat cell lines are available at http://www.genome.wi.mit.edu/MPR/datasets). The human tissue expression profiles contained in the Huge Index dataset were obtained at http://www.hugeindex.org/).

Mouse Microarray Dataset

Mouse tissues were obtained from three adult male C57BL/6J littermates. Mice were killed by cervical dislocation and the tissues rapidly dissected and homogenized in Trizol reagent (Life Technologies). Total cellular RNA was prepared according to the manufacturer's instructions and analyzed by non-denaturing (1% agarose-1xTBE) gel electrophoresis. Probes for the microarray studies were prepared by priming 20 μg of total RNA with 100 pmol of T7- (T) 24 primer (Genosys). The RNA-primer mixture was denatured for 10 minutes at 70°C then chilled on ice. First strand cDNA was synthesized using Superscript II reverse transcriptase (Life Technologies). Second strand synthesis was performed using RNAse H, DNA polymerase I and *E. coli* DNA ligase (Life Technologies). Biotinylated riboprobes were prepared from the entire cDNA reaction using the ENZO Bioarray High Yield RNA Transcript Labeling Kit (ENZO Diagnostics). The average probe length was reduced by incubating the probe in 1X Fragmentation Buffer for 35 minutes at 95°C. Hybridization was performed at 45°C for 16-20 hours using 15μg of biotinylated probe. Following hybridization, the arrays were subjected to 10 low-stringency washes and 4 high-stringency washes using a GeneChip Fluidics

Station 400 (Affymetrix). Bound probe was detected by incubating arrays with SAPE (streptavidin phycoerthryin, Molecular Probes) and scanning the chips using a GeneArray Scanner (Agilent). Scanned images were analyzed using the GeneChip Analysis Suite 3.3 (Affymetrix). Full details of the microarray methods have been described previously[80].

Data analysis

Traditional control genes analyzed in human datasets included: beta-actin (ACTB), beta-2-microglobulin (B2M), phosphofructokinase (PFKP), phosphoglycerate kinase (PGK1), aldolase A (ALDOA), phosphoglycerate mutase (PGAM), alpha-tubulin (TUBA), glyceraldehyde-3 phosphate dehydrogenase (GAPD), glucose-6 phosphate dehydrogenase (G6PD), lactate dehydrogenase A (LDHA), hypoxanthine phosphoribosyltransferase (HPRT), and vimentin (VIM). Traditional control genes analyzed in mouse datasets included: asparagine synthetase (Asns), phosphofructokinase (Pfkp), lactate dehydrogenase A (Ldh1), vimentin (Vim), phosphoglycerate kinase (Pgk1), ubiquitin (Ubc), glucose-6 phosphate dehydrogenase (G6pd), phosphoglycerate mutase (Pgam1), beta-2-microglobulin (B2m), glutamate dehydrogenase (Glud), hypoxanthine phosphoribosyltransferase (Hprt), alpha-tubulin (Tuba1). For accession numbers see Table 1.

Regression scaling was performed only on datapoints assigned a 'P' absolute call by the Affymetrix GeneChip software: the absolute call estimates the hybridization quality for an individual probe set based on measures of background and signal dispersion. The regression scaling algorithm has been described previously[80]: it utilizes normalization to

the regression coefficient of the first sample in each dataset. We rescaled datasets based on mean overall intensity per scan. Mean intensity was calculated on the genes with a minimum average difference of 50 and an absolute call of 'P' by the GeneChip algorithm.

Data analysis was accomplished using Perl or VBScripts in Microsoft Excel. Graphs were created using R (http://www.r-project.org). ANOVA was carried out using SAS (SAS Institute Inc) testing the amount of observed variability in expression of each gene due to replicate (repeat hybridizations of the same RNA sample), mouse (samples from three individual mice), or tissue (samples from 4 different tissues); General linear model was used on a per-gene basis (PROC GLM). P values considered were calculated for each variable individually having adjusted for the variation due to remaining variables (added-last test / SAS Type III F-Test). We conducted ANOVA separately on subsets of the data meeting initial filtering criteria of minimum expression levels of greater than 20, 50, 100 or 200 units across all 12 experiments. ANOVA results must be interpreted with caution as the small sample size makes assessments of normality and homoscedascity difficult. P-values considered were for the added-last F-test (testing each variable individually having adjusted for all other variables). Datasets, figures, tables, and analytical scripts are available at http://www.mcb.mcgill.ca/~pdlee/control_genes .

RESULTS AND DISCUSSION

We initially studied the expression levels of 12 genes commonly employed to normalize RNA levels measured by Northern blots or RT-PCR. The expression levels for many of

these genes fluctuates dramatically both within and across datasets (Figure 1). Within datasets, the maximum fold change (MFC - the ratio of the maximum and minimum values observed within a dataset) ranges from 1.3 for ACTB within the TPA induction dataset to greater than 300 for VIM in the NCI60 dataset (Table 2). All commonly used control genes have MFC of greater than 2.0 in at least one dataset. In addition, the observed coefficients of variation (CVs) are frequently greater than 0.5, reflecting the highly variable levels of expression of these genes within data sets.

We next employed a simple heuristic filter to identify sets of genes showing lower variability. After excluding genes with signal intensities below threshold and with a MFC greater than 2, genes were sorted and ranked according to their CVs. We use this measure of variability as it compensates for the apparent dependence of dispersion on signal[80]. Similar results were obtained using alternate methods of estimating dispersion (data not shown). Of the housekeeping genes analyzed, only GAPD and ACTB rank among the 100 genes with the lowest variability; however, no traditional control genes display consistently low variability across the four datasets. Nine genes identified by filtering have CVs less than 0.7 across all four datasets (Table 3). These are not genes that have commonly been used as controls, but include several ribosomal protein (RP) genes (including RPS27A, RPL19, RPL11, RPS29, RPS3). Even this set of genes shows differing amounts of variability across datasets (Figure 2). For example, while RPS27A has the lowest CV in the NCI60 and leukemia datasets, its MFC ranges from 2.2 in the NCI60 dataset to 5.6 in the TPA induction dataset.

Our failure to identify control genes in the four expression datasets studied might occur if the microarray measurements were associated with high levels of technical variability. In order to assess whether the observed variation could be due to technical variability rather than biological context, we examined expression levels in triplicate for RNA samples obtained from liver, heart, lung and brain of three male C57BL/6J mice reared under identical conditions using MU11KA and B arrays (Affymetrix) containing probe-sets for 11000 mouse genes and ESTs. The expression levels of traditional control genes show greater variability among RNA samples obtained from different tissues than among RNA obtained from the same tissue harvested from different mice, or among identical RNA samples hybridized to replicate microarrays (Figure 3). To determine whether other genes displayed similar behavior, we performed analysis of variance (ANOVA) on a per-gene basis to determine the amount of observed variability that could be attributed to differences among replicates, mice or tissues. Technical replicates using identical RNA samples hybridized to 3 distinct arrays show the least amount of variability: only 3% of genes display significant differences across replicates ($P<0.05$). Among biological replicates using RNA from 3 individual mice, 5-10% of genes show significant differences ($P<0.05$) after adjusting for variation between tissues, and between technical replicates. In contrast, 81-99% of genes show significant variability ($P< 0.05$) among different tissues after adjusting for the variability between technical and biological replicates. This trend remains consistent regardless of the filtering criteria or procedure used to select genes (Tables 4,5,6,7, Figure 4). ANOVA performed on the TPA induction and NCI60 datasets similarly reveals greater variability in gene expression across different tissues than across different time points, cell lines or datasets. Performing our

analysis using multiple normalization methods did not impact our findings (Figure 5A and B). These results indicate that the variability in gene expression detected in this experiment is not due to technical or inter-mouse variability, but rather due to the inherent differences in individual RNA levels present among different tissue types.

It is possible that our failure to identify control genes may result from data filtering techniques that excluded RNA species expressed at low copy number across a wide range of tissues, or, genes that are simply not present on the microarrays used in these studies. These issues may be addressed by the future development of more sensitive complete genome arrays. Despite this, our results clearly show that the expression levels of genes that have been commonly used as controls in classical experiments vary significantly among different cellular and experimental contexts. Furthermore, we fail to identify mammalian genes that qualify as "control genes" based on a definition of ubiquitous and stable expression. While some genes do appear quite stable in expression level within any one experiment, there do not appear to be any genes expressed at stable levels across all four datasets studied in this paper. Furthermore, the extent of variability between datasets appears to differ substantially (Table 8). Hence the traditional use of individual genes, as normalization controls in experiments that compare diverse biological tissues would lead to substantial errors in the derived estimates of fold change in gene expression levels. From inspection of the data it is apparent that some transcripts may serve as control genes for studies performed in a single tissue context, however these conclusions are limited by a study design that does not address the effects of physiologic regulation on the expression of these genes.

The unproven existence of control genes seems to have achieved acceptance in part due to its conceptual simplicity and practical limitations of the past. Recent studies have expressed concern that individual genes or groups of genes may serve as inadequate internal standards for measuring RNA expression levels[72, 81-87]; Measures for data standardization and quality control in microarray databases are currently being reviewed by the MGED working group on Microarray Data Annotations (www.mged.org). The establishment of common frames of reference requires a re-examination of assumptions inherent in the design of biological experiments. From these findings we propose that all genes are differentially expressed in at least one biological context and that the expression of every gene is therefore context dependent. Given the absence of ubiquitous control genes, variation in microarray expression studies must instead be interpreted using statistical characteristics of the data without preconceptions arising from the traditional notions of internal control genes.

Table 1. *Accession numbers of traditional control genes tested.*

ACTB	M10277 X63432
ALDOA	X12447 H24754
G6PD	X55448 X03674
GAPD	X01677 T55131
B2M	S82297 T48041
PFKP	D25328 H28131
PGK1	V00572
PGAM1	J04173
TUBA1	X01703 K03460
HPRT1	M31642 V00530
VIM	Z19554 T51852
LDHA	X02152
RPS27A	S79522 H89983
RPL19	X63527
HSPCA	X15183
RPL11	X79234
RPS29	U14973
NONO	U02493
AAMP	M95627
RPS3	X55715
ARHGDIA	X69550
Asns	U38940
B2m	X01838 AA059700
Tuba1	M13445
Vim	X51438 W21013 AA024049
Pfkp	AA072252
Ubc	D50527
G6pd	Z11911
Pgk1	AA097524 W75817
Hprt	J00423
Glud	X57024
Pgam1	AA065739 AA161799
Ldh1	X02520 Y00309

Table 2. *Traditional control genes across 4 datasets.*

Gene	Description	NCI60		AML/ALL		Huge Index		TPA	
		MFC	CV	MFC	CV	MFC	CV	MFC	CV
ACTB	Actin, beta	7.8	0.39	3.5	0.32	1.3	0.11	4.6	0.35
ALDOA	Aldolase A	8.2	0.37	5.7	0.48	2.9	0.37	2.9	0.31
G6PD	Glucose-6-phosphate dehydrogenase)	7.7	0.90	5.0	0.45	2	0.28	4.6	0.43
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	5.3	0.31	12.3	0.33	2	0.19	1.9	0.17
B2M	Beta-2-Microglobulin	25.6	0.58	14.0	0.42	3.4	0.27	4.6	0.49
PFKP	Phosphofructokinase, platelet	12.4	0.68	>12	1.66	>96	0.08	14.0	0.56
PGK1	Phosphoglycerate kinase 1	6.5	0.44	6.8	0.36	N/A	N/A	4.9	0.41
PGAM1	Phosphoglycerate mutase 1 (brain)	5.2	0.40	12.4	0.48	1.6	0.19	3.6	0.47
TUBA1	Tubulin, alpha 1 (testis specific)	>50	1.09	53.7	0.78	3.1	0.84	6.7	0.46
HPRT1	Hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)	10.7	0.45	43.9	0.53	1.6	0.42	5.6	0.49
VIM	Vimentin	>300	0.68	12.7	0.42	3.8	0.41	28.9	0.94
LDHA	Lactate dehydrogenase A	91.0	0.38	9.2	0.40	5.5	0.60	4.1	0.34

CV – coefficient of variation, MFC – maximum fold change, NCI60 – expression dataset generated from the NCI panel of 60 cancer cell lines[88], AML/ALL – acute myeloid leukemia and acute lymphoblastic leukemia dataset[78], the Huge Index dataset[79], TPA – expression data from HL60 cells treated with TPA[38].

Table 3. *Genes identified by filtering with CV less than 0.7 across all 4 datasets.*

Gene	Description	NCI60		AML/ALL		Huge Index		TPA	
		MFC	CV	MFC	CV	MFC	CV	MFC	CV
RPS27A	Ribosomal protein S27a	2.2	0.17	3.2	0.19	3.1	0.37	5.6	0.61
RPL19	Ribosomal protein L19	3.2	0.24	2.6	0.19	2	0.21	2.6	0.29
HSPCA	Heat shock 90kD protein 1, alpha	9.8	0.25	7.7	0.44	7.3	0.47	4.8	0.32
RPL11	Ribosomal protein L11	3.4	0.25	3.8	0.2	2.8	0.3	11.8	0.54
RPS29	Ribosomal protein S29	3.1	0.26	4.2	0.25	1.6	0.16	3.3	0.33
NONO	Non-POU-domain-containing, octamer-binding	4	0.28	3.7	0.25	1.4	0.1	4.3	0.33
AAMP	Angio-associated migratory cell protein	4	0.31	8.4	0.34	2.2	0.31	4.2	0.43
RPS3	Ribosomal protein S3	4.2	0.32	3.5	0.24	3.9	0.35	2.7	0.26
ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha	7.6	0.38	7	0.29	1.4	0.15	3.8	0.31

CV – coefficient of variation, MFC – maximum fold change, NCI60 – expression dataset generated from the NCI panel of 60 cancer cell lines[88], AML/ALL – acute myeloid leukemia and acute lymphoblastic leukemia dataset[78], the Huge Index dataset[79], TPA – expression data from HL60 cells treated with TPA[38].

Table 4. *Summary of ANOVA conducted on mouse dataset.*

Threshold	Number of genes with expression > threshold	Proportion of genes with P-value < 0.05 when following variables were tested individually*		
		Replicate (n=3)	Mouse (n=3)	Tissue (n=4)
>0	6016	0.035	0.066	0.814
>20	963	0.031	0.086	0.978
>50	773	0.031	0.082	0.978
>100	440	0.030	0.098	0.980
>200	220	0.023	0.055	0.990

General linear model tested on a per-gene basis (PROC GLM; model level=timepoint cell line; by gene). P values considered were calculated for each variable individually having adjusted for the variation due to remaining variables (added-last test / SAS Type III F-Test).

Table 5. *Summary of ANOVA conducted on TPA time course dataset.*

Threshold	Number of genes with expression > threshold	Proportion of genes with P-value < 0.05 when following variables were tested individually*	
		Time Point (n=4)	Cell Line (n=4)
>0	2867	0.120	0.621
>20	1327	0.094	0.502
>50	746	0.132	0.797
>100	434	0.098	0.980
>200	229	0.055	0.990

General linear model tested on a per-gene basis (PROC GLM; model level=timepoint cell line; by gene). P values considered were calculated for each variable individually having adjusted for the variation due to remaining variables (added-last test / SAS Type III F-Test).

Table 6. *Summary of ANOVA conducted within NCI60 dataset.*

Threshold	Number of genes with expression > threshold	Proportion of genes with P-value < 0.05 when following variables were tested individually*	
		Cell Line (n=60)	Tissue (n=8)
>0	7110	0.056	0.214
>20	871	0.044	0.336
>50	543	0.044	0.359
>100	350	0.040	0.380
>200	200	0.030	0.202

General linear model tested on a per-gene basis (PROC GLM; model level=cell line tissue; by gene). P values considered were calculated for each variable individually having adjusted for the variation due to remaining variables (added-last test / SAS Type III F-Test).

Table 7. *Summary of ANOVA conducted across multiple human datasets.*

Threshold	Number of genes with expression > threshold	Proportion of genes with P-value < 0.05 when following variables were tested individually*	
		Dataset (n=3)	Tissue (n=10)
>0	1097	0.066	0.814
>20	585	0.086	0.978
>50	341	0.082	0.978
>100	206	0.098	0.980
>200	139	0.055	0.990

*General linear model tested on a per-gene basis (PROC GLM; model level=dataset tissue; by gene). P values considered were calculated for each variable individually having adjusted for the variation due to remaining variables (added-last test / SAS Type III F-Test).

Table 8. *Outlier detection.*

Dataset	Number of genes detected	Total number of samples	Number of genes with at least one measurement outside +/-3SD	Percentage of genes with expression below threshold
NCI60	7129	60	3598	55.9
AML/ALL	7129	72	5695	50.6
HL60	7129	17	5093	58.5

NCI60 – expression dataset generated from the NCI panel of 60 cancer cell lines[88], AML/ALL – acute myeloid leukemia and acute lymphoblastic leukemia dataset[78], the Huge Index dataset[79], HL60 – expression data from HL60 cells treated with TPA[38].

FIGURE LEGENDS

Figure 1. Gene expression profiles of classic control genes examined across multiple datasets: NCI60 cell line panel, ALL/AML tumor panel, Huge Index and TPA Cell-line induction. Gene expression levels uniformly rescaled are plotted on the y-axis; samples (ordered according to their arrangement given in each respective study) are plotted on the x-axis. All datasets with the exception of the Huge Index were rescaled based on mean intensity per scan.

Figure 2. Replicate samples from 4 mouse tissues. RNA was extracted from the liver, heart, lung and brain of 3 adult male C57BL/6J mice. To assess technical variability, the RNA from each tissue of one mouse was divided and hybridized in replicate to 3 separate arrays. To assess biological variability, RNA from identical tissues of 3 individual mice were hybridized to 3 separate arrays. Points are arranged in the following order for each tissue: mouse1-replicate1, mouse1-replicate2, mouse1-replicate3, mouse2-replicate1, mouse3-replicate1. Multiple probe sets, present for *Glud*, *Pgk1*, *Pgam1* and *Ldh1* show consistency in measurements of expression levels across tissues. Other probe sets for *Tuba1*, *Vim*, and *B2m* show a higher degree of variability indicating issues inherent in probe design. Samples were normalized by mean intensity per scan.

Figure 3. Genes identified by heuristic filtering: Expression profiles for genes of low variability across 4 datasets. Genes were sorted based on degree of variability. Analysis revealed that in any given dataset, gene sets could be identified with less variability in gene expression however the composition of such sets did not remain consistent.

Figure 4. Normal-quantile plot of P-values from ANOVA conducted gene-by-gene on the mouse replicate dataset taking into account variation among technical replicates, individual mice and 4 separate tissues (brain, heart, liver and lung). P-values are those derived from the added-last F-test (Type III GLM in SAS). This tests the contribution of each variable individually towards explaining the observed variation in a given gene's expression level across all samples. Typically genes had lower P-values for tissue indicating this variable is more significant, having adjusted for the variation due to mouse and replicate variables. This effect remains consistent with different data filtering criterion. Data was centered using mean intensity for each scan.

Figure 5A. Rescaling factors across NCI60 and ALL/AML datasets. In order to assess the influence of normalization techniques on the analyses, we compare rescaling factors calculated by various approaches including global mean, median, regression factors, and filtered subsets of genes. Rescaling factors are given for mean and median methods as calculated on filtered subsets of the data (genes with average difference levels greater than 20 and 50). All methods show a similar level variability.

Figure 5B. Effect of rescaling by different methods on the average expression level of traditional housekeeping genes in NCI60, leukemia training and independent datasets. Traditional control genes analyzed included: beta-actin (*ACTB*), beta-2-microglobulin (*B2M*), phosphofructokinase (*PFKP*), phosphoglycerate kinase (*PGK1*), aldolase A (*ALDOA*), phosphoglycerate mutase (*PGAM*), alpha-tubulin (*TUBA*), glyceraldehyde-3 phosphate dehydrogenase (*GAPD*), glucose-6 phosphate dehydrogenase (*G6PD*), lactate dehydrogenase A (*LDHA*), hypoxanthine phosphoribosyltransferase (*HPRT*), and vimentin (*VIM*). Using different normalization methods does not appear to have a significant impact on the observed level of expression variability.

Figure 1.

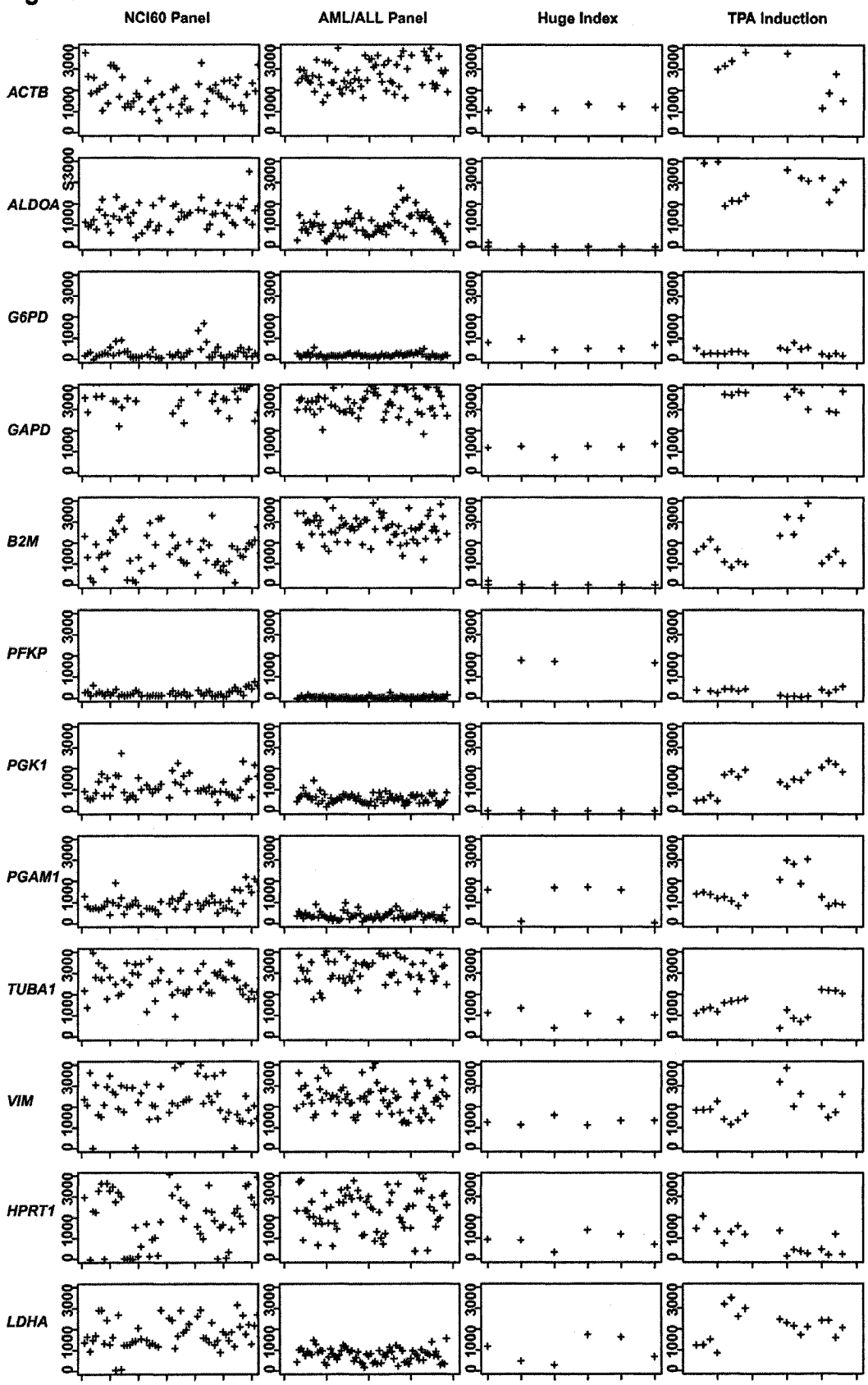

NCI60 Panel
AML/ALL Panel
Huge Index
TPA Induction
ACTB
ALDOA
G6PD
GAPD
B2M
PFKP
PGK1
PGAM1
TUBA1
VIM
HPRT1
LDHA
0
1000
3000

Figure 2.

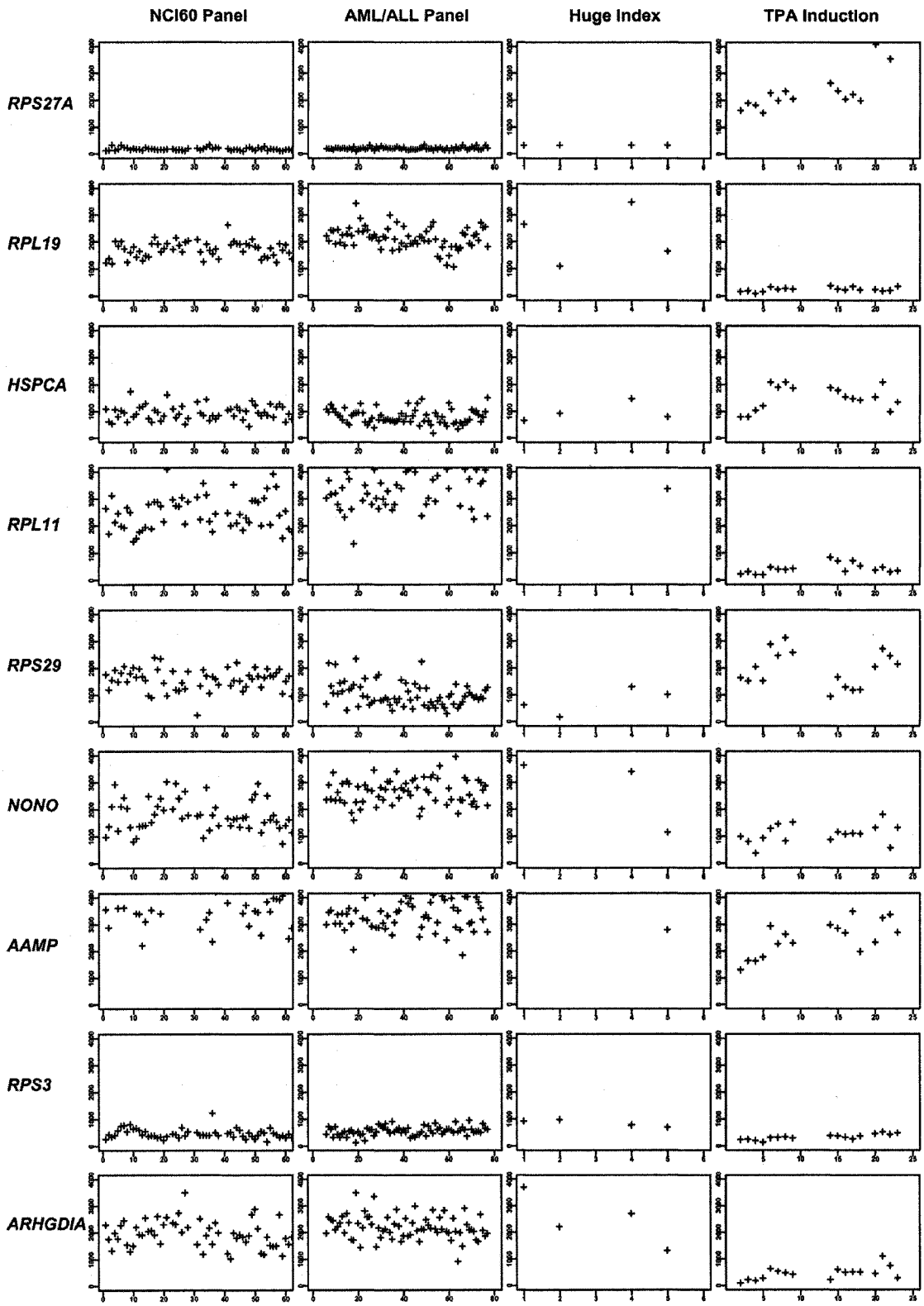

NCI60 Panel
AML/ALL Panel
Huge Index
TPA Induction
RPS27A
RPL19
HSPCA
RPL11
RPS29
NONO
AAMP
RPS3
ARHGDIA

Figure 3.

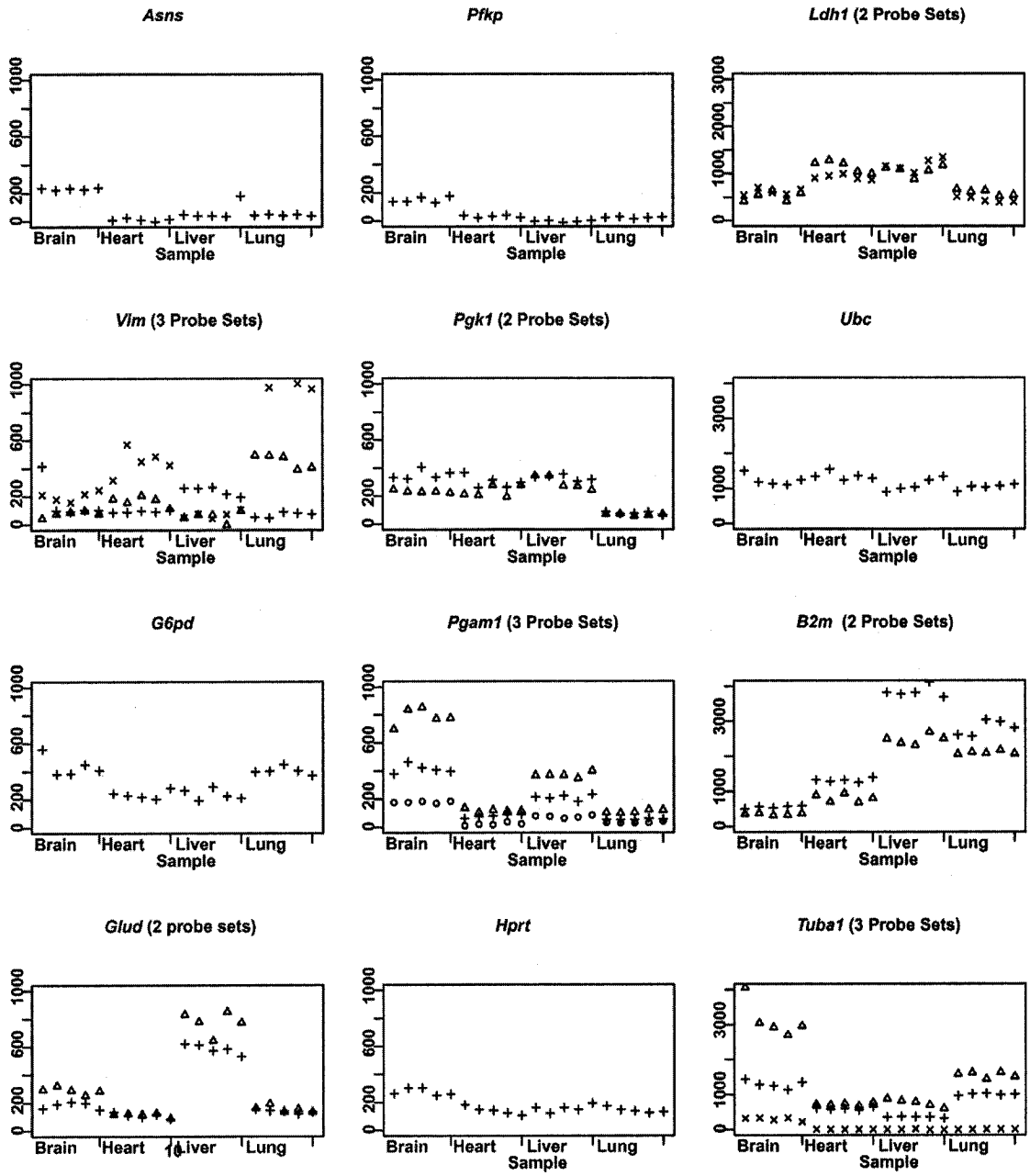

Asns
Pfkp
Ldh1 (2 Probe Sets)
Vim (3 Probe Sets)
Pgk1 (2 Probe Sets)
Ubc
G6pd
Pgam1 (3 Probe Sets)
B2m (2 Probe Sets)
Glud (2 probe sets)
Hprt
Tuba1 (3 Probe Sets)
Brain
Heart
Liver
Lung
Sample

Figure 4.

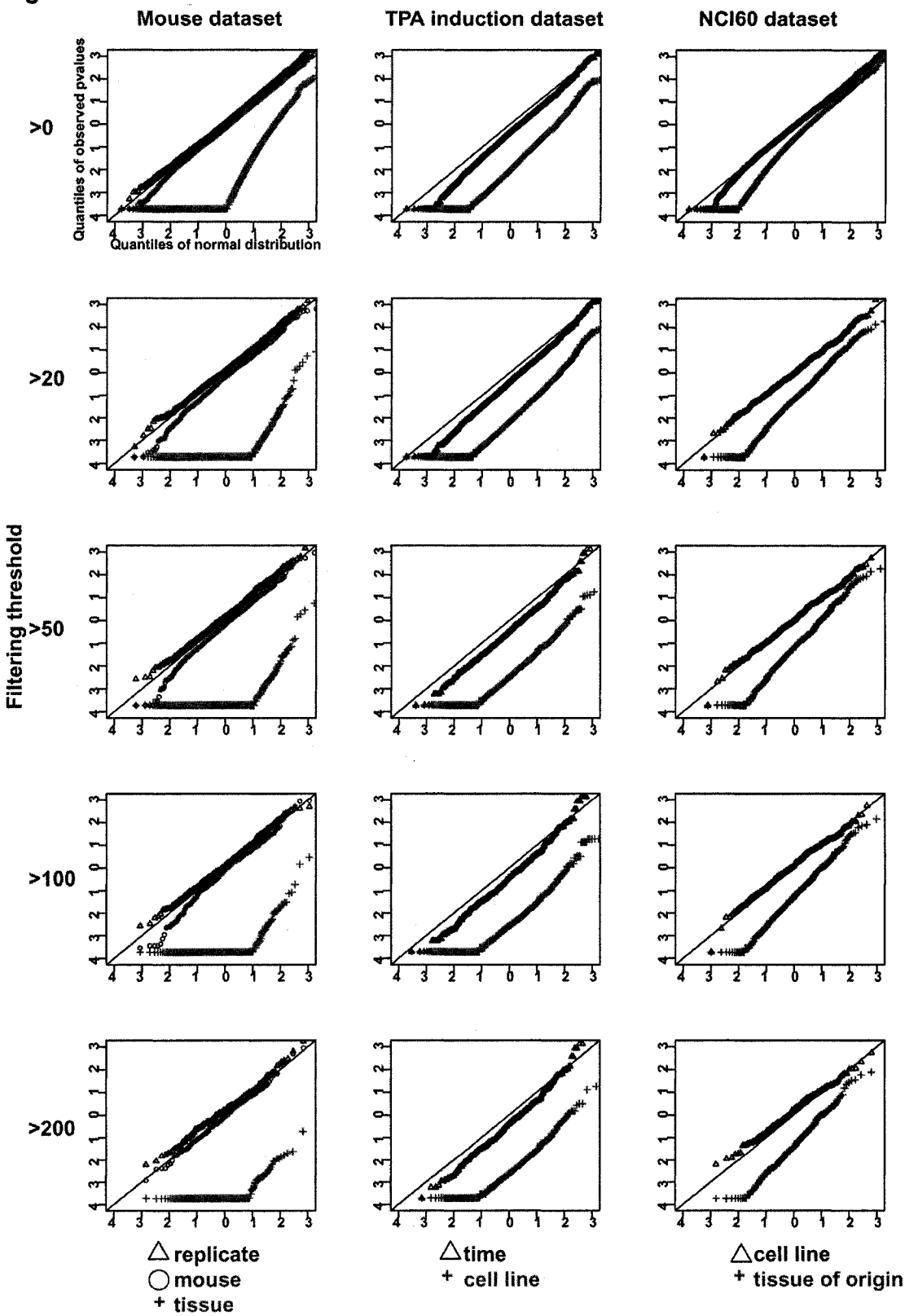

Mouse dataset
TPA induction dataset
NCI60 dataset
Quantiles of observed pvalues
Quantiles of normal distribution
Filtering threshold
>0
>20
>50
>100
>200
△replicate
○mouse
+ tissue
△time
+ cell line
△cell line
+ tissue of origin

Figure 5A.

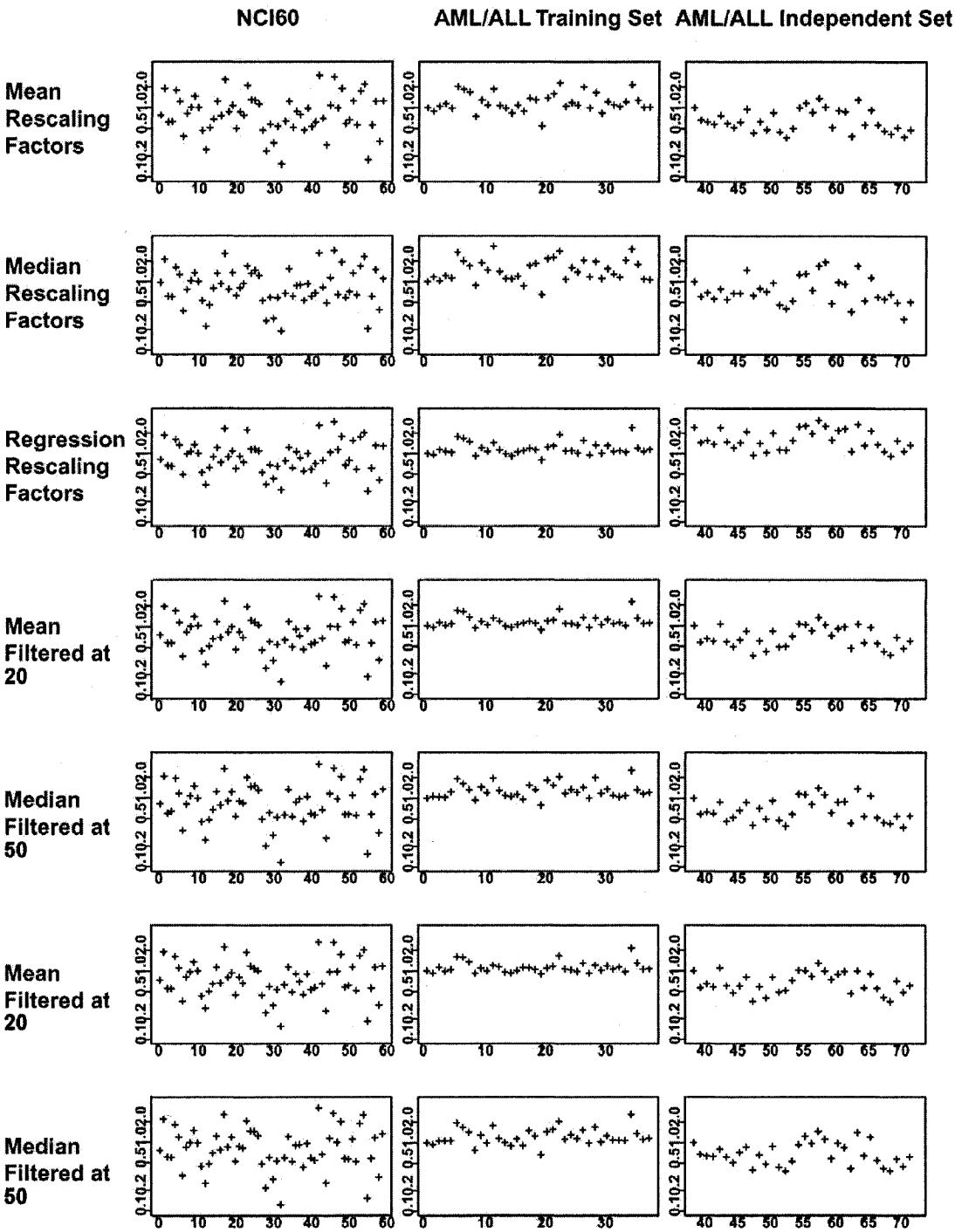
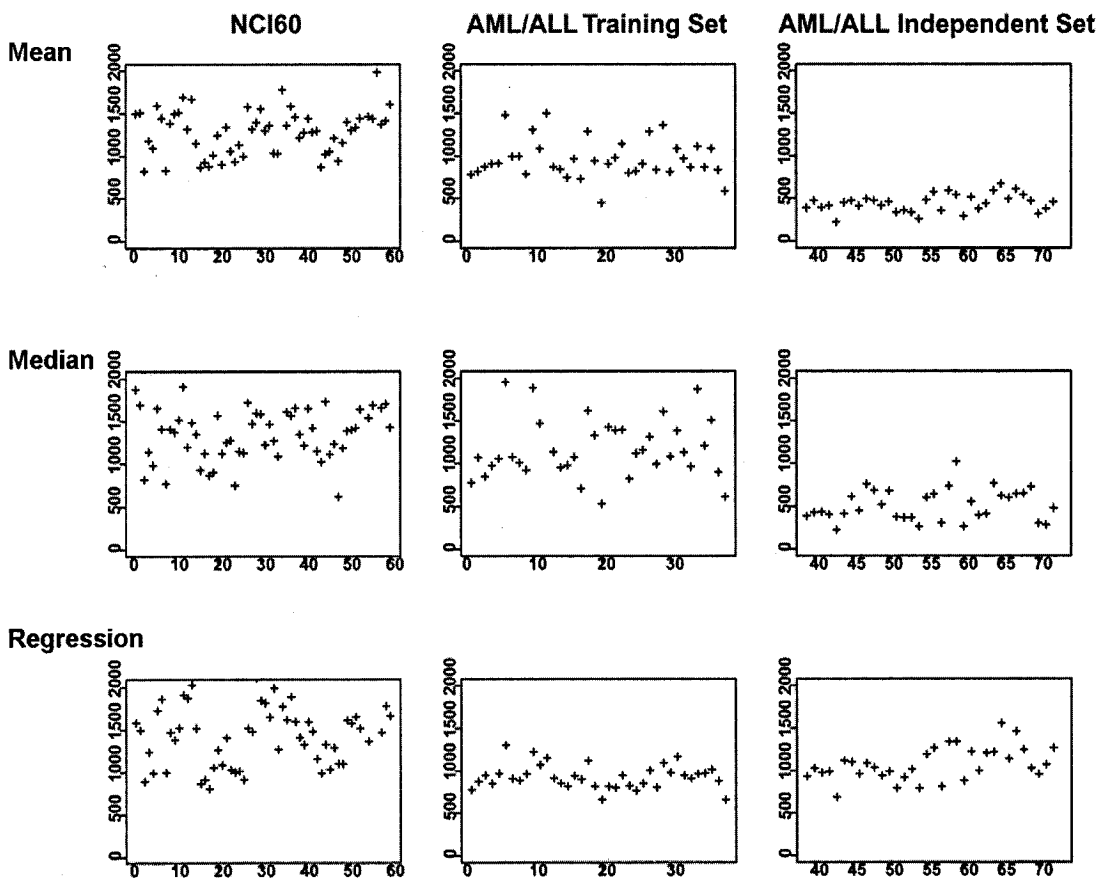

NCI60
AML/ALL Training Set
AML/ALL Independent Set
Mean Rescaling Factors
Median Rescaling Factors
Regression Rescaling Factors
Mean Filtered at 20
Median Filtered at 50
Mean Filtered at 20
Median Filtered at 50

Figure 5B.


NCI60
AML/ALL Training Set
AML/ALL Independent Set
Mean
Median
Regression

CHAPTER 2 – APPLICATION TO EXPERIMENTAL SYSTEMS: GENE EXPRESSION ANALYSIS OF INBRED MOUSE STRAINS

The connection between gene expression and cellular/organismal phenotypes has long been recognized. Early work in mice and flies determined that changes in gene transcription were critical to an organism's development.[89] In these studies, mutations in transcriptional networks disrupted normal limb development, organogenesis, and other developmental processes. Indeed embryonic development has been described in part as the timed activation and repression of gene transcription. Microarrays offer the opportunity to characterize organisms at the level of gene expression. The questions remain, how are expression changes are related to observed differences at the organismal level, and, how does gene expression differ between organisms of differing genetic background?

The second part of this thesis focused on gene expression analysis of inbred mouse strains. Collected by hobbyists around the turn of the century for their coat colours, inbred strains were found to differ dramatically for a wide range of phenotypes[60]. These strains have become a commonly used model system in which to study the genetic components of phenotypes pertaining to human disease. Since then, the genealogy, nomenclature and breeding of mouse strains has been standardized, catalogued and maintained due to the centralization of efforts at the Jackson Laboratories. This study focused on two widely studied strains: A/J and C57BL/6J. A/J (albino) mice have a known propensity for a range of disease phenotypes including low breeding rate, asthma,

susceptibility to cancer and infectious diseases. C57BL/6J are considered normal for many phenotypes and have been used as a standard for comparisons against many of the inbred strains. These comparisons have formed the basis for numerous studies leading to the identification many genetic loci. At the molecular level, the strains have been fully sequenced and are known to differ for a multitude of genetic markers[53]. However, little remains known of how these strains differ at the level of gene expression.

Many phenotypes studied between strains involve some form of treatment. For example, studies of atherosclerosis and obesity often involve administration of diets with varying levels of cholesterol[90-92]. Susceptibility to infection is tested by inoculation with pathogens[93-96], and cancer predisposition has been measured as tumor count following administration of known carcinogens[97, 98]. While most studies focus on characterizing differences in the response to stimuli, little remains known of the context in which these responses are acting, the underlying level of variability in gene expression observed between untreated animals. Baseline differences in gene expression have already been seen to have an heritable component[30] in humans. This chapter provides an analysis of expression profiles in A/J and C57BL/6J untreated adult male mice across four tissues with relevance to previously studied phenotypes. The study aimed to characterize differences in molecular context between these two strains at the level of gene expression.

In the time between this and the previous chapter, the field of microarray analysis advanced considerably. Methods for summarizing probe level data and normalization

techniques evolved from heuristic convention-based approaches to methods based on statistical modeling of the data. Summary statistics are a concern specific to the Affymetrix technology and refer to the methods used to derive a single measure of gene expression from hybridization signals obtained over all probes in a probeset. Normalization refers to methods used to correct for fluctuations in overall hybridization efficiency between arrays. These methods are often referred to as data preprocessing steps since they precede analysis of differential gene expression. An assortment of methods was developed for both tasks within a short period of time. Most significantly, open access to these tools via the Bioconductor project (www.bioconductor.org), a package developed for the R statistical package (www.r-project.org) offered a dramatic improvement in the ability to analyze microarray data. Data preprocessing has been seen to greatly influence the results of subsequent analyses detecting differentially-expressed genes[99]. A portion of the study therefore compared summary and normalization methods available at the time in order to examine the impact of data preprocessing on analysis, and to determine the optimal combination of techniques.

This study furthermore investigated reproducibility by comparing identical experiments conducted 1 year apart, an experimental condition that has been neglected in most microarray studies. A multitude of factors are known to influence the observed variability in a microarray dataset; tissue dissections, RNA preparation steps, changes in personnel. Overall levels of variability due to these sources have not been addressed in traditional RNA measurement assays such as RT-PCR and Northern blots. However, with microarray measurements, statistical determination of differential expression hinges

critically upon the ability to assess the relative contribution of different sources of variability. While inter-experimental variability has been questioned for quite some time, quantitative estimates of the magnitude of the effect remained unaddressed. This study provided a direct comparison of two repeat experiments conducted one year apart, replicating all conditions to the best of abilities. Statistical analysis demonstrated a substantial degree of inter-experimental variation, underlining the necessity for analytical methods to correct for such variability together with an adequate number of replicates. The study furthermore estimated the degree of inter-experimental variability potentially due to biological causes, particularly differences in tissue specific variability, indicating experiments in certain biological systems may exhibit differing sensitivities to environmental variables over time. This study is the first in the field to address these questions.

TISSUE-SPECIFIC DIFFERENCES IN BASAL GENE EXPRESSION BETWEEN A/J AND C57BL/6J INBRED MOUSE STRAINS.

Peter D Lee[1,2,3,4], Robert Sladek[1,2,3], Celia M T Greenwood[5,6], Bing Ge[1,2], Emil Skamene[2,3], Thomas J Hudson[1,2,3§]

[1]McGill University and Genome Quebec Innovation Centre, [2]Research Institute of the McGill University Health Centre, [3]Department of Human Genetics, Faculty of Medicine, McGill University, [4]McGill Centre for Bioinformatics, [5]Program in Genetics and Genomic Biology, Hospital for Sick Children, [6]Department of Public Health Sciences, University of Toronto.

[§]Corresponding author

740 Dr. Penfield Ave, Montreal, QC, H3A 1A4

Tel: (514) 398-3311 ext 00385

Fax: 514-398-2622

tom.hudson@mcgill.ca

Submitted for review to Physiological Genomics, January 2005.

ABSTRACT

Inbred mouse strains A/J and C57BL/6J have long been used to study the genetic basis of disease and are known to differ for a wide variety of simple and complex phenotypes. Despite the large number of loci mapped between these strains, the genetic variants underpinning the majority of these phenotypes remain to be characterized. Here we compare gene expression profiles across 4 tissues of untreated A/J and C57BL/6J mice in order to describe the diversity of these inbred strains at the level of gene expression. We report the levels of inter-experimental variability, demonstrate the effect of analysis methods on the reproducibility of results and provide approaches to adjust for such variability. We also address issues of consistency of our results over time by repeating the experiment. Our results show that when inter-experimental variability is accounted for, 853, 459, 652 and 1229 genes vary significantly between parental strains in heart, liver, lung and spleen respectively ($P<0.01$), having corrected for the effect due to the time the experiment was performed (n=6 replicates). Comparisons among tissues show moderate degrees of overlap with between 9% and 25% of genes being differentially expressed within a given tissue, and with 19 genes being differentially expressed across all tissues. We provide several comparisons of differentially expressed genes located within support intervals of previously mapped quantitative trait loci (QTLs) to illustrate possible applications for prioritizing disease gene candidates underlying genetic loci mapped for complex traits.

INTRODUCTION

Inbred mouse strains have been used for nearly a century as model systems for human disease and to study the genetic basis of complex phenotypes. These strains have been seen to offer the advantage of a confined experimental system whereby associations between phenotypic and genomic sequence variation may be studied in an environment that reduces additional sources of variability. Despite the many physiologic variables catalogued for these strains, little is known about the nature of the differences between inbred strains at the molecular level *in vivo*. One obstacle facing such studies is that for the majority of traits, a multitude of genetic and epigenetic factors interact to cause the observed phenotype. This complexity presents numerous difficulties to fine mapping efforts that attempt to identify genes underlying previously mapped quantitative trait loci (QTL). Compounding these difficulties is the fact that QTL often span large regions of the genome encompassing hundreds or thousands of genes. New approaches are needed to complement mapping studies in order to accelerate the identification of candidate genes[9].

Several directions of research have been proposed to address this challenge. Recently, the diversity between inbred strains has been more finely characterized, improving the information that may be exploited in the design of QTL mapping studies[56, 58, 100]. In addition to increasing the resolution for detecting genetic differences between strains, recent studies suggest that gene expression profiling may provide a complementary approach for prioritizing candidate genes in the search for the molecular determinants of

complex traits[67, 101-103]. It is known that variation in genomic DNA sequence causes genome-wide changes in gene expression levels and that these changes are detectable between populations of individuals[31, 34, 104]. These studies highlight the utility of combining genomic data with expression profiling in order to better understand the molecular biology of complex traits. Another rationale for this approach lies in the multivariate nature of biological systems and the unequivocal context-dependence of the phenotypes studied. Genome-wide expression profiling offers a means to capture the molecular context in which these phenotypes may be manifested. Background or basal variability in gene expression may affect phenotypic measurements especially where treatments are involved that may alter gene regulatory pathways. While interpretation of expression profiles alone presents numerous challenges, methods that measure expression profiles across multiple tissues, strains or species have proven fruitful in identifying genes with related function[70]. Approaches that combine genomic sequence data with expression profiling provide a means to cross-validate experimental results and may advance the study of genetic regulatory mechanisms on a genome-wide scale[30].

Here were present expression profiles of untreated A/J and C57BL/6J mice as an initial sketch of the differences between the two strains at the molecular level. A/J and C57BL/6J are among the most widely used inbred mouse strains in medical research and are known to differ for a wide spectrum of quantitative physiologic traits including diabetes and obesity, cancer, atherosclerosis, asthma, pain sensitivity, alcoholism, as well as host response to a variety of infectious agents[105]. Genetically, these two strains are known to differ at over 3200 microsatellite markers[61, 62] and for at least 120,000 SNPs[63].

These differences have facilitated the mapping of numerous quantitative trait loci (reviewed in Table 1). However, because of the aforementioned challenges inherent in the characterization of these traits[64], few candidate genes have been identified for these traits.

We present expression profiles for 4 tissues (heart, liver, lung, and spleen) with relevance to previously studied phenotypes. We previously studied the expression variability between tissues within individual C57BL/6J mice and show that differences in expression between tissues, within a strain, may be assessed reproducibly[106]. Here we examine differences between strains. Previous microarray gene expression profiling in C57BL/6J mice showed levels of gene expression variability observed between individual mice to be similar to that between technical replicates (identical RNA preparations used for distinct array hybridizations) with the greatest variability occurring between tissues[107]. We compare the variance observed between strains, tissues, and repeat performances of the experiment as well as the effects of various analytical techniques. The differences in gene expression profiles detected between strains are contrasted between tissues and we provide examples where these differences correspond with QTL previously identified in A/J and C57BL/6J. This study provides an illustration of the variability to be expected when applying gene expression profiling to inbred mouse strains and we discuss approaches to deal with this variability by appropriate analysis and experimental design.

MATERIALS AND METHODS

RNA preparation. Mouse tissues (liver, lung, spleen, heart tissue) were harvested from three adult male littermates from each of A/J and C57BL/6J strains. The mice were sacrificed at three months of age by cervical dislocation and the tissues rapidly dissected and homogenized in Trizol reagent. Total cellular RNA was prepared according to the manufacturer's instructions and analyzed by gel electrophoresis. The procedure was repeated 12 months later using mice of the same strains obtained from Jackson Laboratories and sacrificed at the same age.

Microarray hybridizations. Biotinylated probes for the microarray studies were prepared by using 20 μg of total RNA. Hybridization was performed overnight at 45°C using 15μg of biotinylated probe. Following hybridization, the arrays were processed using a GeneChip Fluidics Station 400 (Affymetrix). The experimental protocol has previously been described in detail[108]. In each case, a single chip hybridization was performed for each RNA sample.

Microarray data analysis. Data analysis was accomplished using Perl and Bioconductor in R (www.bioconductor.org). We applied a general linear model on a gene-by-gene basis to determine the effect of strain on gene expression correcting for the effect due to replicate mice (separate samples from three identical mice), and time, (the year that the hybridizations were performed): p values were calculated for each variable individually having adjusted for the variation due to remaining variables. We used a threshold of

$P<0.05$ for significance of differential expression. In order to evaluate the effect of analytical methods on our results, we compared quantile normalization with invariant set and mean-based scaling, as well as comparing RMA estimates of gene expression[109] with the average difference (Microarray Analysis Suite 5.0, Affymetrix) and Li-Wong (Dchip) summary methods[110]. Final results are based on RMA summaries with quantile normalization, since this method gave comparatively the most uniform variance dynamic range of hybridization signal. The dataset has been submitted to the NCBI Gene Expression Omnibus (Series number GSE2148).

Initial analysis used the t-test to compare expression between strains, in each of the two experiments. Initially genes in each tissue were tested separately. An ANOVA model was applied to adjust for inter-individual variability. ANOVA and t-tests were conducted in R on a gene-by-gene basis. Upon observing the inter-individual variability to be a minor component, we tested additional ANOVA models including expression=strain+time, testing interaction terms, expression=strain+time+strain*time. We also compared the contribution of tissue to the dataset testing models expression=strain+time+tissue, and a model containing all possible interaction terms.

Physical localization of genes and comparison with QTL results. Probe sequences from the Affymetrix MU74Av2 oligonucleotide array were aligned against the Feb 2003 build of the NCBI mouse genome using BLAT with default parameters[111]. Of the 12422 probe sequences on the chip, 11105 were localized reliably. Physical positions corresponding to previously described QTL were determined by aligning marker sequences to the same

assembly using BLAST and retrieving all primer matches with intervening sequences less than 500 bp (wordsize=12, p=.90, e=0.1) and iteratively relaxing matching parameters in the event of no match. This procedure matched 90% of the markers from QTL studies. Support intervals for each QTL were estimated as the region with a LOD score greater than 3 on either side of the peak marker. This generated intervals ranging in length from 1.5cM to greater than 60 cM. If flanking marker information was not available, the QTL support intervals were estimated as 40 Mb in either direction from the peak marker. Supplementary data, figures, tables, and analytical scripts are available at http://www.mcb.mcgill.ca/~pdlee/AxB.

RESULTS

Since there are many reasons why *in vivo* gene expression profiling experiments at one time could give different results; (for example, litter effects, changes in handling, new reagents, or different chip lots), we repeated expression profiling at two time points 12 months apart. We initially analyzed each dataset separately before doing a combined analysis, in addition to comparing several combinations of summary statistics (Figure 1A, Table 2). In these analysis we reconfirm that the level of variability between littermates is very close to that of technical replicates[106]. A comparison of observed P-values with calculation of expression differences between strains confirms that statistically significant differences between strains do not necessarily correlate with large fold changes (Figure 1B). However, because of the substantial contribution of time to the observed gene expression variability, we analyzed the two datasets together using an ANOVA that included the time of experiment as a factor (results available at http://www.mcb.mcgill.ca/~pdlee/AxB/). Among genes detected in separate analysis of time points, expression differences between strains exhibit varying directional fold changes (Figure 1C), further indicating the potential for time to confound effects due to strain (Figure 1D). Whereas separate analyses of the time points detects between 121 and 577 genes per tissue (intersections between separate analyses), 459 to 1229 genes display expression differences ($P<0.01$) when a single model is used, correcting for time of the analysis (Table 3). The proportion of genes displaying significant time-dependent effects ($P<0.01$) ranged from 26% to 41% of probesets for each tissue (3545 in heart, 3396 in lung, 5145 in liver, 3250 in spleen). Among the intersections between tissues (Table 4), gene expression in A/J exceed C57BL/6J more often with 77% to 100% of genes

displaying conserved directionality of expression differences between strains in the overlapping sets. Of the total 3193 differentially expressed genes identified, 2109 are differentially expressed exclusively in one tissue indicating a strong tissue-dependent effect. The two tissues with the most number of genes in common that are differentially expressed between strains were heart and spleen (170 genes) while liver and lung have the least number in common (70 genes) (Figure 2A). Only 19 genes display differential expression between strains across all 4 tissues (Table 5, Figure 2B). To further investigate these observations, we conducted an ANOVA over the entire dataset, tissue as a term in the model. These analyses reveal the extent of variability due to tissue to be similar to that due to time (Figure 3), and show the degree of interaction between terms in the model (Figure 4).

In order to investigate the physiological significance of our results, we compared the locations of differentially expressed genes within over 20 previously mapped QTL from an exhaustive search for studies comparing A/J versus C57BL/6J strains (Table 1). We aligned marker and probe sequences against the mouse genome assembly from NCBI using BLAT[111]. Of the 2109 genes differentially expressed in only one tissue, 1931 had chromosomal locations that could be assigned using Feb 2003 assembly of the mouse genome and were distributed across all chromosomes. We present the results of a comparison of our lung data (Figures 5 and 6, Table 6) against QTL for asthma, acute lung injury and lung cancer phenotypes. Of the more than 3400 transcripts on the array that located within these previously mapped QTL intervals, we observe in excess of 750 differentially expressed genes between strains.

DISCUSSION

A/J and C57BL/6J are two of the most widely used inbred mouse strains in genetic research. A/J is a subline of the A strain that was generated by Strong in 1921 and distributed to the Jackson Laboratory in 1947. C57BL/6J originated with Little in 1921 and is a widely used strain as genetic background for mutation and cross breeding experiments[60]. In direct comparisons, the two strains have been characterized for hundreds of phenotypes (www.jax.org/phenome). In this study we report differences in basal gene expression between wild-type A/J and C57BL/6J mice of similar age and sex across key tissues of relevance to phenotypes observed at the physiological level. We present comprehensive results testing for consistency and reproducibility, correcting for the extent of inter-individual variability and for the variability due to time the experiment was performed. In spite of our best attempts to insure accuracy and reproducibility of our analysis, the complex nature of microarray data makes resolution of these issues difficult. While assumptions of normality and homoscedascity in microarray data have been evaluated with respect to identifying the adequate number of replicates[112], the expected collinearity between genes, as well as the potential nonlinear response of hybridization signals to experimental factors warrant caution in the interpretation of the results. Statistical thresholds for significance were chosen to permit a broad comparison of strain-specific differences. While this may result in a greater occurrence of false positives, we propose comparisons with previously determined QTL as a method of cross-validation. However, given the possibility of spurious correlations due to multiple hypothesis testing[113], we advise that putative candidate genes identified in these lists be subsequently validated in appropriate samples and models.

Concerns of reproducibility in microarray analysis have led to many approaches to confirm results including cross-validation using RT-PCR or Northern blots[43]. To address issues of reproducibility we chose to repeat the experiment, duplicating to the best of our abilities all experimental conditions. Our results show that the amount of variability between experiments is substantial. There are several possible explanations for these results. Technical factors included relocation of the laboratory, recalibrating the scanner, chip lot, reagents and changes in the team performing RNA extractions and hybridizations. The choice of analytical method and the number of sample replicates are known to jointly affect consistency of results; in particular, estimates of standard deviation based on only 3 replicates are not stable, and could easily vary substantially between the first and second experiments. However, within the intersection of gene sets obtained from separate analysis of the two experiments, the direction of gene expression changes between strains was well preserved (77-100% of genes) between replicated experiments for all tissues (Figure 1B), indicating that such genes may reliably be measured when correcting for environmental variables. As expected, analysis of the 2 experiments together using a single model factoring for the effect due to time of analysis identified far more genes as differentially expressed in each tissue as compared to separate analyses (Tables 2 and 3). Explanations include the improved power to detect differentially-expressed genes with larger sample sizes[114] and the ability of the ANOVA to correct for the confounding effect of time on strain-specific variation. While questions persist as to the appropriateness of parametric methods applied to microarray analysis, we propose that the higher number of replicates in the combined dataset provides increased

power to detect differentially-expressed genes and that our analysis represents an accurate capture of gene expression variability in the experimental system over time.

The variability observed between experiments may furthermore be due to biological causes. While this environment replicated as many variables as possible, both handling and housing are known to affect biological variables[115]. Interestingly, an ANOVA that included the effect of time-strain interaction detected 2677 genes for which this interaction was significant at $P<0.01$ (Table 7, Figure 5). This suggests that a proportion of the time-dependent gene expression variability will vary depending on strain, and supports the hypothesis of a biological component to the time-dependent effect. Furthermore, individual tissues appear to show different sensitivity; in particular, there were over 1000 liver genes showing evidence for strain-time interaction. This agrees with previous findings[116] and suggests that liver displays the most sensitivity to inter-experimental variation, possibly as a result of experimental technique (i.e.: tissue extractions) and/or due to biological responses within the mice. These findings suggest that studies measuring molecular phenotypes must take into account the inherent dependence of each tissue on environmental factors. The number of uncharacterized potential sources of variability, both biological and technical, renders interpretation of these results difficult. We report the levels of variation to be expected between microarray experiments to inform users of this data and strongly suggest using analytical methods that adjust for such variability.

The varying degree of overlap in differential expression among tissues suggests that the majority of gene expression changes reflect tissue-specific effects and that gene regulation differences between inbred strains are largely dependent on tissue context. Differential expression between strains common to all tissues was seen for only 19 genes upon separate analysis of tissues (Table 3, Figure 2B), a result echoed when tissues were analyzed together (Figures 5 and 6). Common strain-specific expression differences between tissues may indicate involvement of these genes in steady-state conditions important to all tissues rather than transient roles in more specialized cellular processes. The results point to possible strategies for exploiting the degree of tissue similarity at the molecular level for more efficient querying of the biological system. Overlapping molecular signatures may underlie the pleiotropy observed for many phenotypes and which may be testable using informative combinations of tissues. Our inability to find a large amount of strain-specific variability common to all tissues suggests that distinct subnetworks regulate gene expression in the different tissues.

We believe our estimates of gene expression variability between inbred strains represents a lower limit for several reasons; First, while the experiment was repeated at two time points in adult male mice, transcriptional switches are known to regulate many of the transition points in growth and development. Therefore a comparison of strains at different ages will likely expose more gene expression differences. Second, sex-dependent differences between genetically divergent strains, which were not studied in our experiments, have been observed to be substantial in other model systems[117]. Third, an analysis of basal gene expression for a wider range of tissues may expose far more

differences in transcriptional activity resulting from differences in genetic sequence. Finally, while we focus on comparing two inbred strains, the diversity observed between other commonly used inbred strains[100] would suggest similar tissue-specific diversity of gene expression may be observed across a broader sampling of strains.

The application of gene expression profiling has been suggested as a means for informed prioritization of gene candidates within QTL. The tissues chosen in this study represent a subset of those that may be affected by various phenotypes characterized in these strains. In order to present examples of how the expression data results may be used by laboratories studying disease phenotypes in these strains, we compared the location of differentially expressed genes an exhaustive search of QTL support intervals previously mapped in AxB comparisons. We observe interesting correlations (Table 8) when comparing the spleen data for genes relevant to malaria resistance[118]. Vcam1, differentially expressed in spleen, localizes within the Char4 locus for susceptibility to *P. chabaudi* infection on chromosome 3, and is upregulated in mice infected with *P. falciparum* [119]. We furthermore detect differential expression of Pon3; this gene is located within the region of the Aliq4 locus for acute lung injury on chromosome 6[120]. Paraoxonases are implicated in response to oxidative stress[121] and have been implicated in a variety of other disease[122]. Other examples include Kras1, a putative candidate within the Pas1 locus for predisposition to lung adenoma[123, 124], Anxa2, Casp2, Ctsc and Gpx3 implicated in the pathogenesis of asthma[125-128], and Infgr, Il6ra and Csf3r detected in spleen and implicated in resistance to Listeria infection[129]. Our study detects differentially regulated genes that have been identified in previous studies[130]. However

our ability to detect more differentially expressed candidates within lung QTL most likely rests upon our increased number of replicates (n=12), as well as the sensitivity offered by combined application of RMA and ANOVA methods. While such candidates remain to be tested experimentally, the coincidence of our results with previous studies illustrates the potential utility of gene expression analysis in the prioritization of disease gene candidates within previously mapped loci. However, the large number of genes within QTLs suggests that microarray expression profiling alone will not be sufficient and that a more in-depth of analysis of the relationship between specific genetic variability and expression differences shall be required.

While polymorphisms in coding sequences have been identified underlying complex traits, recent studies indicate that regulatory polymorphisms may be more prevalent than previously anticipated, and that gene expression profiling provides the opportunity to detect such mechanisms on a genome-wide scale[29, 104, 131]. Our results lend support to the theory that phenotypic differences between A/J and C57BL/6J may in part be due to differences in gene regulation. Evidence from other studies highlight this prevalence of regulatory variants in complex traits[132] and demonstrate the utility of gene expression in uncovering genetic regulatory mechanisms using inbred mouse strains[67, 133-135]. Similar comparisons of gene expression profiles between inbred strains, such as C57BL6 versus 129SvEv in brain[136], comparison of transgenic strains[137], and thymic gene expression profiles in NOD strains[67] confirm the utility of the approach for studying complex traits. Further studies attribute the diversity of gene expression between populations to regulatory polymorphisms[34] with allele specific variation in gene expression dependent

upon tissue context[135]. The genome-wide expression changes observed in this study, many of which lie within previously identified QTL, suggest a more thorough investigation of the association between specific genetic variations and gene expression may better elucidate the molecular mechanisms of complex traits. An investigation is currently underway to determine the relationship between genetic variation, gene expression and gene regulation in a panel of recombinant congenic strains.

Table 1. *QTL previously mapped in A/J versus C57BL/6J comparisons.*

Trait	Tissue	LOD	Chr	Marker	Locus	Reference
TB susceptibility	Lung					[138]
Cocaine locomotor activation	Brain	5.71	12	D5Mit32	Lapls2-10	[139]
Cocaine locomotor activation	Brain	4.72	15	D15Mit105	Pdgfb	[139]
Airway responsiveness	Lung	3.1	2	D2Mit409	Bhr1	[140]
Airway responsiveness	Lung	3.8	15	D15Mit107	Bhr2	[140]
Airway responsiveness	Lung	2.9	17	D17Mit26	Bhr3	[140]
Nitrosurea-induced lung adenoma	Lung		6		Pas1	[123]
Nitrosurea-induced lung adenoma	Lung		17		Pas2	[123]
Nitrosurea-induced lung adenoma	Lung	7	19		Pas3	[123]
Legionella susceptibility	Macrophages	9.9	13	D13Mit146	Lgn1	[141]
PKC activity	Lung	3.4	3	D3Mit19		[142]
PKC activity	Lung	2.7	11	D11Mit333		[142]
Lung adenoma	Lung	6.46	4			[143]
Malaria susceptibility	RBC	6.57	3	D3Mit109	Char4	[118]
Malaria susceptibility	RBC	2.53	10	D10Mit189		[118]
Fear-like behaviour	Brain	7.7	1	D1Mit144		[144]
Fear-like behaviour	Brain	9.3	10	D10Mit237		[144]
Fear-like behaviour	Brain	3.95	19	D19Mit86		[144]
Fear-like behaviour	Brain	3.48	14	D14Mit133		[144]
Fear-like behaviour	Brain	2.77	6	D6mit86		[144]
Fear-like behaviour	Brain	2.84	X	DXMit172		[144]
Alcohol locomotor activation	Brain	3.02	16	D16Mit47		[145]
Alcohol locomotor activation	Brain	3.36	18	Iapls3-7		[145]

Trait	Tissue	LOD	Chr	Marker	Locus	Reference
Alcohol preference	Brain	3.5	2	D2Mit74		[146]
Alcohol preference	Brain		4	D4Mit172		[146]
Alcohol preference	Brain		7	D7Mit31		[146]
Alcohol preference	Brain		11	D11Mit35	Alcp2	[146]
Helicobacter hepatic inflammation	Liver		19			[147]
Listeria susceptibility	Spleen					[129]
Age-related hearing loss	Nervous	70	10	D10Mit112		[148]
Obesity	White adipose	8.59	2	D2Mit66	Itg	[149]
Obesity	White adipose	4.98	3	D3Mit10	Ngfb	[149]
Obesity	White adipose	12.54	8	D8Mit128		[149]
Obesity	WAT	3.63	19	D19Mit86	Rln	[149]
Axonal regeneration	Nervous					[150]
Urethane-induced adenomas	Lung		6		Pas1	[151]
LPS response	Spleen	7.2	1	D1Mit132	Mol2	[152]
LPS response	Spleen		7	D7Mit155	Mol1	[152]
LPS response	Spleen	6.5	11	D11Mit299	Mol4	[152]
LPS response	Spleen	8	13	D13Mit	Mol3	[152]
LPS response	Liver	2.76	5	D5Mit233	Hpi1	[153]
LPS response	Liver	4.8	13	D13Mit88	Hpi2	[153]
Benzodiazepine sensitivity	Brain		1	Xmv-41		[154]
Benzodiazepine sensitivity	Brain		10	D10Mit2		[154]
Benzodiazepine sensitivity	Brain		15	D15Mit5		[154]
Sensitivity to Inflammation	Brain					[155]
Resistance to endotoxin	Lymphoid					[156]
Lung injury ozone	Lung	6.8	11	D11Mit289	Ngfr	[120]

Trait	Tissue	LOD	Chr	Marker	Locus	Reference
Lung injury ozone	Lung	6.8	11	D11Mit179	Ali1	[157]
Lung injury ozone	Lung		13		Ali2	[157]
Lung injury ozone	Lung	4.3	17	D17Mit2	Ali3	[157]
Lung injury nickel	Lung	2.58	1	D1Mit213		[158]
Lung injury nickel	Lung	3.02	6	D6Mit185	Ali4	[158]
Lung injury nickel	Lung	2.33	12	D12Mit112		[158]
Hormone-induced ovulation rate	Ovary	2.084	2	D2Mit433	Oriq2	[159]
Hormone-induced ovulation rate	Ovary	1.979	6	D6Mit316	Oriq3	[159]
Hormone-induced ovulation rate	Ovary	2.228	9	D9Mit4	Oriq1	[159]
Hormone-induced ovulation rate	Ovary	2.743	X	DXmit22	Oriq4	[159]
Hormone-induced ovulation rate	Ovary		7			[159]
Hormone-induced ovulation rate	Ovary		10			[159]
Blood pressure	Cardiovascular		1	D1Mit334	Bpq1	[160]
Blood pressure	Cardiovascular		1	D1Mit14	Bpq2	[160]
Blood pressure	Cardiovascular		4	D4Mit164	Bpq3	[160]
Blood pressure	Cardiovascular		5	D5Mit31	Bpq4	[160]
Blood pressure	Cardiovascular		6	D6Mit15	Bpq5	[160]
Blood pressure	Cardiovascular		15	D15Mit152	Bpq6	[160]
Hyperglycemia, hyperinsulinemia	Pancreatic					[161]
Diet-induced obesity	Adipose					[90]
Histoplasma capsulatum resistance	Lung					[162]

Table 2. *Number of genes found to be differentially expressed (P<0.05) via separate analysis of old and new datasets and comparing RMA with Dchip (intersections - ∩ between datasets are italicized).*

Tissue	Method	Year 1 t-test P<0.05	Year 2 t-test P<0.05	T-test Year1 ∩ Year2	Year 1 ANOVA P<0.05	Year 2 ANOVA P<0.05	ANOVA Year1 ∩ Year2	T-test ∩ ANOVA Year1/ Year2
Heart	RMA	1723	1492	*456*	1716	1369	*420*	*1402/ 1154*
	Dchip	1800	1298	*403*	1783	1251	*381*	*1459/ 1056*
	∩	*992*	*732*	*230*	*955*	*659*	*301*	*190*
Liver	RMA	1118	2547	*357*	1117	2361	*338*	*841/ 2024*
	Dchip	970	2936	*344*	983	2763	*337*	*733/ 2731*
	∩	*487*	*1468*	*171*	*459*	*1339*	*142*	*112*
Lung	RMA	1164	1518	*311*	1201	1560	*300*	*920/ 1170*
	Dchip	1100	1528	*295*	1124	1561	*270*	*855/ 1182*
	∩	*562*	*693*	*142*	*557*	*665*	*121*	*88*
Spleen	RMA	2051	2063	*518*	2034	2082	*510*	*1624/ 1688*
	Dchip	2253	2217	*577*	2261	2333	*569*	*1794/ 1888*
	∩	*1047*	*1300*	*263*	*1026*	*1293*	*239*	*180*

Table 3. *Number of genes identified by ANOVA correcting for the time.*

Tissue	P<0.01 Strain P>0.01 Time	P<0.01 Strain P<0.01 Time	P>0.01 Strain P<0.01 Time
Heart	853	475	3070
Liver	459	449	4696
Lung	652	336	3058
Spleen	1229	453	2797

ANOVA model tested: Expression ~ Time + Strain

Table 4. *Differentially expressed genes in common between tissues.*

	Heart	Liver	Lung	Spleen
Heart	853	93	141	170
	(312/541)	(72/21)	(67/74)	(101/69)
Liver	93	459	70	116
	(51/42)	(265/194)	(41/29)	(70/46)
Lung	141	70	652	160
	(66/75)	(41/29)	(323/329)	(95/65)
Spleen	170	116	160	1229
	(89/81)	(73/43)	(101/59)	(630/559)

Numbers of differentially-expressed genes identified by an ANOVA correcting for time, $P<0.01$ are indicated for each tissue comparison. The numbers in parentheses indicate numbers of genes in the intersecting set of each tissue in the column header where mean expression in A/J exceeds C57BL/6J versus the opposite.

Table 5. *Genes commonly differentially expressed across all tissues (P<0.01) using an ANOVA correcting for time effects.*

Gene	Description	Chr	Heart Pval	Liver Pval	Lung Pval	Spleen Pval
Psmb5	proteasome (prosome, macropain) subunit, beta type 5	11	5.70E-06	2.44E-04	7.14E-07	2.20E-04
2810452K22Rik	RIKEN cDNA 2810452K22 gene	12	3.58E-03	2.57E-04	2.62E-03	1.69E-03
4932416N17Rik	Mus musculus mRNA for mKIAA0350 protein	16	1.49E-07	2.15E-03	8.85E-08	8.51E-04
Galnt11	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 11	5	1.23E-03	9.01E-07	2.31E-03	5.61E-04
Ctsc	cathepsin C	7	3.41E-03	6.17E-04	9.97E-05	6.16E-05
Clu	Clusterin	14	5.06E-04	5.83E-03	2.16E-04	4.44E-03
Mapre1	microtubule-associated protein, RP/EB family, member 1	2	8.87E-03	8.03E-03	2.93E-04	1.82E-04
Glo1	glyoxalase 1	17	5.50E-11	1.25E-06	3.69E-08	1.65E-07
Cap1	adenylyl cyclase-associated CAP protein homolog 1 (S. cerevisiae, S. pombe)	4	6.48E-03	1.28E-05	1.88E-04	5.97E-05
Thumpd1	THUMP domain containing 1	7	3.12E-06	3.26E-04	3.84E-07	1.92E-07
Ifi202a	interferon activated gene 202A	1	2.17E-03	7.53E-03	3.62E-06	1.05E-04
Gnb1	guanine nucleotide binding protein, beta 1	4	8.19E-06	1.53E-06	9.34E-09	7.76E-03
D9Wsu18e	DNA segment, Chr 9, Wayne State University 18, expressed	9	3.55E-08	3.52E-04	2.56E-04	5.14E-06
Zfp68	Zinc finger protein 68	5	5.76E-04	2.01E-03	3.27E-04	4.52E-03
Tceb3	transcription elongation factor B (SIII), polypeptide 3 (110kD)	4	1.68E-05	3.13E-04	8.10E-05	1.72E-03
Hod	homeobox only domain	5	1.09E-09	4.28E-05	1.86E-09	2.31E-04
Arpp19	cyclic AMP-regulated phosphoprotein	9	8.50E-05	4.35E-05	1.53E-04	9.29E-03
1110033J19Rik	RIKEN cDNA 1110033J19 gene	6	2.62E-06	3.04E-04	7.94E-05	8.09E-05
Gas5	growth arrest specific 5	1	4.91E-11	1.86E-07	3.32E-07	9.19E-08

Table 6. *Coincidence of differentially expressed genes within QTL LOD intervals previously mapped in lung.*

Tissue	Trait	Locus	Chr	LOD	Number of differentially expressed genes, P<0.01 (vs. total) within QTL interval	Reference
Lung	Acute lung injury		1	2.58	88 (417)	[158]
Lung	Airway responsiveness	Bhr1	2	3.8	86 (411)	[140]
Lung	PKC activity		3	2.7	58 (275)	[142]
Lung	Acute lung injury	Ali4	6	3.02	125 (602)	[158]
Lung	Lung adenoma	Pas1	6	9	41 (166)	[124]
Lung	Acute lung injury	Ali1	11	6.8	68 (203)	[120]
Lung	Acute lung injury		12	2.33	36 (164)	[158]
Lung	Airway responsiveness	Bhr2	15	3.8	35 (193)	[140]
Lung	Airway responsiveness	Bhr3	17	2.9	112 (508)	[140]
Lung	Acute lung injury	Ali3	17	4.3	19 (105)	[157]
Lung	Lung adenoma	Pas2	17	3.0	23 (97)	[123]
Lung	Lung adenoma	Pas3	19	7	65 (322)	[123]
Total					756 (3463)	

Total number of genes on MGU74Av2 array contained within the QTL interval is indicated in parentheses.

Table 7. *Interaction of strain with time-dependent effects. Analysis of old and new datasets together - ANOVA correcting for the time experiment was conducted with the model: Expression ~ Time + Strain + Strain*Time (Number of genes with P>0.01 for Strain*Time).*

Tissue	<0.01 Strain >0.01 Time	<0.01 Strain <0.01 Time	>0.01 Strain <0.01 Time	<0.01 Strain*Time
Heart	881	570	3063	442
	813	475	2928	
Liver	567	681	4911	1116
	431	479	4513	
Lung	725	440	3204	463
	667	361	3056	
Spleen	1229	453	2954	656
	1105	468	2777	

Table 8. *Examples of genes detected by this study in agreement with previous phenotypic studies in A/J and C57BL/6J mice.*

Gene	Desc	Chr	Tissue	P Value	Trait	QTL	Ref
Ifnrg	Interferon gamma receptor	10	Spleen	7.58E-06	Listeria resistance		[129]
Il6ra	Interleukin receptor alpha	3	Spleen	0.00035	Listeria resistance		[129]
Csf3r	Colony stimulating factor3 receptor (granulocyte)	4	Spleen	0.00805	Listeria resistance		[129]
Kras2	Kirsten rat sarcoma oncogene 2, expressed	6	Lung	0.00034	Lung adenoma	Pas1	[123]
Pon3	paraoxonase 3	6	Lung	0.00260	Lung injury	Aliq4	[120]
Vcam1	vascular cell adhesion molecule 1	3	Spleen	0.00813	Malaria resistance	Char4	[118]

FIGURE LEGENDS

Figure 1A. Separate analysis of experiments conducted at two time points: Histogram of the number of differentially-expressed genes retrieved by various combinations of analysis and summary statistics. Differential expression was determined either by T-test ($P<0.05$) on data summarized by Dchip with invariant set normalization, or RMA with quantile normalization (Bioconductor).

Figure 1B. Volcano plots for genes differentially expressed in each lung across A/J and C57BL/6J strains. P-values are plotted on the Y-axis versus the maximum fold change between strains on the X-axis. Maximum fold changes (MFC) were calculated between A/J and C57BL/6J at time 1 versus at time 2. P-values were obtained by applying either a Student's t-test or an ANOVA (testing the model expression=replicate+strain). Analyses furthermore compared probeset summary and normalization techniques, Dchip with invariant set normalization, and RMA with quantile normalization. Genes detected as differentially expressed are displayed in red.

Figure 1C. Comparison of gene expression changes in datasets from both time points for lung: Mean fold changes between A/J and C57BL/6J at time 1 versus at time 2. Genes detected as differentially expressed at both time points (in red) display a consistent change of direction between strains.

Figure 1D. QQ plots of P-values obtained from applying ANOVA to each tissue separately using the model: Expression ~ Strain + Time+ Strain*Time.

Figure 2A. Comparison of genes commonly differentially expressed across A/J and C57BL/6J across tissues (heart, liver, lung and spleen). Histogram of the number genes differentially expressed in common between tissues.

Figure 2B. Histograms of 19 genes found to be differentially expressed ($P<0.01$) between A/J and C57BL/6J across all tissues: log of the ratio between mean expression levels between stains is expressed on the vertical Z-axis, the first horizontal axis lists genes from Table 3 ranked by log ratio in heart, and the second horizontal axis denotes tissues.

Figure 3. QQ plots of P-values obtained from applying ANOVA to the combined dataset of all tissues using the model: Expression ~ Strain + Time + Tissue

Figure 4. QQ plots of P-values obtained from applying ANOVA to the combined dataset of all tissues using the model including all possible interaction terms: Expression ~ Strain + Time + Tissue + Strain*Time + Strain*Tissue + Tissue*Time + Strain*Tissue*Time

Figure 5. Chromosomal position of genes differentially expressed between strains compared with QTL intervals previously mapped for asthma in blue [140, 142, 163], acute lung injury in green [120, 157, 158, 164], and lung cancer in purple [123, 165]. Left Y axis: P value (inverted), Right Y axis: LOD score of QTL, X axis: Chromosomal location expressed as a fraction of chromosome length.

Figure 6. Genes differentially expressed between strains in each tissue. P-values (Y-axis inverted) versus fractional chromosomal position for A) chromosomes 1-5, B) chromosomes 6-10, C) chromosomes 11-15, D) chromosomes 16-19 and X.

Figure 1A.

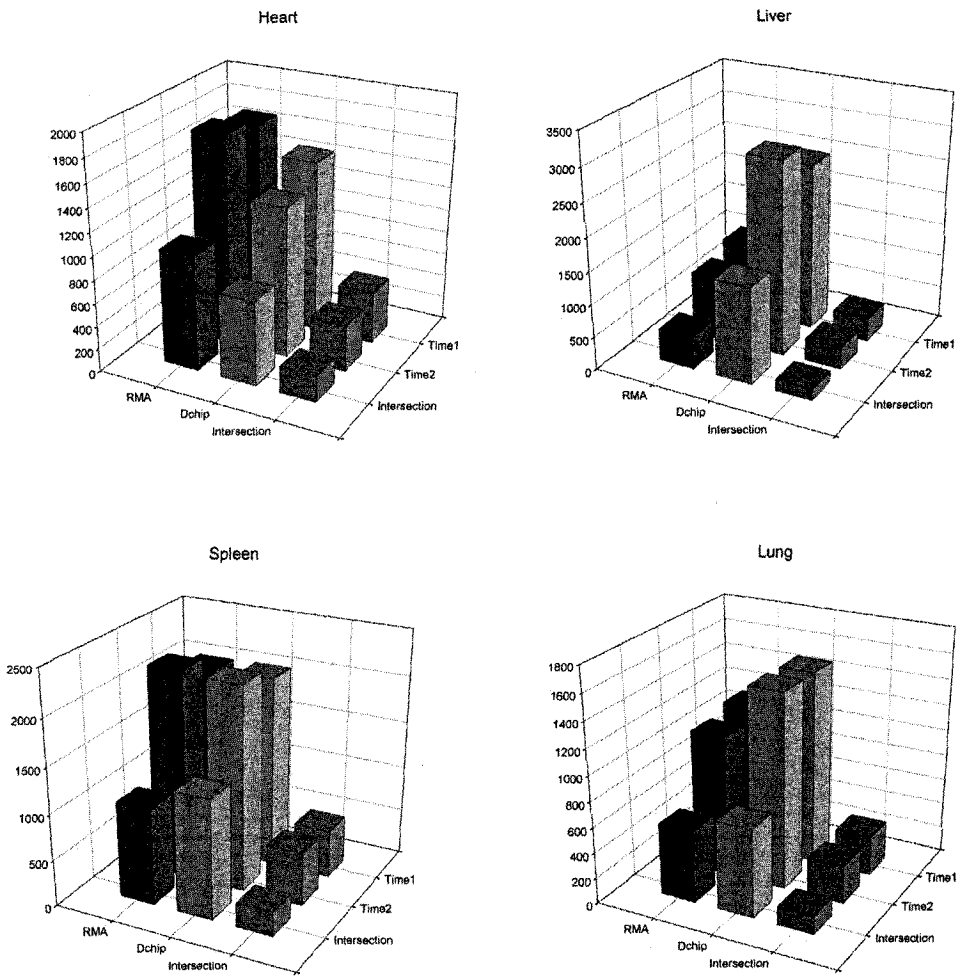

Heart
2000
1800
1600
1400
1200
1000
800
600
400
200
0
RMA
Dchip
Intersection
Time1
Time2
Intersection
Liver
3500
3000
2500
2000
1500
1000
500
0
RMA
Dchip
Intersection
Time1
Time2
Intersection
Spleen
2500
2000
1500
1000
500
0
RMA
Dchip
Intersection
Time1
Time2
Intersection
Lung
1800
1600
1400
1200
1000
800
600
400
200
0
RMA
Dchip
Intersection
Time1
Time2
Intersection

Figure 1B.

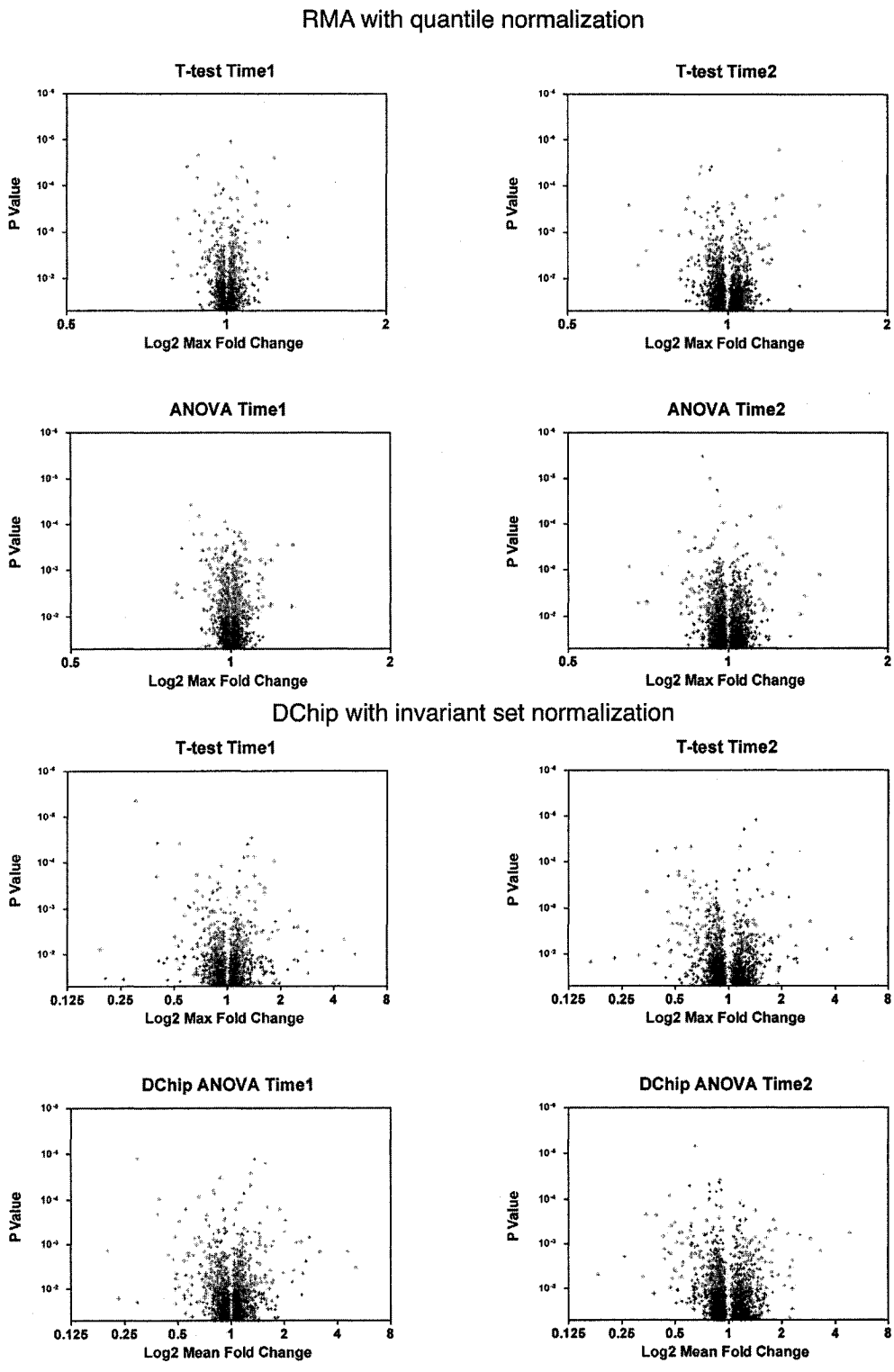

RMA with quantile normalization
T-test Time1
T-test Time2
ANOVA Time1
ANOVA Time2
P Value
Log2 Max Fold Change
DChip with invariant set normalization
DChip ANOVA Time1
DChip ANOVA Time2
Log2 Mean Fold Change
+ P < 0.01
+ P > 0.01
P < 0.01 in Time1 & Time2

Figure 1C.

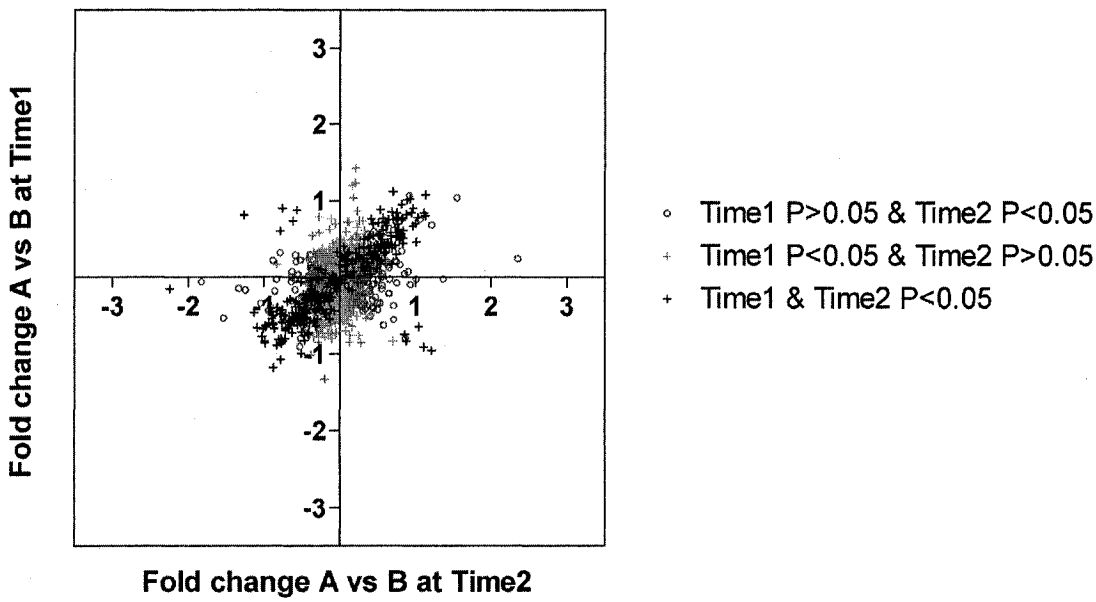

Fold change A vs B at Time1
Fold change A vs B at Time2
3
2
1
-1
-2
-3
-3
-2
1
2
3
Time1 P>0.05 & Time2 P<0.05
Time1 P<0.05 & Time2 P>0.05
Time1 & Time2 P<0.05

Figure 1D.

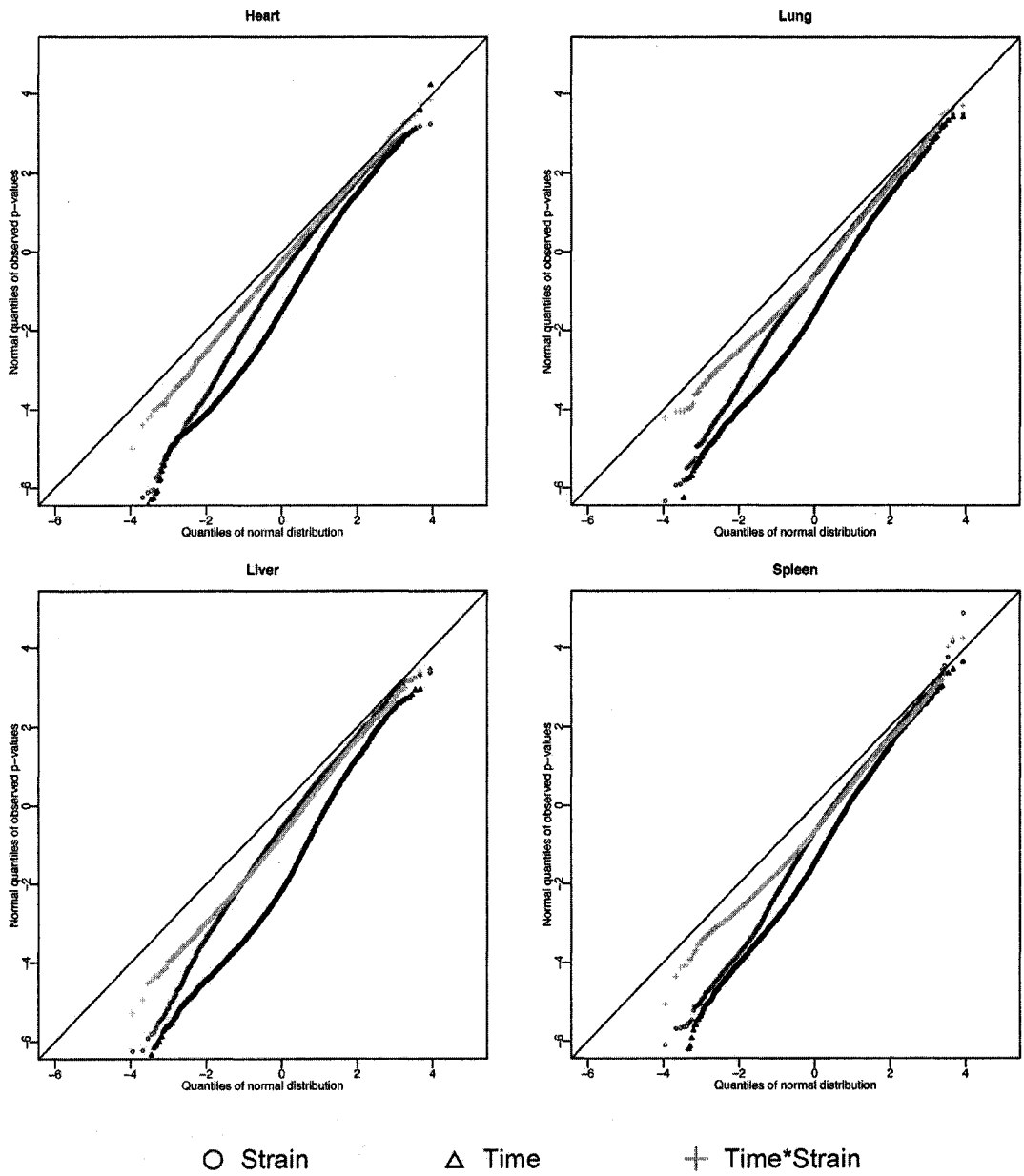

Heart
Lung
Liver
Spleen
Normal quantiles of observed p-values
Quantiles of normal distribution
−6
−4
−2
0
2
4
○ Strain
△ Time
+ Time*Strain

Figure 2A.

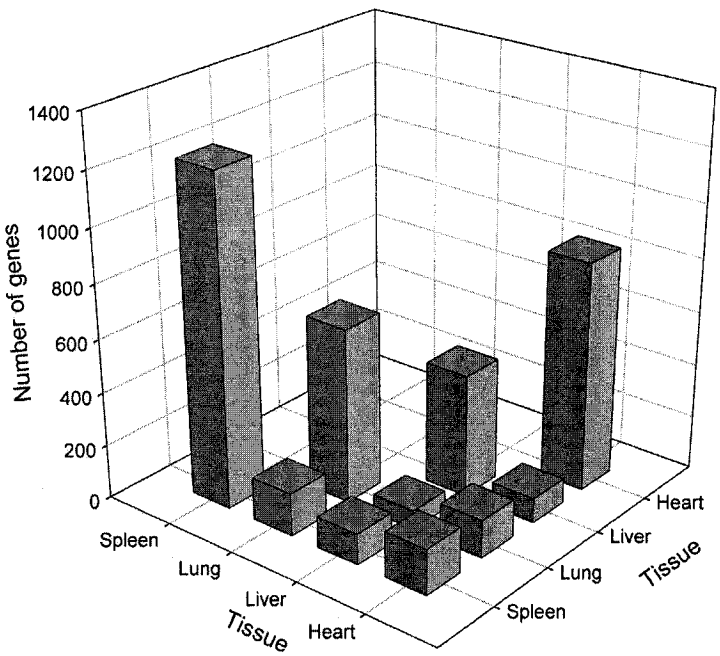

1400
1200
1000
800
600
400
200
0
Number of genes
Spleen
Lung
Liver
Heart
Tissue
Heart
Liver
Lung
Spleen
Tissue

Figure 2B.

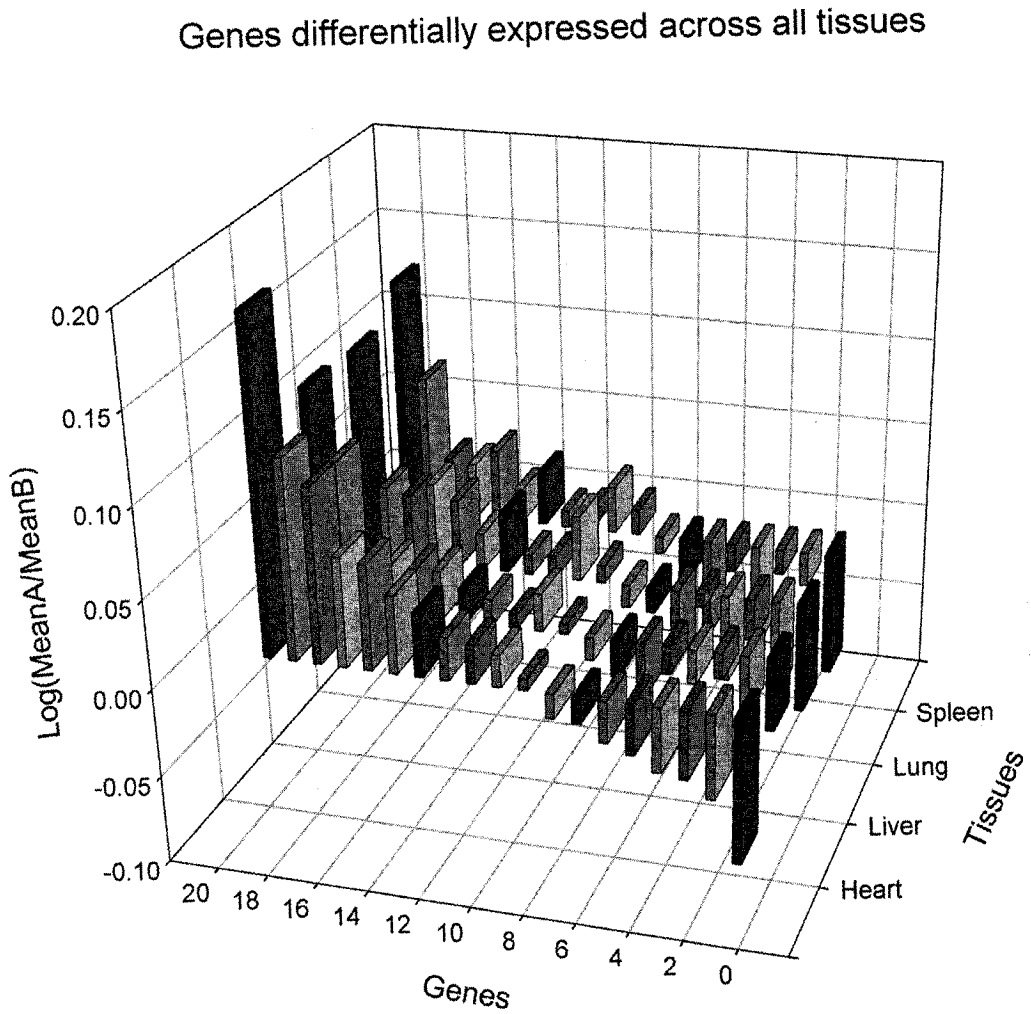

Genes differentially expressed across all tissues
Log(MeanA/MeanB)
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
20
18
16
14
12
10
8
6
4
2
0
Genes
Spleen
Lung
Liver
Heart
Tissues

Figure 3.

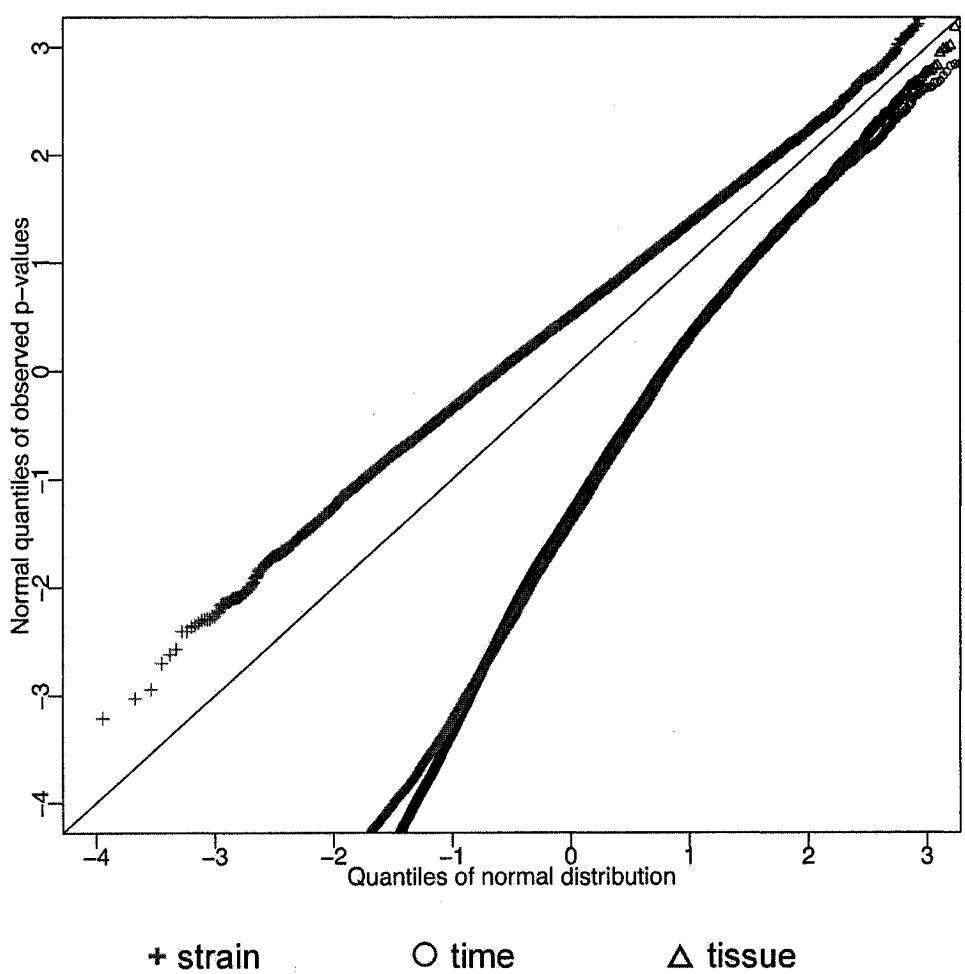

Normal quantiles of observed p-values
Quantiles of normal distribution
+ strain
O time
Δ tissue

Figure 4.

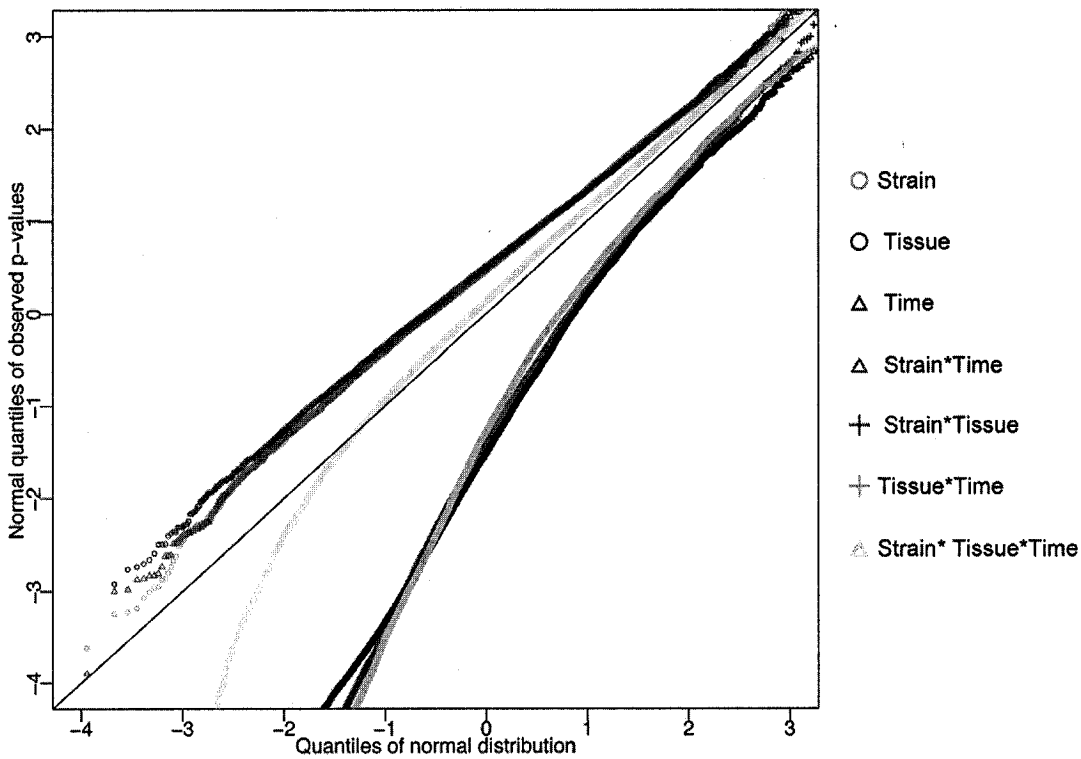

Normal quantiles of observed p−values
Quantiles of normal distribution
−4
−3
−2
−1
0
1
2
3
Strain
Tissue
Time
Strain*Time
Strain*Tissue
Tissue*Time
Strain* Tissue*Time

Figure 5.

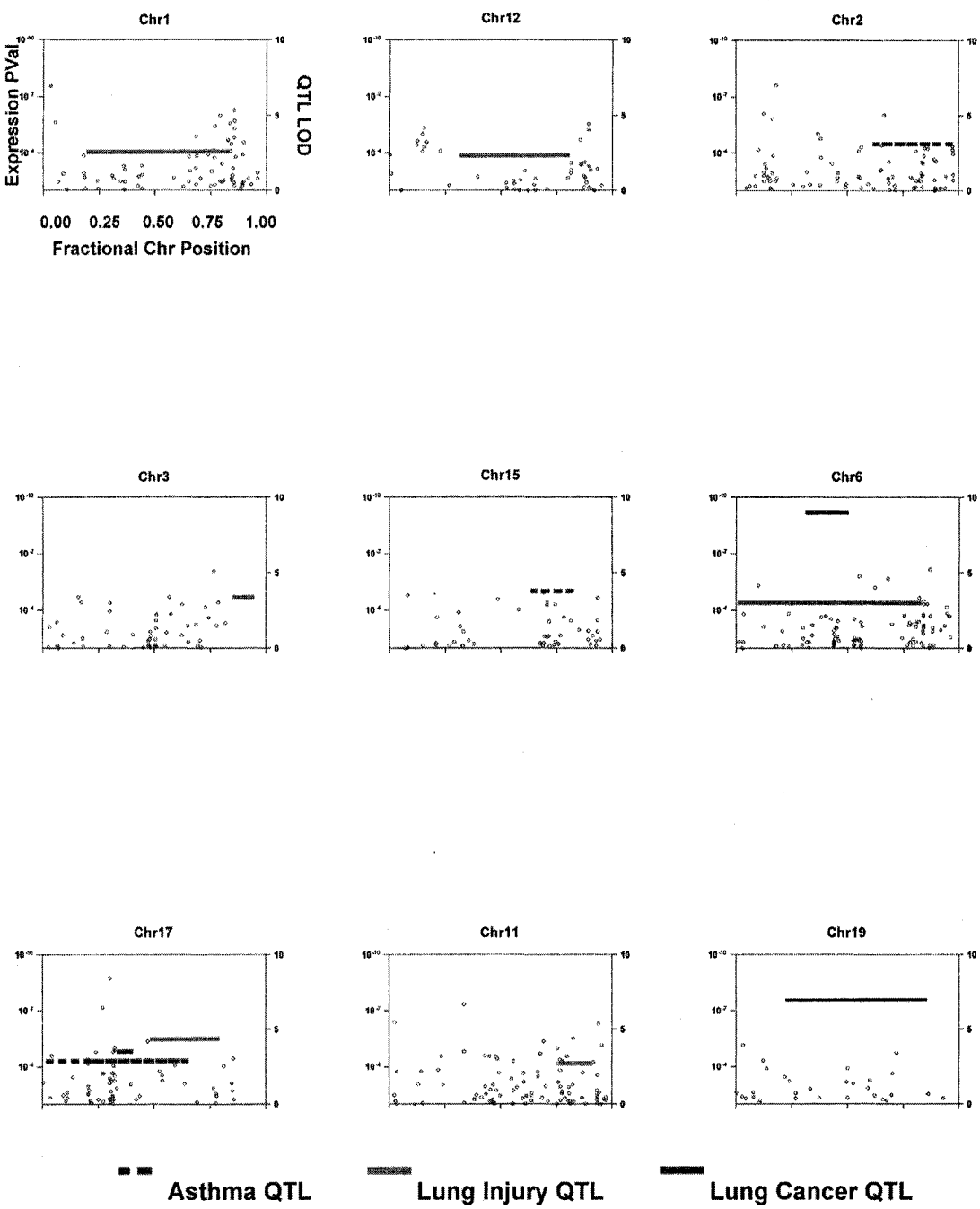

Chr1
Chr12
Chr2
Chr3
Chr15
Chr6
Chr17
Chr11
Chr19
Expression PVal
QTL LOD
0.00 0.25 0.50 0.75 1.00
Fractional Chr Position
Asthma QTL
Lung Injury QTL
Lung Cancer QTL

Figure 6A.

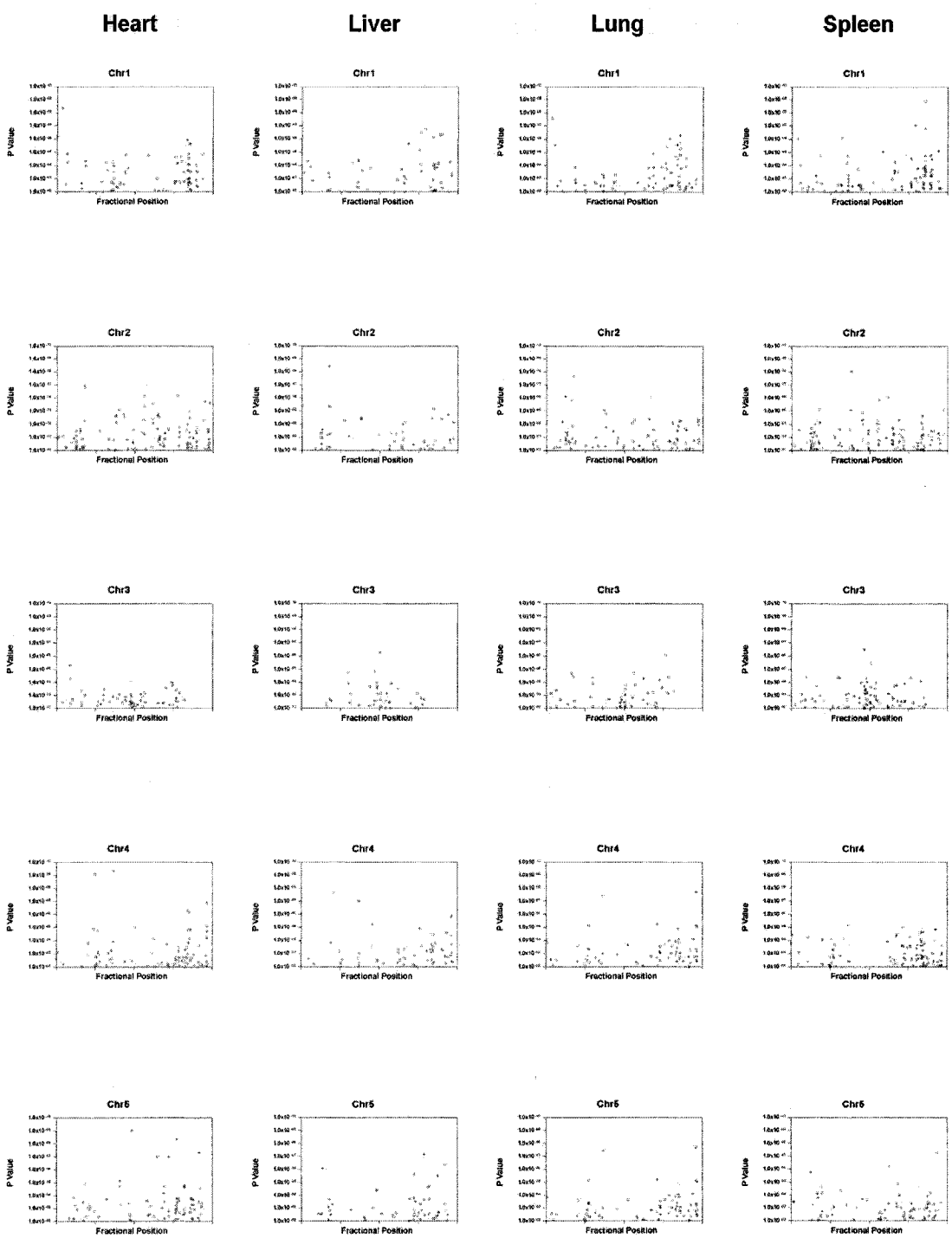

Heart
Liver
Lung
Spleen
Chr1
Chr2
Chr3
Chr4
Chr5
P Value
Fractional Position

Figure 6B.

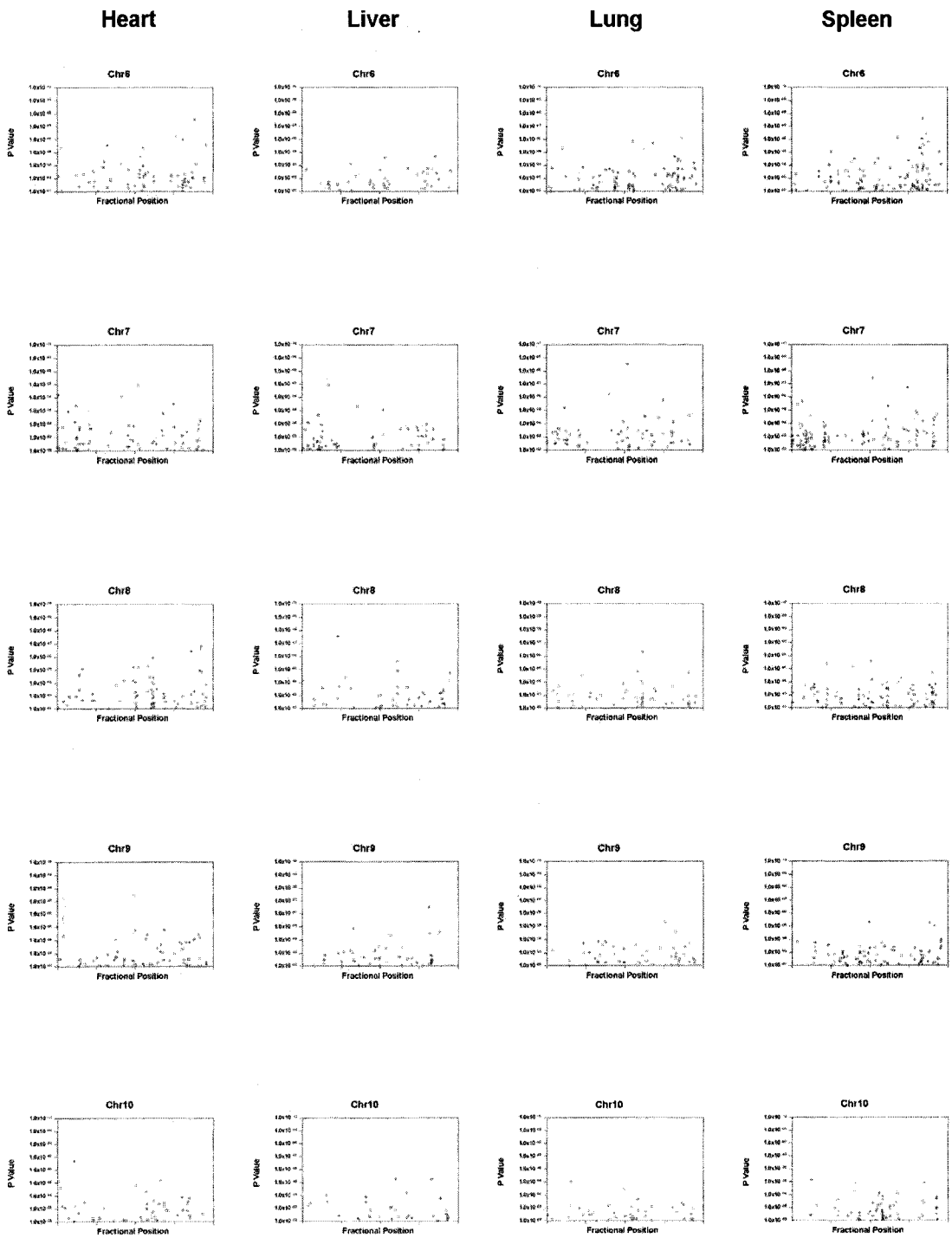

Heart
Liver
Lung
Spleen
Chr6
Chr7
Chr8
Chr9
Chr10
P Value
Fractional Position

Figure 6C.

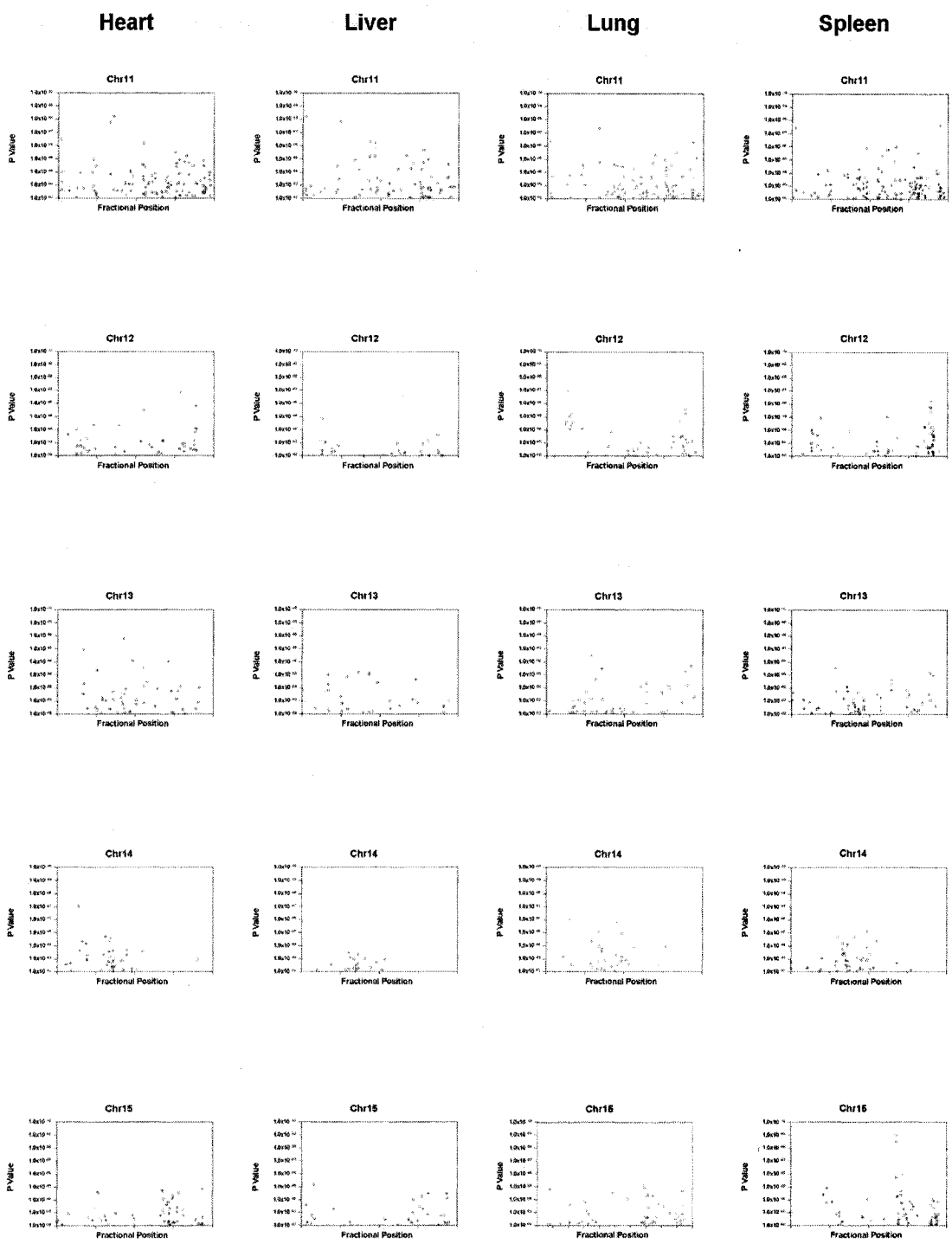
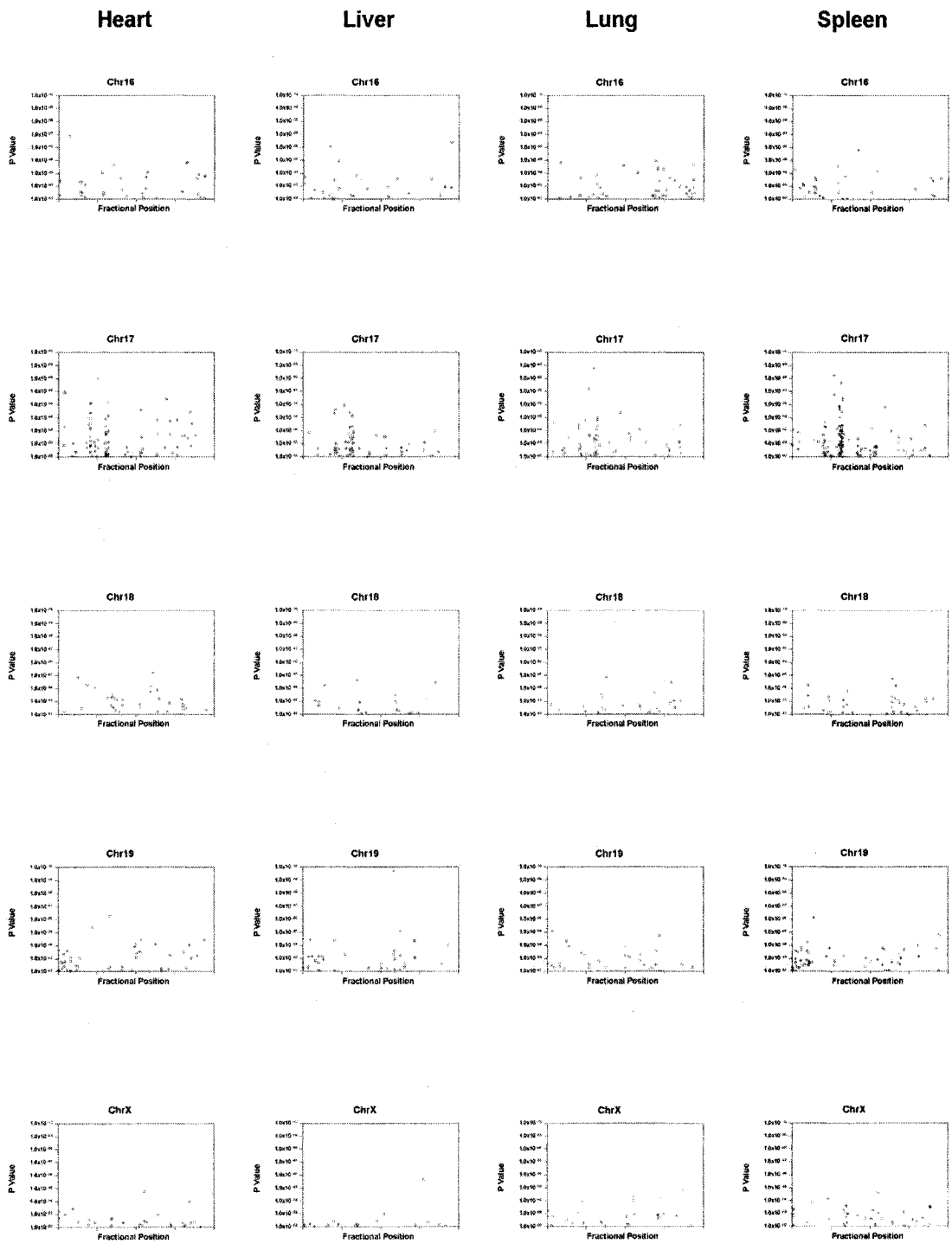

Heart
Liver
Lung
Spleen
Chr11
Chr12
Chr13
Chr14
Chr15
P Value
Fractional Position

Figure 6D.


Heart
Liver
Lung
Spleen
Chr16
Chr17
Chr18
Chr19
ChrX
P Value
Fractional Position

CHAPTER 3 – INTEGRATIVE APPROACHES TO DETECTING CIS-REGULATORY VARIATION ACROSS THE MOUSE GENOME

The previous chapter demonstrated that, like studies comparing more distantly related organisms[31, 34, 166] (separated by millions of years of evolution), expression profiling revealed a substantial degree of gene expression variability between mouse inbred strains (separated by one hundred of years of breeding). Such evidence points to a connection between differences in gene expression the evolution of heritable traits[167, 168]. The heritability of gene expression differences has been demonstrated[30, 104, 169]. However, these and other studies have established few direct links between gene regulation and changes in gene expression. The work of the previous chapter indirectly demonstrated the correlation between gene expression differences between A/J and C57BL/6J mice, strains known to differ genetically. Direct association between gene expression and regulatory genetic variation is the focus of this chapter.

Changes in genetic sequence may bring out phenotypic changes through either coding polymorphisms, affecting the structure of a gene product, or by regulatory polymorphisms, affecting the expression of a gene through the quantity of transcript. Regulatory polymorphisms were previously believed to reside primarily in the promoter region of the gene, affecting the binding sites for transcriptional activating proteins upstream from the transcription start site[170]. We now believe that regulatory variants may reside over large genomic regions, and can act via several mechanisms such as alternative splicing of transcripts[171], and differential rates of transcript degradation[170]. Regulatory polymorphisms located proximal to the affected gene are termed cis-acting whereas

variants that affect the expression of distal genes are termed trans-acting. For the sake of this discussion, cis-acting shall be used to describe regulatory elements located on the same chromosome as the target gene. Genes affected by cis-acting regulatory polymorphisms may be revealed by measuring allele-specific expression ratios in heterozygous individuals. The method for measuring this, allelic imbalance (AI), uses coding SNPs to measure the relative transcription levels of different alleles in individuals heterozygous for the given locus. AI is currently considered a method for the detection of cis-regulated genes because it provides a comparison of the expression levels of two alleles under the identical cellular context[172].

The work described in the current chapter attempts to detect potential cis-regulated genes by combining expression profiling data with genotyping data from a panel of RCS mice. The RCS panel was initially developed as an experimental tool for more refined mapping of quantitative traits[18, 173]. By providing a panel of strains, each genetically composed of variable segments of one inbred strain's genome on the background of another, the panel provides an experimental system with reduced genetic background. The smaller size of segments distributed over the genome for all strains enables finer mapping of phenotypic associates, reducing the initial size of mapped loci[174-176], and offering the possibility to better study interactions between multiple loci, since they are separated by the breeding process and will be present in varied combinations across the panel of strains[15, 16, 177].

By applying expression profiling to RCS, this study represents a combination of functional genomics and genetic approaches. The RCS panel used in this study were

generated from A/J and C57BL/6J parental strains and were previously genotyped for 621 microsatellite markers separated on average by 2.6 cM (~5Mb)[62]. Our analysis began by aligning probe sequences from the Affymetrix U74Av2 oligonucleotide array and microsatellite markers with the Feb 2003 UCSC assembly of the mouse genome using BLAT[178]. Genotype information from these markers enabled the parental strain (which we termed donor strain of origin or DSO) of each probeset on the array to be inferred for each strain. This allowed determination of the association between expression profiles over all strains with DSO profile.

The use of these strains furthermore permitted measuring the association of expression with background strain of origin (BG). Because approximately half of the strains contained genomic segments primarily from one or the other parental strain (to a ratio averaging 1:7) the contribution of BG to the observed variability in expression could be measured by including a term for BG in the ANOVA model. Evidence for the potential role of BG in expression variability came from the previous chapter's comparison of the parental strains where over 600 genes showed differential expression between strains (having adjusted for time effects). While it was not possible to determine the proportion of expression changes between parental strains that were due to cis or trans acting variants, the RCS panel provides the experimental tool to make this distinction. The parental comparison indicated that a substantial portion of the expression variability observed across the RCS panel could be due to BG. Therefore BG constituted a potential confounding variable for which adjustment was necessary.

ENRICHED DETECTION OF GENES WITH ALLELE-SPECIFIC EXPRESSION DIFFERENCES BY EXPRESSION PROFILING IN RECOMBINANT CONGENIC STRAINS

Peter D Lee[1,3,4], Tomi Pastinen[1], Celia M T Greenwood[5,6], Anny Fortin[7], Bing Ge[1,2], Yannick Fortin[1], Marina Takane[4], Emil Skamene[2,3,7], Michael Hallett[4], Thomas J Hudson[1,2,3§], and Robert Sladek[1,2,3]

[1]McGill University and Genome Quebec Innovation Centre, [2]Research Institute of the McGill University Health Centre, [3]Department of Human Genetics, Faculty of Medicine, McGill University, [4]McGill Centre for Bioinformatics, [5]Program in Genetics and Genomic Biology, Hospital for Sick Children, [6]Department of Public Health Sciences, University of Toronto, Emerillon Therapeutics[7].

[§]Corresponding author

740 Dr. Penfield Ave, Montreal, QC, H3A 1A4

Tel: (514) 398-3311 ext 00385

Fax: 514-398-2622

tom.hudson@mcgill.ca

Submitted for review to Genome Research, May 2005.

ABSTRACT

A high proportion of genetic variants modifying complex traits are believed to influence the regulation of gene expression levels. Cis-acting variation in gene regulation, estimated to affect 30-50% of human genes, remains largely uncharacterized. Better methods are required for genome-wide discovery of genes affected by cis-acting regulatory variation. We conducted expression profiling on lung tissue obtained from a panel of 30 AcB/BcA recombinant congenic strains (RCS) generated from A/J and C57BL/6J mice. By applying an ANOVA model adjusting for predominant background strain, we detected over 1500 genes displaying an association between expression levels and the donor strain of origin (DSO) for the corresponding locus. To further characterize associations detected by this model, we conducted SNP discovery by resequencing genes randomly selected from those whose expression did or did not associate with DSO. Within 1kb of 3'UTR resequenced, SNPs were found in 52% of the positive genes versus 27.5 % of negative genes ($P<0.05$, Fisher Exact Test). To investigate whether the frequency of regulatory variants might differ among these subsets, we measured allelic imbalance (AI) using cDNA from F1 mice generated from an A/J X C57BL/6J cross. We detect a significant enrichment of AI in genes identified by expression profiling; 63% of positive genes showed AI versus 23% of negative genes ($P<0.01$, Fisher Exact Test) indicating a higher rate of cis-acting regulatory variation in genes displaying expression association with DSO across the RCS panel. This study demonstrates an integrated genome-wide approach for enriched detection of genes affected by cis-regulatory variation in mice.

INTRODUCTION

While genetic research has identified many examples of diseases caused by genetic variants affecting coding sequences, genetic variants affecting gene regulation are now believed to account for a large proportion of changes contributing to the evolution of complex traits[25]. Since current techniques that study gene regulation are best suited for investigating individual genes, new approaches are necessary to identify genetic variability as it affects gene regulatory mechanisms on a genome-wide scale. The importance of characterizing genetic regulation using large-scale approaches is furthermore emphasized by recent studies demonstrating the heritability of gene expression[104], mapping of expression traits[28], as well as the impact of regulatory variants in human disease [29, 30, 104, 131, 179].

Increasing evidence suggests that variation in gene regulation affecting gene expression plays an important role in complex phenotypes[180]. The search for regulatory variants underlying complex traits lags behind the identification of protein coding variants, perhaps due to the lack of generalized knowledge about transcriptional regulation on a genome-wide scale as well as to the complexity of interactions between activators and sequence elements seen to regulate gene transcription[21]. This has led to the development of numerous experimental systems that attempt to augment the identification of candidate genes underlying complex traits by reducing genetic complexity of the sample population. Recombinant congenic strains (RCS) represent one such murine system, where specific breeding of two inbred strains generates a panel of strains, each containing variable congenic segments from the genome of the donor strain (averaging 12%) on that

of the background strain[173]. An RCS panel of sufficient size (>20 lines) insures that 95% of genes are contained on donor congenic segments and provides a system for separating loci involved in multigenic traits, permitting each locus to be analyzed separately[175]. The increase in mapping efficiency afforded by RCS has led to identification of candidate genes[181] as well as multilocus interaction effects[177].

The combination of gene expression profiling with such experimental systems offers numerous advantages over and above either technique alone[68]. Correlations between gene expression levels and sequence variation permit genetic analysis at the level of individual genes rather than broad genomic regions typical of QTL analysis. Furthermore, this approach permits the identification of interactions and potential pathways suggestive of gene regulation events[131]. Recent studies in yeast using integrated approaches have introduced a classification of regulatory effects across the genome[28, 36] ranging in complexity from a small proportion of genes affected by single locus cis-effects, to a larger number affected by two loci, and an even higher proportion displaying complex modes of inheritance. Studies in humans show that gene expression is a heritable trait[30] and that regulatory variants affecting the level of transcripts can have phenotypic effects[180].

Gene regulatory mechanisms may be broadly categorized into cis-acting, operating proximal to the interrogated gene, and trans-acting, exerting their effects distally at one or more genes on other chromosomes, usually occurring via the expression of other genes such as transcription factors. The presence of cis-acting variants may be detected by

analysis of allele-specific expression ratios for a specific gene[170]. Where transcripts contain informative SNPs, the relative expression levels of the two alleles may be measured in individuals heterozygous for intragenic polymorphisms using a variety of techniques[135, 172, 182]. Allelic imbalance (AI) describes the state where one allele is overexpressed with respect to the other. A difference in allele-specific transcript levels for a given gene indicates the presence of cis-acting regulatory variants since transcription levels of the two alleles are compared under identical cellular contexts *in vivo*[135]. The technique is furthermore sufficiently scalable to allow surveying of cis-acting gene regulation on a genome-wide scale [172].

In this study we compare 2 inbred strains of mice, A/J and C57BL/6J, both used for close to a century as model systems for human disease and previously characterized for numerous complex phenotypes. We have previously reported the baseline gene expression differences between adult males of these strains and demonstrated widespread tissue-specific expression variability between these strains in 4 tissues[183]. Other studies have also demonstrated the presence of allele specific expression differences between these strains across 3 tissues[135]. This study involves genome-wide expression profiling on lung tissue across a panel of 12 AcB and 18 BcA RCS previously derived by reciprocal backcrossing of A/J and C57BL/6J mice[62]. We further investigated the correspondence of these results with genetic variability between the strains by resequencing for SNPs in genes identified by expression analysis. To further investigate the presence of cis-acting regulatory variation among the donor strains, genes identified by expression analysis

were subsequently evaluated for allelic imbalance using a sequence-based method in F1 offspring of an A/J x C57BL/6J cross.

MATERIAL AND METHODS

Mice: This study used a RCS panel previously developed and genotyped in duplicate for 621 microsatellite markers at an average spacing of 2.6cM[62]. Mice were housed with free access to food and water in the animal colony of the Montreal General Hospital Research Institute under conditions of 12 hours of darkness and 12 hours of light. Animals were fasted for 24 hours before sacrificing. Genotying data was obtained by virtue of a Material Transfer Agreement with Emerillon Therapeutics, Montreal. Out of the 26746 genotypes, 25571 results agreed between replicates, 520 results disagreed between replicates (where one result was definite and the other undefined), 19 results were contradictory between replicates (showing results from opposing parental origins for the same locus) and 33 results were undetermined. Disagreeing or contradictory results were discarded by assigning an undetermined status for those loci. As these results were dispersed throughout the dataset (over all markers and strains), the effect on individual genes and strains in the analysis was minimal.

Microarray studies: Lung tissue was obtained from two 3-month-old male mice from each of 12 AcB and 18 BcA RCS by rapid dissection and freezing in liquid nitrogen. The duplicate samples were homogenized in Trizol reagent and RNA was prepared and hybridized to Affymetrix MGU74Av2 oligonucleotide arrays as described previously[80]. Expression data was summarized with RMA and quantile normalized using Bioconductor (www.bioconductor.org). These methods were chosen after comparison of MVA plots for various combinations of normalization and summary methods (data not shown). Plots that displayed the least change in variance over the average intensity were chosen so as to

minimize departures from assumptions of constant variance in parametric analysis methods used subsequently.

Probe sequences from the Affymetrix MGU74Av2 oligonucleotide array were aligned against the Feb 2003 NCBI build of the mouse genome using BLAT with default parameters[111]. Of the 12422 probe sequences on the chip, 11293 were localized reliably. Physical positions of SSLPs from the RCS genotype data were retrieved by aligning marker sequences to the same assembly using BLAST and retrieving all primer matches with intervening sequences less than 500bp (wordsize=12, p=.90, e=0.1) and iteratively relaxing matching parameters in the event of no match. This procedure matched 566 of 621 (90%) markers from the genotyping data. We inferred DSO for each gene based on the DSO of surrounding SSLPs; Genes flanked by markers of identical DSO were assigned the same DSO. If flanking markers differed (e.g. the boundary of a recombined segment), genes in between were assigned an unknown DSO status. Genes at the ends of chromosomes were assigned the DSO of the closest SSLP.

ANOVA was conducted gene by gene assuming independence of loci, using the linear model: EXPRESSION ~ DSO + BG + DSO*BG (where background, BG, was calculated for each strain as the ratio of total segment lengths over the entire genome with DSO from one parent over the other). Positive association between expression and DSO was assigned if the $P<0.05$ for DSO and $P>0.1$ for DSO*BG. All analyses were conducted using R (www.r-project.org).

Validation studies: From sets of genes positive and negative for DSO association, 50 positive genes and 80 negative genes were selected randomly for resequencing 1kb of 3' UTR in A/J and C57BL/6J genomic DNA. Randomization was done on the subset filtered for a MAS5.0 mean value of 500 or more over all strains. This was done based on observations that this signal level corresponds to greater success of the sequence-based AI assay[184]. Selection was restricted to genes without documented alternative splicing in the Alternative Splicing Database[185] (http://www.ebi.ac.uk/asd/), and without complex loci (overlapping transcripts, reversed overlapping transcripts) in the UCSC Genome Browser[186] (http://genome.ucsc.edu). Primers for resequencing were designed using Primer3.0[187] using default parameters.

AI was measured in F1 mice (5 males) generated by crossing an A/J male with a C57BL/6J female mouse. Lung tissue was harvested at 13 weeks and RNA extracted as above. cDNA was generated using reverse-transcriptase. Sequence traces were generated using the sequencing primers in F1 cDNA and gDNA for all SNP-containing genes using methods previously described[172]. SNP peak heights in cDNA from these samples versus gDNA were measured using PeakPicker0.5, developed in-house, which normalizes SNP peak heights against those of the surrounding sequence. This method can detect differences in allelic expression >1.2 fold[170]. AI was determined by a paired Student's t-test comparing peak height ratios of the two alleles in gDNA versus cDNA over 5 replicate F1 mice. Association of SNP frequency and AI frequency with DSO expression analysis results was done in R using a one-tailed Fisher exact test and logistic regression models: AI ~ PVAL, and SNP ~ PVAL.

RESULTS

We first used the genotypes provided for 621 SSLPs spaced on average 2.6 cM apart[62] to infer DSO in 12 AcB and 18 BcA RCS lines for the probesets contained on the MGU74Av2 oligonucleotide array. We define donor strain of origin (DSO) as the parental strain from which a congenic segment derives. The overview of our experimental approach is outlined in Figure 1. Because of uncertainty as to the position of the recombination site between SSLPs of differing DSO (indicative of a recombination), we assigned an unknown DSO status to genes falling within such regions (Figure 2). This still permitted assignment of DSO to over 90% of the probesets on the array over all strains. We assigned unknown status 6.8% of the time (22882 of 338790 results) with the number of unknown assignments ranging from 220 for BcA82 and 1503 for BcA83 (genome-wide), and a maximum per chromosome of 366 for BcA74 on chromosome 8. The largest stretch of genes affected was on chromosome 10 for BcA70 where 46% of genes were assigned an unknown DSO status. Over all the entire dataset (all chromosomes) unknown DSO ranged from 4% for BcA6 to 38% for AcB51, averaging 21% over all strains. Strains contained between 26 and 56 congenic segments genome-wide, averaging 43.7 (between 4 and 38 A/J segments, 7 and 35 C57BL/6J segments, both averaging 22 over all strains). The maximum number of segments per chromosome was 5 (for BcA83 on chromosome 6 and AcB56 on chromosome 11). As expected the rate of unknown DSO assignment correlates with the number of recombined segments (Figure 3) indicating that higher frequencies of recombination decreased the amount of data included in our analysis.

To identify genes displaying associations between expression and DSO for the surrounding locus, we applied an ANOVA on a per-gene basis using a model that included terms for DSO, background strain (BG) and their interaction (DSO*BG). We included BG in the model based on our previous observations of numerous differences in baseline gene expression between A/J and C57BL/6J strains in multiple tissues including lung. The model assumes independence between loci and genes. This assumption provided between 38 and 62 effective replicates per gene (the number of results in the dataset for which DSO = A/J or C57BL/6J for any locus). The variability between individual replicate mice was factored into the error term for the ANOVA model. We note large chromosomal segments where most RCS are derived from one parental strain. This correlates with our observation of uneven partitioning of the dataset by the two terms in the ANOVA model (DSO and BG). As a result, a proportion of genes (30%) could not be tested. The association test was thus performed for 8860 genes.

Of the genes tested, we identify over 1500 genes with significant association between expression and DSO of $P<0.05$ having adjusted for BG and DSO*BG interaction (Table 1). This includes over 130 probesets with documented involvement in transcription (Table 2). A comparison of the extent to which each variable contributes to the overall variability reveals that DSO contributes the most, followed by BG and their interaction (Figure 4). We further note that 1213 genes display significant association ($P<0.05$) between expression and BG, and 651 for the interaction term DSO*BG. These results may indicate the presence of elements located distal to the locus affected. We chose

relaxed thresholds purposely in order to maximize the sensitivity of detecting genes affected by cis-regulatory polymorphisms.

Previous studies have attributed 97% of the genetic variability between inbred strains to a minority of regions throughout the genome thought to diverge from a common ancestral genome[58, 100]. These divergent haplotype blocks have been proposed as an additional tool for the genetic mapping of complex traits and correspond to regions of higher polymorphism between strains[55, 56, 59]. The contribution of this variable becomes further evident given the number of genes found to display significant contribution of effects due to BG. To confirm whether the results of our expression analysis corresponded to genetic differences at the surrounding locus we resequenced 1kb of 3'UTR for randomly selected genes from those negative and positive for DSO association to assess SNP content. We find a significantly increased occurrence of SNPs in genes positive for association between DSO and expression (Table 3). Resequencing of genomic DNA from parental strains A/J and C57BL/6J showed that 54% genes with positive DSO association contained SNPs as opposed to 27.5% of negative genes ($P<0.05$ Fisher Exact Test). We further observe a trend of increasing likelihood for SNP occurrence with significance of association between gene expression association and DSO in the ANOVA model (Figure 5). Because the frequency of SNPs in the negative set was lower, we resequenced 30 more genes in this set versus the positive.

The association of expression with DSO suggests the presence of factors affecting expression proximal to the affected gene. To investigate this observation further, we

sought to determine whether the set of genes identified by the model showed differences in allelic expression (indicative of cis-acting gene regulation) by measuring AI in SNP-containing genes using 5 replicate F1 mice generated from the two parental strains A/J and C57BL/6J (Figure 6). We selected 50 genes positive for DSO association and 80 without because of the lower rate of SNP discovery in the negative set. The genes containing SNPs displayed AI in 63% of positive genes versus 23% of negative genes ($P<0.01$ Fisher Exact Test). The presence of AI also displays dependence on the significance of DSO expression association by logistic regression (Figure 7), suggesting that the likelihood of AI increases with higher significance of association` between expression and DSO in the ANOVA model (Figure 8A-C). Results for AI are summarized in Table 3 and Figure 9.

DISCUSSION

We demonstrate an integrated approach for the genome-wide determination of genes subject to likely cis-acting genetic variation. The large-scale categorization and classification of genes based on modes of regulation is needed to understand the genetics of complex traits and may eventually lead to genome-wide models of gene regulation. The RCS expression dataset, combined with the F1 studies of allelic imbalance, allow us to estimate that 8-11% of genes are affected by cis-acting regulatory variation in one tissue among these strains. This estimate is higher than that of a previous study across 3 tissues and 4 mouse inbred strains where allelic expression differences were found in 3-6% of randomly selected genes, depending on the combination of tissue and strain[135]. A significant proportion of genes have been estimated to be subject to cis-acting regulation in yeast and humans[29, 169]. Demonstration of the heritability of gene expression, together with widespread gene-expression differences observed between subspecies further implies an evolutionary role for variation in gene regulation[26, 167, 188].

The RCS expression approach enriches for the detection of cis-acting regulatory variation. RCS mice provided a unique opportunity to link knowledge of genetic variation and gene expression due to the restricted and well-characterized level of variation across the derived strains. Since RCS are homozygous at every locus, a simple test for association between gene expression and DSO of the surrounding locus could be applied. The detection of allelic imbalance validates the presence of a functional cis-acting variant influencing the gene of interest. We demonstrate AI in 63% of genes showing an association between expression and DSO. This represents between a 10 and

20-fold efficiency of detection as compared to random screening of genes[135, 172]. We note that the detection of association between expression and DSO was greatly enhanced by our ability to subtract the effect due to the predominant strain of background. By contrast, an initial analysis without correction for background detected far fewer genes of smaller effect indicating the substantial contribution of background to the overall variability and the potential for background to confound proximal effects tested by the DSO term (data not shown). Logistic regression suggests that higher thresholds for significance in the ANOVA may correlate with a higher frequency of AI suggesting greater enrichment with more stringent thresholds.

The estimated 8-11% of genes subject to cis-acting regulation represents a lower limit for a number of reasons. A single tissue (lung) was analyzed at one developmental stage (adult). Since cis-regulation is known to act in a tissue-dependent fashion[135], and transcriptional changes abound throughout development, an analysis of more tissues over a greater number of developmental stages is likely to uncover additional genes affected by cis-regulatory variants. Our design excluded genes falling within segments containing a recombination since the positions of recombination sites were not characterized. While this affected a small percentage of results dispersed throughout the dataset (6%), this exclusion together with varied levels of heterogeneity in the dataset led to 70% of the genes on the microarray being tested by the ANOVA model. Improvements in map resolution would reduce the amount of missing data points by permitting DSO to be assigned for a higher percentage of genes. The sequence-based AI assay is seen to detect differences between allele-specific transcription as low as 1.2-fold[189]. Subtler differences

may reflect the present of compensatory effects such as negative feedback control mechanisms acting to tightly regulate transcript levels. AI detection depended upon the variance observed among replicate samples in cDNA and gDNA; larger gDNA variances rendered an indeterminate measurement of AI. In addition, the study assumed independence for each locus tested; this may not be appropriate since dependencies between genes and loci may exist that are on the same chromosome, largely dependent on the distance between loci[190]. An analytical strategy involving more traditional mapping techniques such as QTL mapping[191] and linkage analysis[192] would take into account the dependence on location as well as recombination rate throughout the genome.

There are additional causes of AI besides upstream cis-acting regulatory polymorphisms. Imprinting, whereby the allele of one parent is silenced by that of the other, was formerly thought to be the sole cause of differential allele expression. This phenomena, however, is quite rare thus far accounting for about 80 genes in humans[193] and 60 in mice[193] which have been catalogued[194]. To date, there is no evidence that known or new cases of imprinting were observed by AI in this study.

Current efforts to characterize the ancestral haplotype structure of inbred mice suggest that a major proportion of the variability between strains may be attributable to segments of the genome that differ ancestrally[58], which are embedded in the long segments created by the breeding process used to generate the RCS panel. What effects this might have on the sensitivity of the method used in this study are difficult to state in advance. However, one can only speculate that inclusion of this factor would increase our sensitivity to detect

cis-regulated genes. The higher number of SNPs found in the 3' UTR of genes displaying association between expression and DSO suggests a higher rate of polymorphism in the loci of affected genes. These more polymorphic genes are probably in regions that are most divergent between inbred strains reflecting haplotypes not inherited from a common ancestor[100]. This also implies that genes located in shared ancestral haplotypes should be less likely to exhibit an association between DSO and transcript level, further refining the location of cis-acting regulatory variants in the associations detected by our model. A comprehensive set of publicly available SNPs polymorphic between A/J and C57BL/6J strains is required to investigate this question thoroughly.

Our results suggest that a proportion of genes with cis-acting variants do not display expression differences that are detectable across the RCS panel. We observe five genes negative for association ($P>0.05$ for DSO) display AI. One explanation is that gene expression levels on the array may not have been sensitive enough to detect subtler changes in gene expression over the strains. However, owing to the transformations we used in the analysis, signal intensity-dependent effects are not likely to have exerted a major effect on the results[109]. Subtle effects may also indicate the presence of one or more collinear variables that confound the expression results, such as epistatic trans interactions suppressing variability. Given the extent of gene interactions believed to exist across the genome, collinear variables are more likely to be the norm rather than the exception. We note the substantial contribution of the background genetic composition to the overall expression variability observed across the strains; over 1200 genes displayed significant associations with BG ($P<0.05$), in addition to 600 genes that displayed

significant association with the interaction term, DSO*BG. These observations point to many determinants of expression variability that is distal to the affected gene, suggestive of trans-acting regulatory mechanisms. The estimation of the size of the background effect may prove useful in future determinations of trans effects and the extent to which genes may be affected by both forms of regulation. Other analytical methods designed specifically to detect epistatic interactions are required to provide corroborating evidence in this area.

While genes displaying AI in this study are more likely to contain cis-acting regulatory variants[195], identification of the causative polymorphisms remains to be determined empirically. The ability to identify potentially cis-regulated genes genome-wide would provide a key resource in the search for genetic determinants of complex traits and act as a prioritization tool for further detailed analysis. Mapping strategies to dissect such traits consistently face the difficulty of tracking multiple variables simultaneously, testing numerous interaction terms with little prior knowledge upon which to base hypotheses. Recent studies in yeast have shown progression towards classification of genes based on gene regulatory interactions[28, 36]. Construction of a systematic catalogue would not only favor the discovery of candidate genes, but also initiate the formation of a common vocabulary by which to describe this area of biology where little is known at the genome-wide level.

Table 1. *Differentially expressed genes in RCS.*

	DSO	BG	DSO*BG
P<0.01	894	515	178
P<0.05	1591	1213	651
P<0.10	2190	1816	1159

Numbers of differentially expressed genes identified for each term in the ANOVA model: Expression ~ DSO + BG + DSO*BG.

Table 2. *Transcription factors with association between expression and DSO by the ANOVA model.*

AFFYID	Gene	DSO P-value	BG P-value	DSO*BG P-value	Accession	Description
92399_at	Runx1	8.77E-12	0.1590	0.6970	D26532	runt related transcription factor 1
101980_at	Rpo2tc1	1.96E-11	0.1450	0.5600	J03750	RNA polymerase II transcriptional coactivator
95536_at	Tceb3	4.00E-10	0.3360	0.1310	AB025015	transcription elongation factor B (SIII), polypeptide 3
160616_at	Whsc2h	8.11E-10	0.2340	0.6470	AW047201	Wolf-Hirschhorn syndrome candidate 2 homolog (human)
104340_at	Mbd1	1.75E-09	0.0107	0.1450	AF072240	methyl-CpG binding domain protein 1
161466_r_at	Asb3	2.59E-09	0.2610	0.2730	AV347947	ankyrin repeat and SOCS box-containing protein 3
101943_at	Tceb3	1.40E-07	0.8810	0.3270	AA960259	transcription elongation factor B (SIII), polypeptide 3 (110kD)
93728_at	Tgfb1i4	2.97E-07	0.7160	0.5500	X62940	transforming growth factor beta 1 induced transcript 4
101382_at	Pbx2	3.56E-07	0.4890	0.0352	AF020198	pre B-cell leukemia transcription factor 2
104536_at	Madh2	1.41E-06	0.6290	0.9230	U60530	MAD homolog 2 (Drosophila)
95616_at	Crsp3	2.12E-06	0.1490	0.5620	AA674714	cofactor required for Sp1 transcriptional activation, subunit 3
96672_at	Hop-pending	4.64E-06	0.0656	0.5650	AW123564	homeodomain only protein
94804_at	Pbx1	6.10E-06	0.0166	0.5180	L27453	pre B-cell leukemia transcription factor 1
100010_at	Klf3	7.13E-06	0.7740	0.9030	U36340	Kruppel-like factor 3 (basic)
102641_at	Sfpi1	7.15E-06	3.77E-03	0.0241	L03215	SFFV proviral integration 1
94406_at	Phtf	1.07E-05	0.0426	0.7650	AJ242864	putative homeodomain transcription factor
99901_at	Ptrf	1.18E-05	6.57E-03	1.25E-04	AF036249	polymerase I and transcript release factor
93656_g_at	Usf1	2.49E-05	0.1500	0.1310	X95316	upstream transcription factor 1
102864_at	Hoxa7	2.69E-05	4.38E-03	0.0144	M17192	homeo box A7
100513_at	Ddefl	5.17E-05	0.3580	0.2000	AF075461	Development and differentiation enhancing
94189_at	Bcl6b	6.41E-05	0.1990	0.4900	AB011665	B-cell CLL/lymphoma 6, member B

AFFYID	Gene	DSO P-value	BG P-value	DSO*BG P-value	Accession	Description
101631_at	Sox11	8.92E-05	8.62E-08	0.8710	AF009414	SRY-box containing gene 11
103925_at	Mllt3	9.37E-05	5.06E-03	0.3750	AW120605	myeloid/lymphoid or mixed lineage-leukemia translocation to 3 homolog (Drosophila)
100011_at	Klf3	9.90E-05	0.3820	0.3930	AI851658	Kruppel-like factor 3 (basic)
99917_at	Ezh2	1.04E-04	0.8810	0.2750	U52951	enhancer of zeste homolog 2 (Drosophila)
92804_at	Polr2h	1.81E-04	0.7580	0.0117	AW122864	polymerase (RNA) II (DNA directed) polypeptide H
100307_at	Nfix	2.16E-04	0.9840	0.5930	AA002843	Mus musculus 4 days neonate male adipose cDNA, RIKEN full-length enriched library, clone:B430214H24 product:nuclear factor I/X,
92927_at	Etv1	2.20E-04	0.0556	0.4020	L10426	ets variant gene 1
99846_at	Foxf2	2.42E-04	0.0367	0.3690	Y12293	forkhead box F2
103236_at	Ring1	2.85E-04	0.0901	0.8940	Y12881	ring finger protein 1
98032_at	Zfp35	2.90E-04	0.4530	0.9250	M36146	zinc finger protein 35
97926_s_at	Pparg	5.31E-04	0.8100	0.0803	U10374	peroxisome proliferator activated receptor gamma
92443_i_at	Zfp1	6.44E-04	0.3970	0.0857	X16493	zinc finger protein 1
101034_at	Grb2	8.17E-04	0.9290	0.8740	U07617	growth factor receptor bound protein 2
92935_at	Cbfa2t1h	1.41E-03	0.3190	0.9770	D32007	CBFA2T1 identified gene homolog (human)
100553_at	Trim27	1.44E-03	9.87E-04	0.9300	L46855	tripartite motif protein 27
92782_at	Tmpo	1.47E-03	0.1180	0.0935	U39074	thymopoietin
98030_at	Trim30	1.49E-03	0.0152	0.0239	J03776	tripartite motif protein 30
161067_at	Ifld2	1.51E-03	0.0533	0.9680	AA770736	induced in fatty liver dystrophy 2
97969_at	Nr1h4	1.69E-03	0.5910	0.1250	U09416	nuclear receptor subfamily 1, group H, member 4
92974_at	Zfp37	1.93E-03	0.0189	0.5500	X52533	zinc finger protein 37
103437_at	Zfp57	1.94E-03	7.36E-03	0.3050	D21850	zinc finger protein 57
98963_at	Trpv2	2.24E-03	1.68E-04	0.2550	AB021665	transient receptor potential cation channel, subfamily V, member 2
103387_at	Tctex3	2.34E-03	0.7230	0.8710	AB011550	t-complex testis-expressed 3
103288_at	Nrip1	2.39E-03	0.8420	0.1980	AF053062	nuclear receptor interacting protein 1

AFFYID	Gene	DSO P-value	BG P-value	DSO*BG P-value	Accession	Description
99111_at	Skd3	2.42E-03	4.96E-04	0.3430	U09874	suppressor of K+ transport defect 3
103634_at	Isgf3g	3.24E-03	0.0176	0.4540	U51992	interferon dependent positive acting transcription factor 3 gamma
101014_at	Ifnar2	3.43E-03	0.6610	0.5790	Y09864	interferon (alpha and beta) receptor 2
101930_at	Nfix	3.89E-03	0.1860	0.9040	Y07688	nuclear factor I/X
92882_at	Rab1	4.37E-03	0.3040	0.0205	Y00094	RAB1, member RAS oncogene family
100422_i_at	Stat5a	4.79E-03	0.8530	0.8810	AJ237939	signal transducer and activator of transcription 5A
161139_f_at	Ddef1	5.01E-03	0.1770	0.0374	AV175719	Development and differentiation enhancing
93547_at	Cbfb	5.56E-03	0.6390	0.3620	L03279	core binding factor beta
99635_at	Ing4	6.92E-03	0.1950	0.6170	AI845183	inhibitor of growth family, member 4
99622_at	Klf4	7.53E-03	0.7860	0.0563	U20344	Kruppel-like factor 4 (gut)
96327_at	Skz1-pending	7.62E-03	0.8490	0.9210	AI852535	SCAN-KRAB-zinc finger gene 1
96824_at	Sox15	8.71E-03	0.2470	0.5420	AB025354	SRY-box containing gene 15
92658_at	Foxq1	9.12E-03	0.0123	0.7320	AF010405	forkhead box Q1
92652_at	Notch4	1.00E-02	0.2630	0.8840	AF030001	Notch gene homolog 4, (Drosophila)
102893_at	Pou2f1	0.0109	0.4120	0.1910	X68363	POU domain, class 2, transcription factor 1
102069_at	Mtf2	0.0110	0.2750	0.9090	S78454	metal response element binding transcription factor 2
102901_at	Six3	0.0117	0.3180	0.1610	D83144	sine oculis-related homeobox 3 homolog (Drosophila)
160160_at	Dedd	0.0122	0.1480	0.3820	AF100342	death effector domain-containing
94408_at	Nab1	0.0122	0.0413	0.5460	U47008	Ngfi-A binding protein 1
92991_at	Sp4	0.0125	0.8610	0.8900	U62522	trans-acting transcription factor 4
160225_at	LOC229906	0.0127	0.0764	0.0228	AI840450	similar to TFIIB
102039_at	Gtf2h4	0.0151	0.0475	0.5880	AI850881	general transcription factor II H, polypeptide 4
103057_at	Pold1	0.0157	0.8260	8.50E-03	AF024570	polymerase (DNA directed), delta 1, catalytic subunit (125kDa)
93697_at	Cbx4	0.0160	0.1390	0.6870	U63387	chromobox homolog 4 (Drosophila Pc class)

AFFYID	Gene	DSO P-value	BG P-value	DSO*BG P-value	Accession	Description
101529_g_at	Tcea1	0.0162	0.8100	0.1970	D00925	transcription elongation factor A (SII) 1
93546_s_at	Cbfb	0.0167	0.7900	0.7330	D14572	core binding factor beta
93250_r_at	Hmgb2	0.0169	0.0227	0.3330	X67668	high mobility group box 2
99169_at	Carm1-pending	0.0171	0.5330	0.9100	AW122165	coactivator-associated arginine methyltransferase 1
100554_at	Pdlim1	0.0174	0.1030	0.4340	AF053367	PDZ and LIM domain 1 (elfin)
102363_r_at	Junb	0.0177	0.3810	0.1540	U20735	Jun-B oncogene
102344_s_at	Tcea3	0.0177	0.3560	0.3600	AI132239	transcription elongation factor A (SII), 3
103354_at	Mrps31	0.0180	0.3960	0.1850	Z46966	mitochondrial ribosomal protein S31
102362_i_at	Junb	0.0190	0.4720	0.1570	U20735	Jun-B oncogene
98790_s_at	Meis1	0.0197	0.0664	0.1820	U33629	myeloid ecotropic viral integration site 1
102048_at	Crap	0.0214	0.5020	0.7700	AF041847	cardiac responsive adriamycin protein
102955_at	Nfil3	0.0215	0.2560	0.9560	U83148	nuclear factor, interleukin 3, regulated
94821_at	Xbp1	0.0228	0.8510	0.0321	AW123880	X-box binding protein 1
162016_f_at	Foxc2	0.0233	0.8380	0.3240	AV251191	forkhead box C2
97550_at	Hdac7a	0.0243	0.1080	0.6180	AW047228	histone deacetylase 7A
98336_s_at	Recc1	0.0251	0.8700	0.3500	M88489	replication factor C, 140 kDa
102917_at	C2ta	0.0274	0.1360	0.6860	AF042158	class II transactivator
94198_at	Ppard	0.0292	0.9200	0.2870	L28116	peroxisome proliferator activator receptor delta
102671_at	Creb1	0.0301	0.1040	0.7990	X67719	cAMP responsive element binding protein 1
103283_at	Elf5	0.0304	0.7240	0.7790	AF049702	E74-like factor 5
99891_at	Pou6f1	0.0311	0.0806	0.0432	L13763	POU domain, class 6, transcription factor 1
97157_at	Nkx3-1	0.0313	0.8600	0.2160	U88542	NK-3 transcription factor, locus 1 (Drosophila)
97695_s_at	Rpl7	0.0314	0.0935	0.5210	M29015	ribosomal protein L7
162204_r_at	Notch1	0.0317	0.5140	0.0186	AV374287	Notch gene homolog 1, (Drosophila)
92956_at	Notch3	0.0336	0.4340	0.5820	X74760	Notch gene homolog 3, (Drosophila)
160833_at	Mbd2	0.0341	0.3290	0.3470	AF072243	methyl-CpG binding domain protein 2
93856_at	Wt1	0.0346	0.0486	0.4230	M55512	Wilms tumor homolog
93008_at	Lsm4-pending	0.0351	0.7710	0.8120	AW120557	U6 snRNA-associated SM-like protein 4

AFFYID	Gene	DSO P-value	BG P-value	DSO*BG P-value	Accession	Description
101902_at	Rbpsuh	0.0357	0.4990	0.7680	X17459	recombining binding protein suppressor of hairless (Drosophila)
95297_at	Hoxa1	0.0370	0.6240	0.5790	M22115	homeo box A1
93918_at	Taf9	0.0378	0.8270	0.7010	AA673500	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32 kDa
98726_at	Pgr	0.0383	6.06E-03	0.9240	M68915	progesterone receptor
102364_at	Jund1	0.0402	3.77E-03	0.1390	J04509	Jun proto-oncogene related gene d1
97159_at	Aire	0.0417	0.7690	0.5500	AJ007715	autoimmune regulator (autoimmune polyendocrinopathy candidiasis ectodermal dystrophy)
101903_at	Bop	0.0418	0.8450	0.2140	U76371	CD8beta opposite strand
160068_at	Sap30	0.0420	0.3080	0.7700	AF075136	sin3 associated polypeptide, 30kD
160780_at	Tcf3	0.0441	0.0482	0.4210	AJ223069	transcription factor 3
98040_at	Tcfe2a	0.0450	0.1240	0.3430	D16631	transcription factor E2a
98150_at	Rab11b	0.0462	0.3230	0.2510	L26528	RAB11B, member RAS oncogene family
98816_s_at	Evx1	0.0476	0.5320	0.3180	AW049988	even skipped homeotic gene 1 homolog
97994_at	Tcf7	0.0497	0.2030	0.2870	AI019193	transcription factor 7, T-cell specific

Table 3. *SNP discovery and allelic imbalance results in 130 genes.*

Association between gene expression and DSO	Total number of genes tested	Number of genes with SNPs	SNPs per gene per base pair sequenced	Number of genes with AI
Positive	50	27	2.21E-03	17
Negative	80	22	0.79E-03	5

SNPs discovered by sequencing 1000bp of 3' UTR of genes randomly selected from sets with and without association between DSO and gene expression. Association is significant if $P<0.05$, using a one-tailed Fisher exact test. Genes were called positive for AI if at least one SNP displayed significant differential expression between alleles in a comparison of peak heights in gDNA versus cDNA from F1 mice ($P<0.05$, paired one-tailed Student's t-test, n=5 F1 replicates) and no contradictory results between SNPs (where multiple SNPs show opposite allele ratios in the same gene). Association between genes with expression associated with DSO and AI is significant $P<0.01$, one-tailed Fisher exact test.

FIGURE LEGENDS

Figure 1. Overview of the approach for the detection of genes with putative cis-regulatory variation. AI is detected using a sequence-based approach previously described[172]. On the right side of the figure, allelic imbalance is illustrated using a sequencing-based assay. Sequence traces obtained using genomic DNA from A/J (upper left) and C57BL6 (lower left) parental lines show a G/A SNP (black arrows). Differences in relative peak heights in sequence traces from AxB F1 genomic (upper right) and cDNA (lower right) reflect allele-specific differences in transcript levels.

Figure 2. Demonstration of rules for assigning donor strain of origin (DSO) for genes using SSLP data. SSLPs and oligonucleotide probesets were aligned to the UCSC Feb 2003 mouse genome assembly. Probesets flanked by SSLPs originating from the same parental strain were assigned the same DSO. Genes flanked by SSLPs from different parents were assigned an unknown DSO status.

Figure 3. The percentage of unknown DSO assignments versus number of segments (recombinations); the percentage of unknown DSO assignments was calculated over the entire genome for each of the 44 strains contained in the original RCS genotyping dataset. The number of congenic segments was calculated over the entire genome for each of the 44 strains.

Figure 4. QQ plot of P-values obtained from ANOVA testing the model:

Expression ~ DSO + BG +DSO*BG. Experimental P-values are sorted and plotted versus the sorted P-values of a normal distribution.

Figure 5. Fitted values of logistic regression for SNP occurrence and frequency versus P-value of association between gene expression and DSO. Experimental P-values are sorted and plotted versus the sorted P-values of a normal distribution.

Figure 6. Demonstration of allelic imbalance using a sequencing-based assay. Sequence traces obtained from 1kb of 3' UTR using genomic DNA from A/J (upper left) and C57BL6 (lower left) parental lines show a G/A SNP (black arrows). Differences in the relative peak heights of the two alleles in sequence traces from AxB F1 genomic (upper right) versus cDNA (lower right) were used to calculate the allele-specific differences in transcript levels.

Figure 7. Fitted values of logistic regression for occurrence of AI versus P-value of association between gene expression and DSO in the ANOVA model. Experimental P-values are sorted and plotted versus the sorted P-values of a normal distribution.

Figure 8. Genes ranked by P-value for DSO in the ANOVA model where $P<0.05$ are indicated in red and $p>0.05$ in blue. A) Randomly selected genes for resequencing indicated by crosses. B) Genes found to contain SNPs indicated by closed circles. C) Genes showing AI indicated by triangles.

Figure 9. Summary of AI results expressed as a percentage of genes tested from sets of genes positive or negative for association between DSO and expression.

Figure 1.

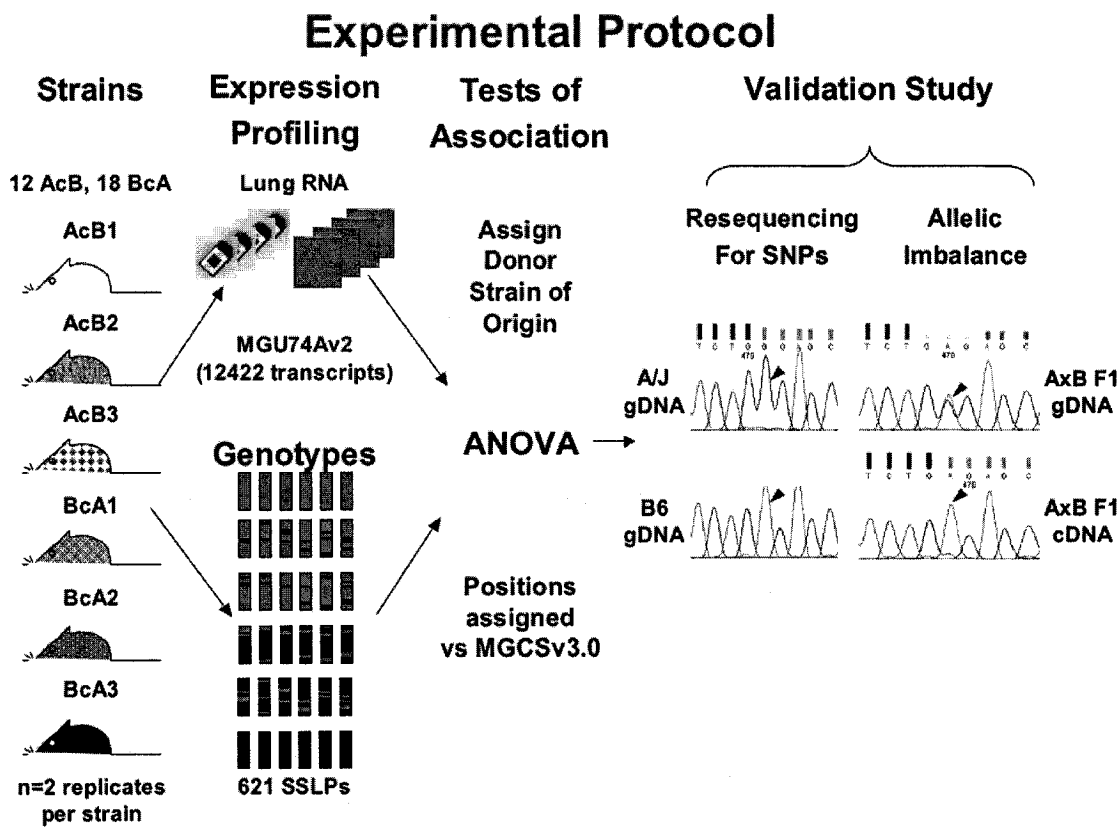

Experimental Protocol
Strains
Expression Profiling
Tests of Association
Validation Study
12 AcB, 18 BcA
Lung RNA
AcB1
AcB2
AcB3
BcA1
BcA2
BcA3
MGU74Av2
(12422 transcripts)
Genotypes
621 SSLPs
n=2 replicates per strain
Assign Donor Strain of Origin
ANOVA
Positions assigned vs MGCSv3.0
Resequencing For SNPs
Allelic Imbalance
A/J gDNA
AxB F1 gDNA
B6 gDNA
AxB F1 cDNA

Figure 2.

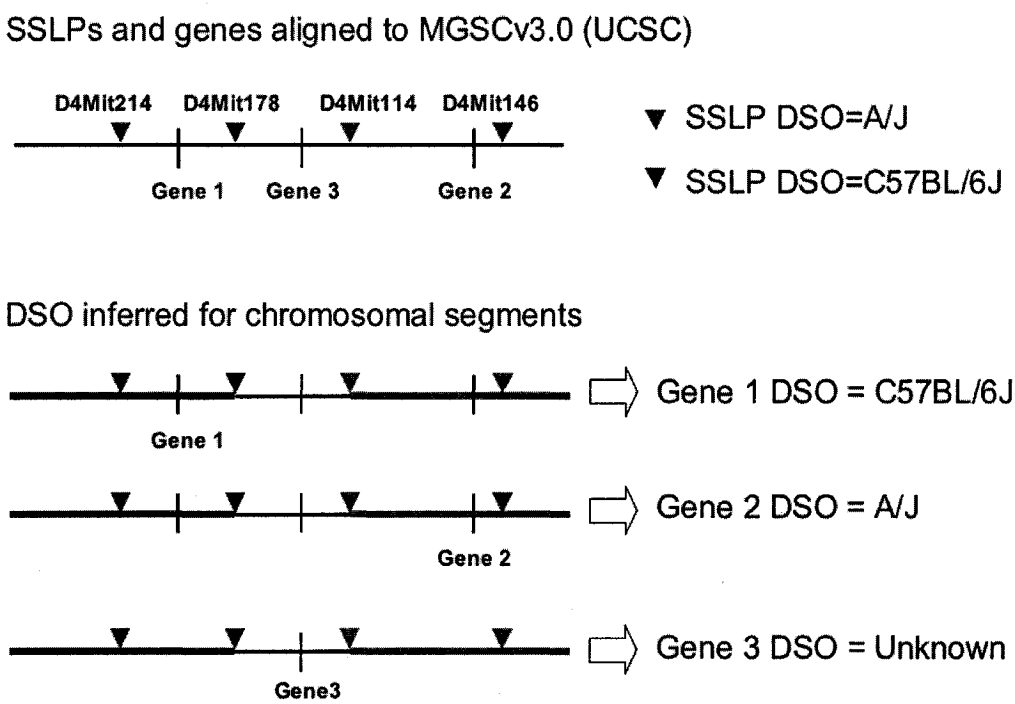

SSLPs and genes aligned to MGSCv3.0 (UCSC)
D4Mit214
D4Mit178
D4Mit114
D4Mit146
Gene 1
Gene 3
Gene 2
SSLP DSO=A/J
SSLP DSO=C57BL/6J
DSO inferred for chromosomal segments
Gene 1
Gene 1 DSO = C57BL/6J
Gene 2
Gene 2 DSO = A/J
Gene3
Gene 3 DSO = Unknown

Figure 3.

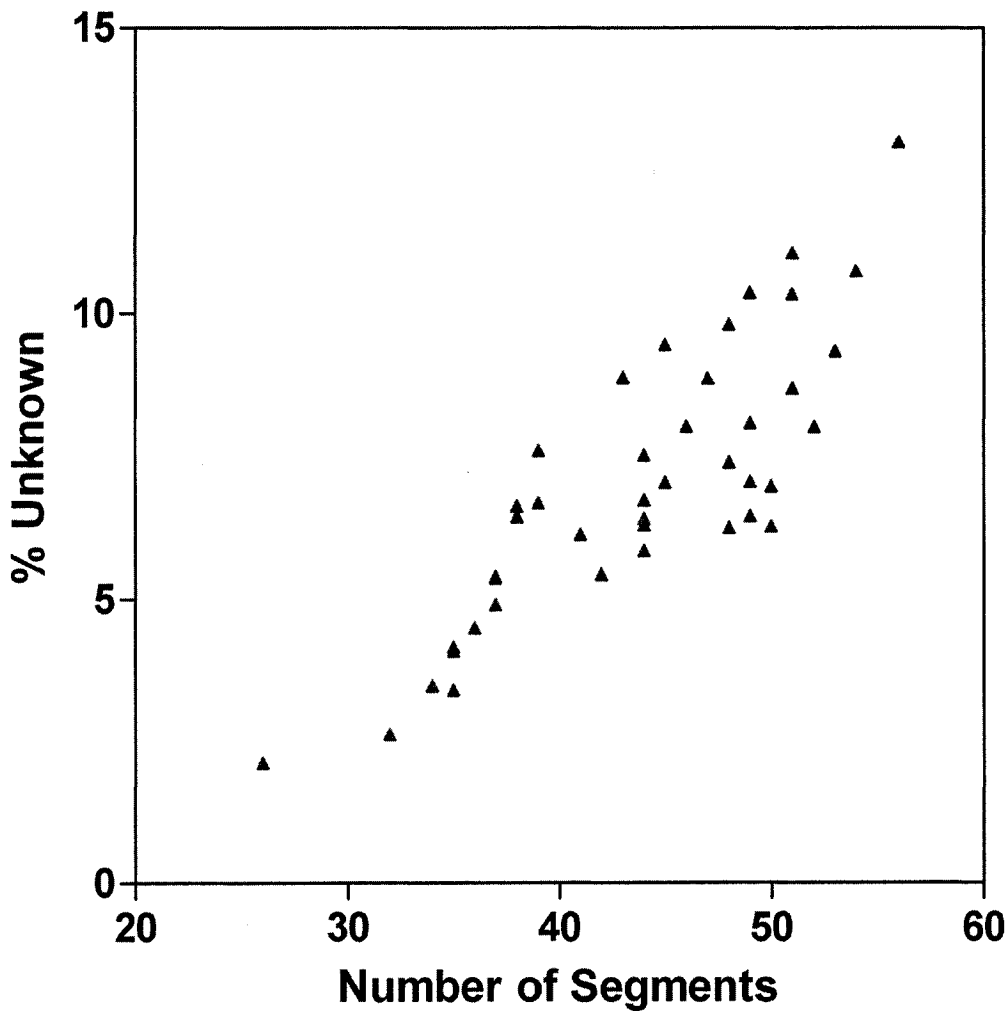


Figure 4.

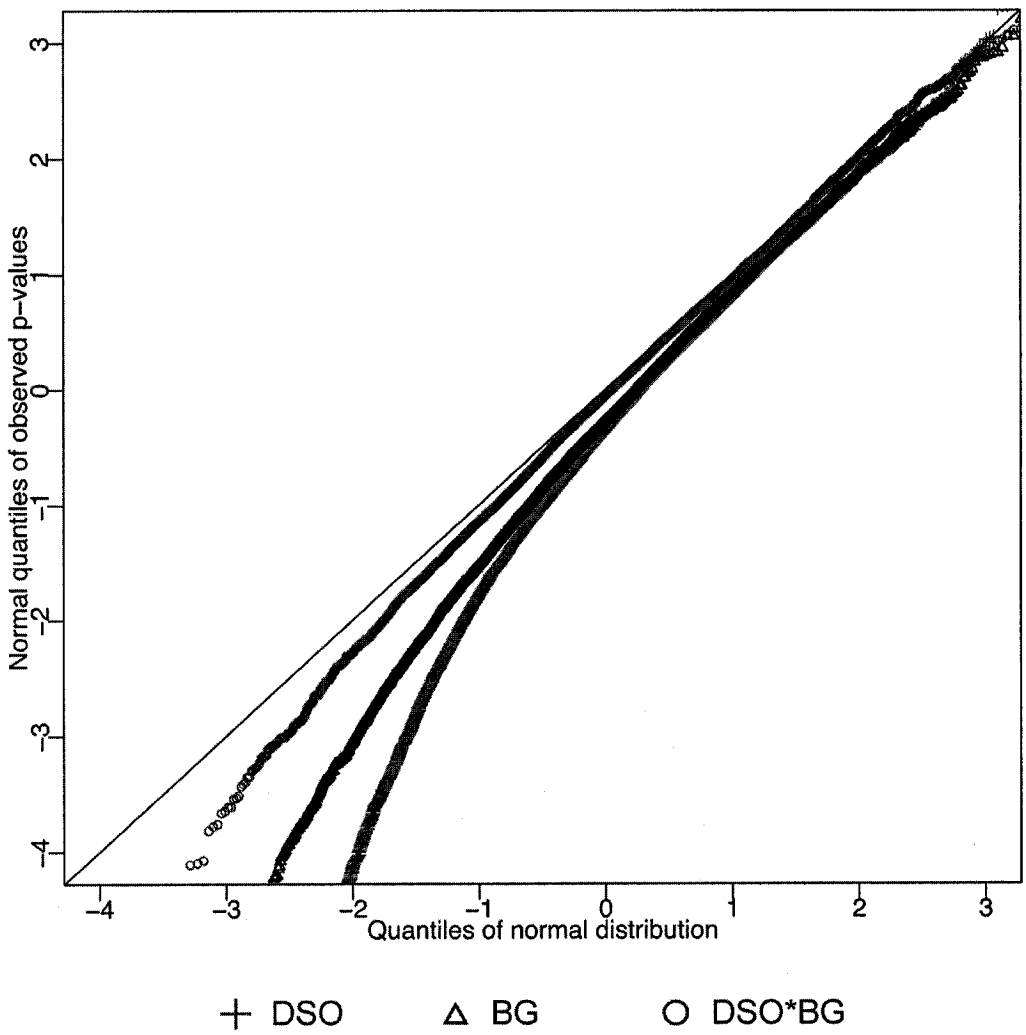

Normal quantiles of observed p-values
Quantiles of normal distribution
+ DSO
Δ BG
O DSO*BG

Figure 5.

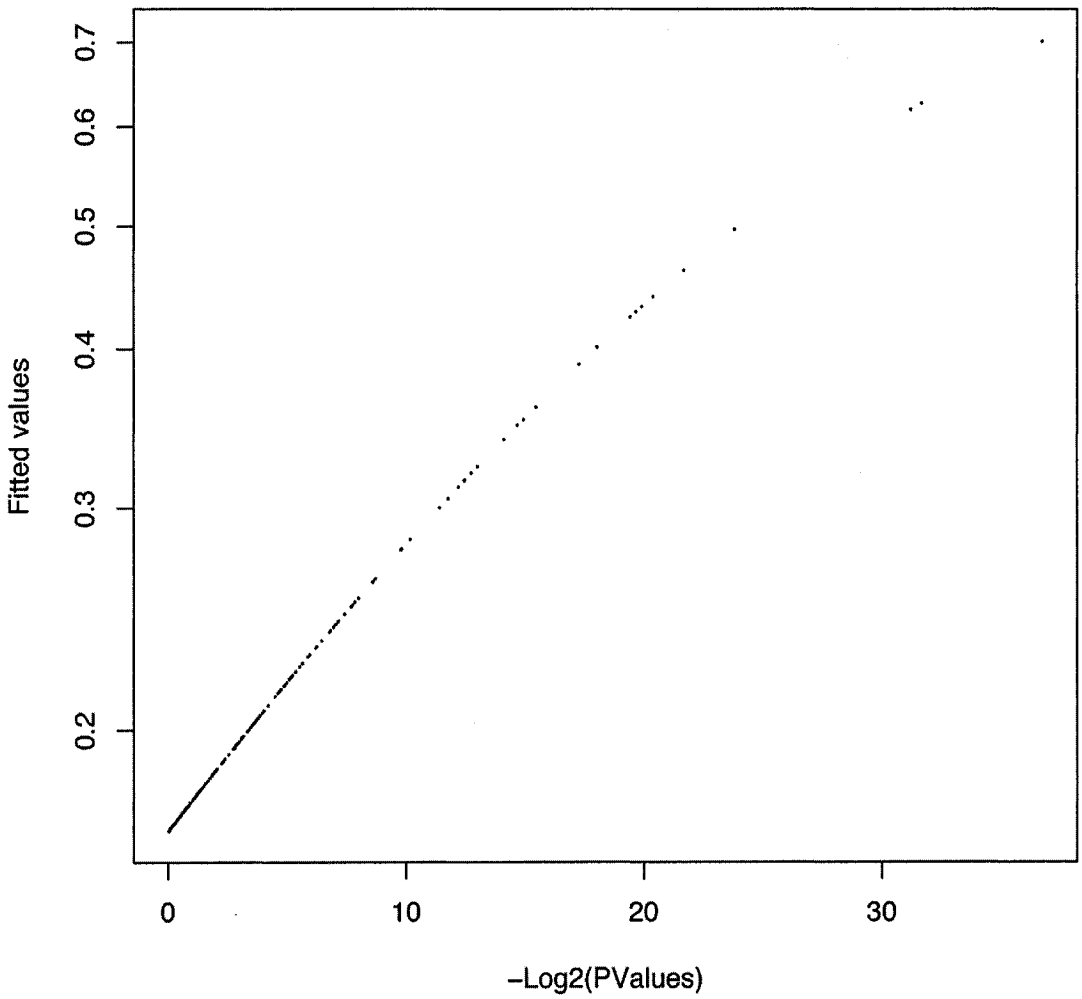


Figure 6.

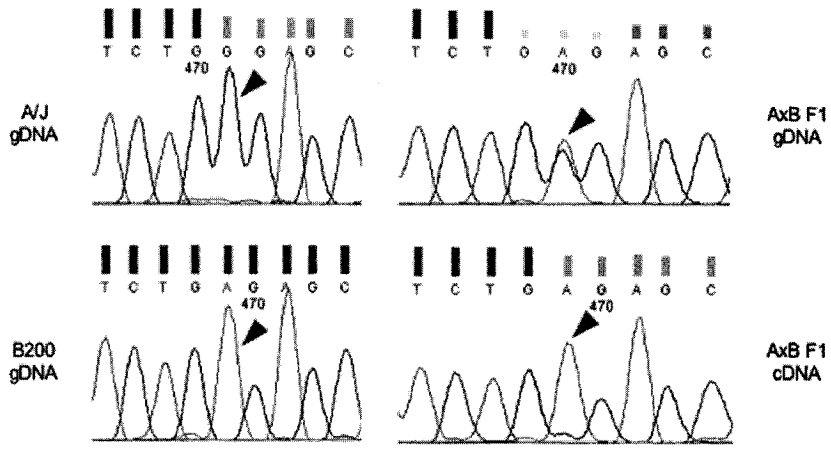

A/J
gDNA
B200
gDNA
AxB F1
gDNA
AxB F1
cDNA
T C T G G G A G C
T C T G A G A G C
T C T G A G G A G C
T C T G A G A G C
470

Figure 7.

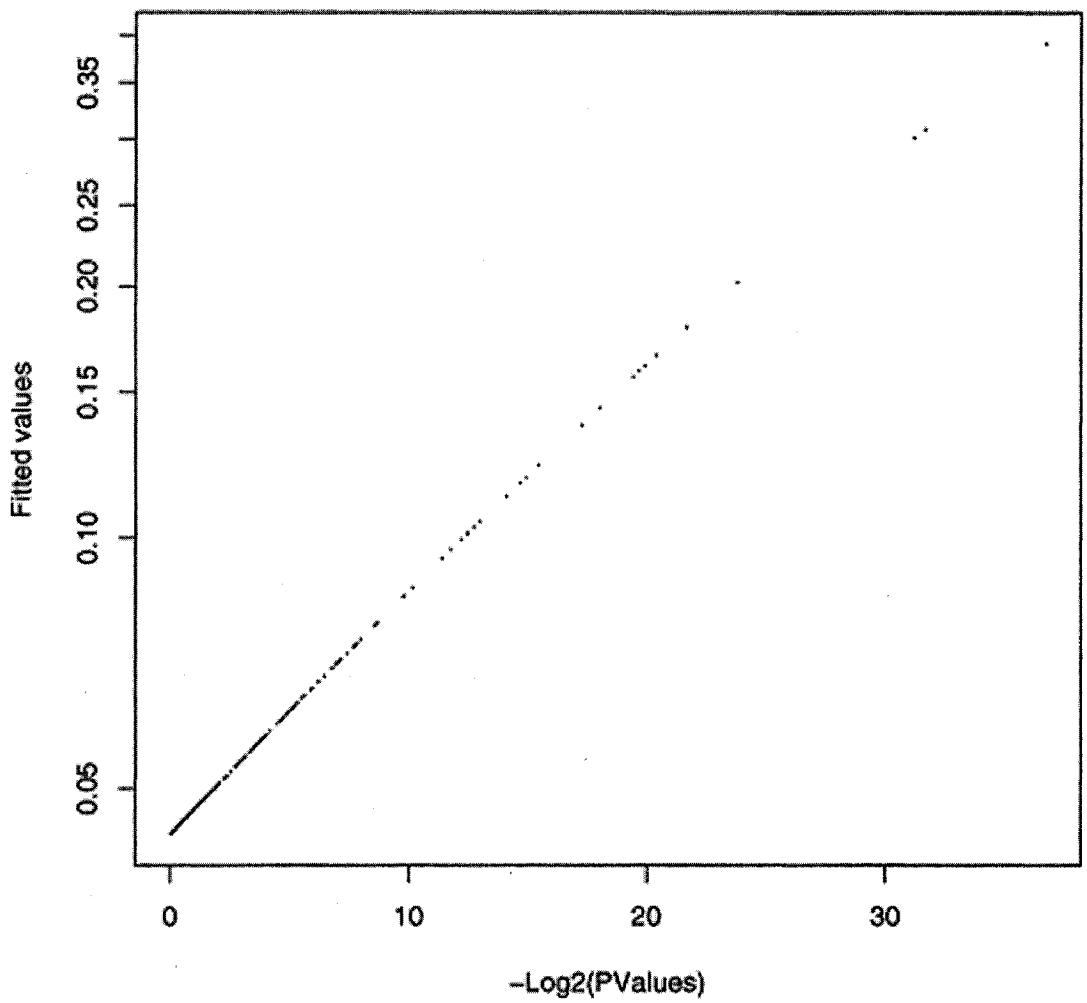


Figure 8A.

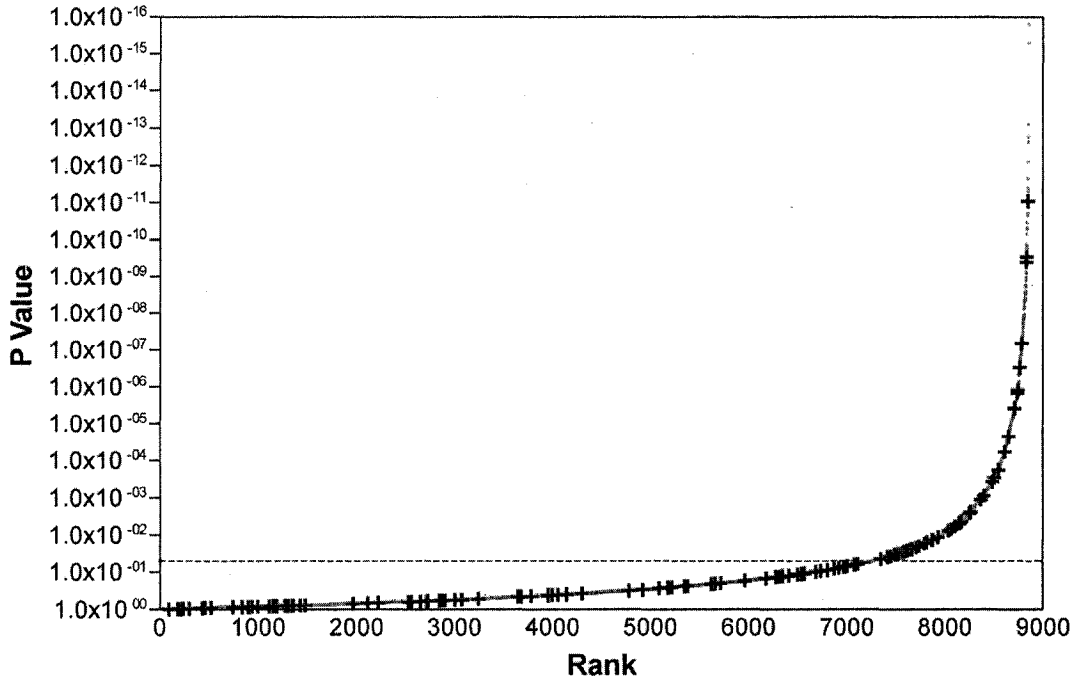


Figure 8B.

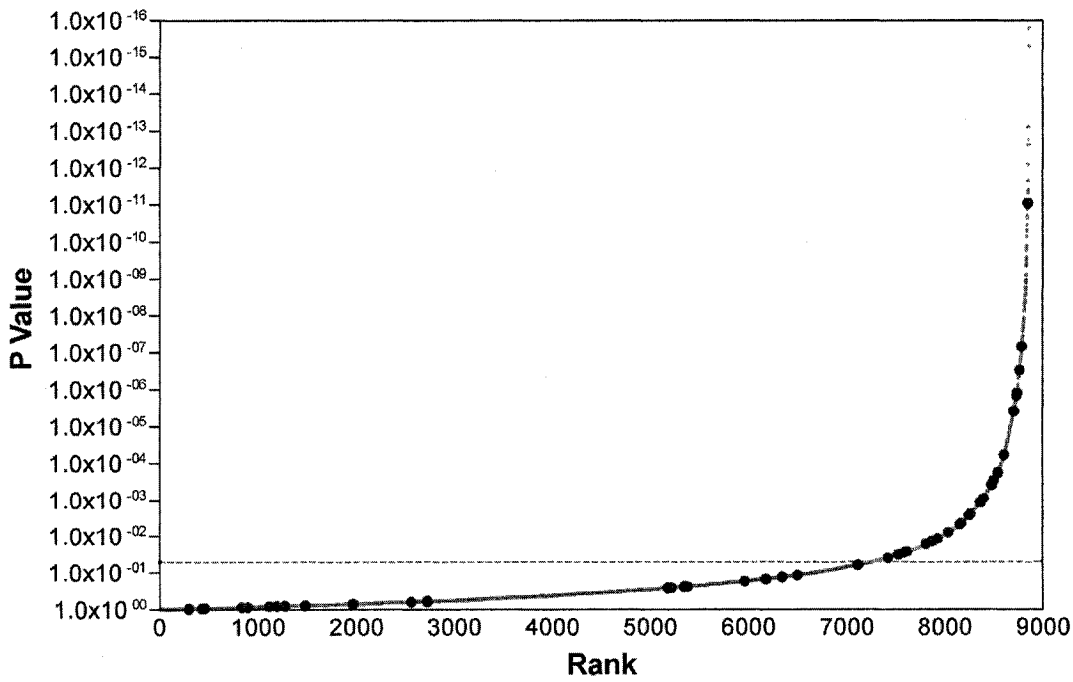


Figure 8C.

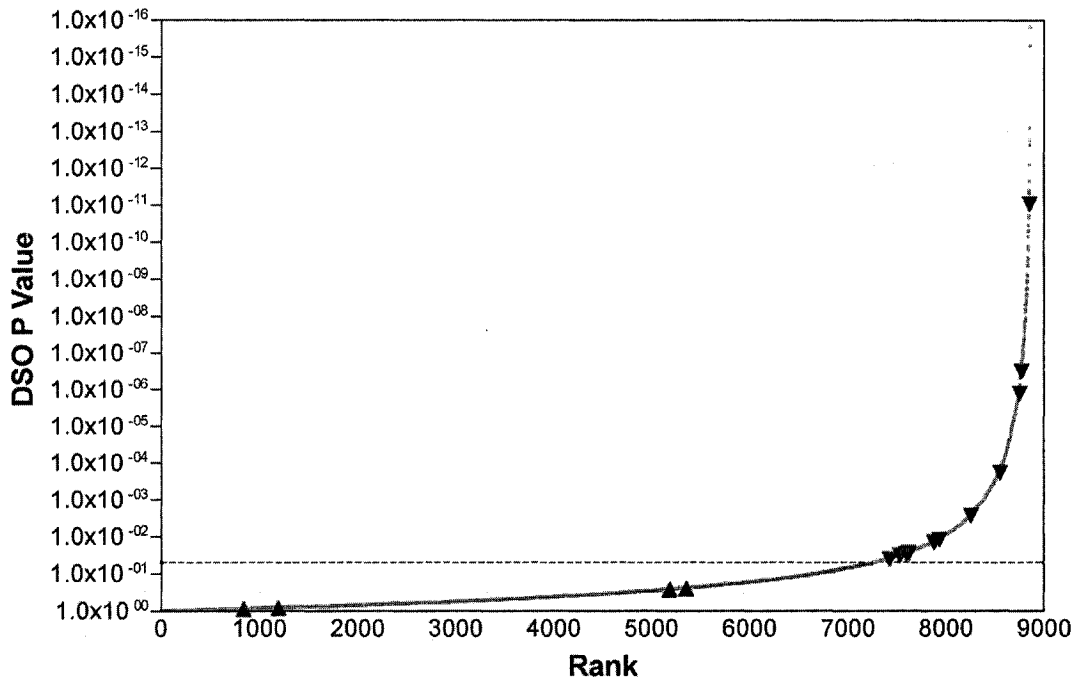
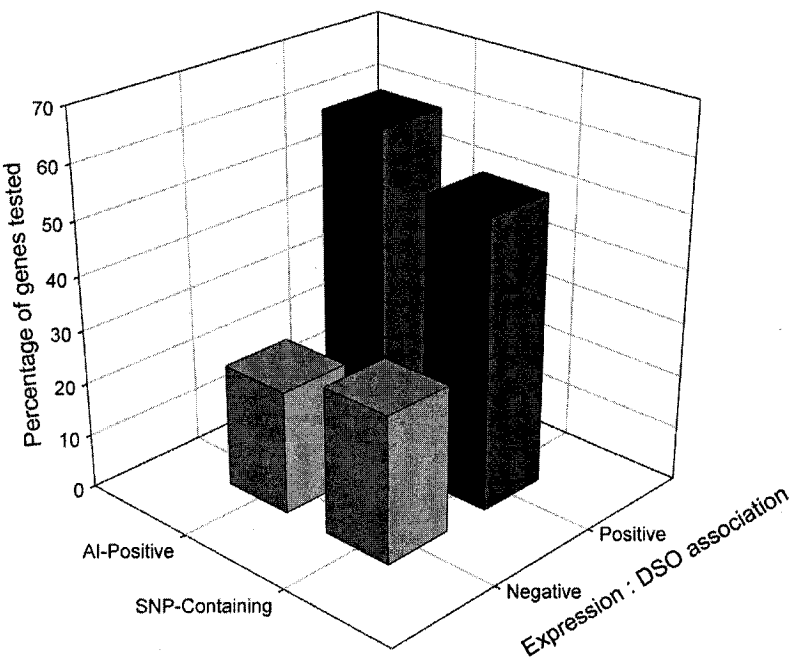

DSO P Value
1.0x10 -16
1.0x10 -15
1.0x10 -14
1.0x10 -13
1.0x10 -12
1.0x10 -11
1.0x10 -10
1.0x10 -09
1.0x10 -08
1.0x10 -07
1.0x10 -06
1.0x10 -05
1.0x10 -04
1.0x10 -03
1.0x10 -02
1.0x10 -01
1.0x10 00
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Rank

Figure 9.


70
60
50
40
30
20
10
0
Percentage of genes tested
AI-Positive
SNP-Containing
Positive
Negative
Expression : DSO association

GENERAL DISCUSSION

The research in this thesis has provided a framework for understanding microarray data analysis, how to apply this understanding to the design of expression profiling experiments, and how to extend their application by combining genetic and genomic approaches. Applications focused on inbred mice with the goal of relating gene expression differences to gene regulation in a mammalian system. This study established gene expression differences between inbred mouse strains and addressed long-standing concerns about the consistency of microarray results when applied to an in vivo mammalian system.

The goal of this thesis was to develop genomic approaches that may be used to eventually construct a genome-wide model for genetic regulation in mammalian systems. This work focused specifically on understanding what was, at the start of this degree, a novel technology (DNA microarrays) and applying these approaches to the question of gene expression and genetic variability in inbred mice. The goal of this work has always been to attempt an understanding of the biological system at a genome-wide level. The snapshots we have gained from high-throughput genomics have given us glimpses of complexity that exceeded all previous expectations. There shall be no shortage of discoveries to be made for quite some time.

Chapter 1 proposed a naïve question pertinent to experimental controls and in doing so debunked common conception and methods in practice at the time. Previously, controls in microarray experiments were accomplished by calculation of fold changes with

normalization against one or a handful of traditionally recognized control genes, a technique derived from traditional RNA assays. This study asked, are there genes that in fact do not vary across experimental systems? The answer was a resounding no. There were no genes that could be said to display a "stable" degree of expression across tissues or cell lines in an analysis of previously generated datasets, as well as in a comparison of tissues obtained from mice. Prior to this, it was thought that there existed a set of genes, ubiquitously expressed in all tissues at a constant level. These so-called "housekeeping genes" were thought to be expressed in this manner because of their hypothesized function (that of essential maintenance). This study once and for all disproved the existence of such genes and exposed the long-standing pitfalls of relying on any one gene as an invariant control to determine expression change.

These findings impacted upon extant opinions on microarray normalization, the process whereby chip-to-chip variability is adjusted via an attempt to equalize their overall hybridization intensity. Previously it was thought that utilization of control genes or some subset of genes with lower variability would provide the optimal technique for inter-hybridization correction. This study demonstrated that there was no added improvement offered by using a smaller subset, and in fact, that this approach actually impeded control of experiments by hindering auditing of analytical steps. Since the time of this study, there have been many welcome advances in the field. The advent of probe-level analysis [196] and the development normalization methods such as quantile normalization [197] and loess [198], now in widespread use, attest to the benefits of whole dataset methods, rather than a tailored subset.

This chapter furthermore provided a picture of variability across biological test systems. Previously, experimental systems whether derived from tissue samples or cell lines were used without consideration of the different degrees of variability inherent in the systems. By comparing gene expression profiling in multiple datasets including the NCI60 cancer cell line panel [88], and the Huge Index, a panel human tissues [79], this study demonstrated differences in the inherent variability contained in these datasets. The baseline level of expression variability impacts heavily upon subsequent analyses. Where study designs at the time may have assumed a constant degree of background variability, this was the first study to draw attention to these differences and the need to first characterize such levels of variability in experimental systems.

The study was one of the first to employ a study design employing replicates (not common at the time), and one of the first to apply a gene-by-gene ANOVA to assess the effects due to multiple simultaneous factors in a study design. Prior to this, the majority of microarray studies used fold-change cutoffs to determine differentially expressed genes. Indeed, use of such cutoffs persists to this day. This study, as well as many others since[33, 35], maintain that fold change cutoffs are inadequate for microarray studies and are based on notions of gene expression change derived from conventional RNA assays, rather than statistically-valid concepts. High fold changes may not correspond to statistically significant differences (spurious results), whereas genes displaying highly significant differences may show relatively subtle changes. The methods and lessons

learned in this part of the study formed the foundation for our analysis of microarray data in all subsequent studies.

The comparison of the variability between technological and biological replicates established a precedent for later experiments. The study showed a remarkable similarity in the degree of variability between genetically identical mice raised in the same cage, and that observed between repeat hybridizations of the same RNA preparation. This initial estimation of the intra-strain variability represented an important step for subsequent studies employing replicates. Where a comparable level of variation was observed between replicates, this factor could safely be factored into the error terms of ANOVA models. This did not abrogate the need for replicates, but rather provide greater degrees of freedom with which to measure associations with experimental treatment variables. More formal analysis has since examined the effect of replicate number upon the application of parametric analysis methods[112]. In exposing flaws in the extant concepts of control genes, normalization and methods to define differentially expressed genes, this study demonstrated the need for improvements in these areas, many of which have since been addressed by the field as a whole.

Where Chapter 1 demonstrated analytical methods and the need to characterize the baseline variability in any experimental systems, chapter 2 focused on applying these principles to the model system used throughout the remainder of the thesis, inbred mice strains A/J and C57BL/6J. In characterizing the baseline gene expression variability between strains across 4 tissues, this study was able to show substantial differences in

gene expression between these two strains, echoing the differences observed between these strains at the physiological level. Characterization of the extent of expression variability in healthy, untreated adult mice of each strain establishes the context in which subsequent studies of pathological states may be compared. At the start of this study, experiments comparing normal subjects were rare with most studies focusing on pathological conditions. Since then, the importance of knowing the baseline variability in normal tissues has been acknowledged as a fundamental first step towards studying the diseased state[199]. In mice, where disease phenotypes are studied by comparison between strains, baseline gene expression differences between strains remain largely uncharacterized. This baseline variability may act as a confounding variable in studies attempting to relate gene expression to disease phenotypes. Expression profiling across more tissues and strains will likely become necessary as more disease phenotypes are studied at the level of gene expression.

This chapter also addressed a long-standing concern with microarray studies, mainly that of reproducibility over time. While experiments repeated one year apart were performed in a controlled manner to the best of our knowledge, time-specific effects were seen to be substantial. In spite of rigorous controls, other variables yet unidentified could play a role in the variability observed from one time point to the next. The multitude of nonautomated procedures separating the lab animal from the final hybridization intensity measurement renders determination of each source of variability difficult. However, this study represents one of the first to estimate the magnitude of inter-experimental variability in an *in vivo* microarray experiment, and the first to demonstrate the ability to

adjust for the confounding effect of this variable. This study furthermore identified a biological component to the time variable in the experiment, in particular, the tissue-time and the tissue-strain-time interaction terms showing significant contributions to the overall variability observed in the experiment. This result cautions future experimenters about the sensitivities of the different tissues to inter-experimental variability, an issue affecting studies involving simultaneous comparisons of multiple tissues.

Across the 4 tissues, over 750 genes localized within previously mapped QTL. While these included several likely candidates identified by previous studies, the large number of genes suggests that gene expression profiling on its own may not be sufficient to determine the genes underlying QTL. Gene expression changes do not automatically imply involvement of these genes in a phenotype. The complex mixture of gene expression changes resulting from direct and indirect activation insures that this remains a difficult hypothesis to prove. Likewise, subtle variations in expression, typical of genes whose transcription is tightly regulated (e.g. negative feedback regulation), will remain unidentified in expression studies because such changes fall outside the detection limits of the technology. Because gene interactions are believed to exist to such a high degree, correlations between differential expression and its causes shall remain difficult to identify. By its very design, this experiment did not attempt to address potential genetic causes of the observed expression variability. Instead, Chapter 2 established the baseline expression variability for future comparisons of experimental treatments or genetic recombinants involving A/J vs. C57BL/6J comparisons.

Whereas Chapter 2 provided indirect evidence of an association between expression differences across a system known to contain genetic variation, Chapter 3 provided direct evidence. Here, association of gene expression variability with the genetic composition of the surrounding locus provided compelling evidence for a correlation between transcription levels and cis-acting genetic variants. Application of an independent genomic approach, allelic imbalance, further confirmed hypotheses of cis-acting gene regulation for these genes. The RCS panel represents an experimental system controlled for a level of genetic heterogeneity permitting the differentiation of proximal and distal factors affecting gene expression.

Genetic variants underlying complex traits may affect the quality of the gene product (via alterations in protein coding sequences), or quantity (altering the level of transcript). While numerous genetic variants affecting protein sequence have been identified, far fewer regulatory variants have been isolated, mainly due to the lack of methods to find such variants in large-scale. Observations of large intergenic regions in the mammalian genome and the heritability of gene expression differences between individuals suggest many of the variants underlying complex traits are regulatory affecting gene expression. While gene regulation appears to function in a highly coordinated fashion across the genome, the picture of gene regulation on a genome-wide scale is sparsely populated with a minority of genes that are functionally characterized. Integrated approaches are required to better characterize gene regulation and to understand the genetics of complex traits.

The combination of gene expression profiling in experimental systems traditionally used in genetic research exposes new avenues that would not be accessible through either technology alone. Where genetic mapping studies have traditionally dealt with the problem of narrowing down phenotypic associations observed over large regions of the genome, expression profiling immediately brings the focus of study to the functional molecular level. Knowledge of genetic variation enables us to explore the potential causes of differential gene expression and to begin addressing questions of gene regulation on a genome-wide scale. One criticism of genomic approaches pertains to their validity with respect to previously documented evidence. Inbred mice offer a wealth of information accumulated over decades of research at the phenotypic level providing a link to observations at the molecular level. A combination of genomic approaches may furthermore provide independent methods for verification of results. Our integration of expression profiling in RCS mice together with screening for allelic expression differences represents such an approach.

Gene regulation is currently classified into categories of cis-acting and trans-acting, differentiating variants located proximally versus distally to the affected gene. The distinction proves useful in genetic research where cis-acting factors may display simple modes of inheritance while trans-acting factors correspond to more complex inheritance. Characterizing the extent of each form of regulation across the genome may prove informative in fine-mapping strategies and efforts to decipher traits arising from numerous loci. To date, genome-wide mapping of either form of regulation remains to be done. However, cis-acting variants may be more readily mapped by combined genomic

and genetic approaches. Trans-acting variants describing the epistatic interactions between genes are more difficult to ascertain, but are believed to predominate. Whereas the study in Chapter 2 identified differential expression containing a complex mixture of cis and trans effects, the RCS panel used in Chapter 3 offered the unique opportunity to distinguish differential expression due to cis or trans effects. In the future, this strategy may help prioritize genes for further empirical study and functional annotation. This classification of genes may provide the first steps towards building a model of genome-wide gene regulation.

We have demonstrated the feasibility of an integrated genomic approach suitable for genome-wide cataloguing of cis-acting regulatory effects. As awareness of gene regulatory variants and gene interactions in complex phenotypes increases, so shall the need to chart gene regulation on a genome-wide scale. Recent studies estimate between 25 and 50% of genes in humans display evidence of allelic expression indicative of cis-regulation[135]. A random screening study across 3 tissues in 4 mouse inbred strains previously found allelic expression differences in 3-6% of genes tested, depending on the combination of tissue and strain[135]. This study arrived at a slightly higher estimate of 8-11% of genes. This falls close to the latest estimates in yeast[28] and humans[30]. While more accurate estimations await the development of higher resolution SNP maps and whole-genome arrays, this study demonstrates the efficiency of the approach towards finding the cis-regulated genes in the mammalian system.

One of the primary questions regarding complex traits is whether a small number of loci account for the majority of variation, or whether it is due to a large number, each contributing a small amount. Furthermore, are the effects of multiple QTL additive, or do they interact in a nonadditive way? This comes from the observations that most heritable traits show a continuous spectrum of variation across a population[200]. Judging from recent studies, mostly in yeast, the picture is likely a mixture of all mechanisms in varying proportions. A recent study mapping gene expression traits in yeast determined that 3% of transcripts demonstrated heritability consistent with a single-locus model, 17-18% with a two-locus model, and greater than 50% with extensive genetic complexity consistent with 3 or more loci[28]. Descriptive work of this kind enables a map to be generated of the overall topology of the genetic network, establishing a framework for future discussion, and provides a rough model of genetic regulation across the genome. Determining the same distributions for more complex organisms such as mice shall be the logical next step.

From the onset, it appears that the majority of gene regulatory variations are believed to be trans-acting or epistatic, involving the interactions between multiple genes. Owing to the number of genes in the genome, analytical determination of these interactions is extremely challenging[201]. The number of possible outcomes increases exponentially with number of interacting genes, and study designs often suffer from inadequate sample sizes to account for the combinatorial explosion[202]. The task is further complicated by the question of multiple hypotheses testing, which is magnified exponentially in a test for trans effects. The number of permutations required to test so many hypotheses renders

analyses to be computationally intensive involving compute farms with hundreds of processors. Finally, so little is known about the trans-acting network on a genome-wide scale that the very dimensionality of the dataset remains to be determined – how many variables are important at any given time, and in what proportions? Directional or more intricate logical dependencies between genes and regulatory elements are even more difficult to delineate. If the level of complexity is anything like that seen in simpler model systems such as yeast[203, 204], the task of sorting out trans-interaction networks is likely to remain a challenge for years to come.

One goal of mapping trans-regulatory effects is to construct network diagrams of gene interactions or models of the genetic network. These are generally displayed as directed graphs where genes are represented by nodes and the interactions by edges. These interactions include any mechanism whereby the input or affector node affects the output or effector. In this sense, genes may be represented as inputs and outputs. The information that passes between genes may be complex. For example, transcriptional activation of one gene by another (where one gene's protein product binds to the promoter region of another gene) may entail additional protein-protein interactions, such as recruiting of multiple tethering factors or chaperones which may act on enhancer sites far from the actual transcription start site (e.g. activation mechanisms known to govern glucocorticoid-regulated genes[205]). While identification of the precise mechanisms of specific gene interactions may lie outside the scope of system-wide surveys, confirming the presence of such interactions on a genome-wide scale, together with determining conditional dependencies across experimental systems establishes a framework for

integrating empirical studies from varied sources. Once individual interactions are identified between pairs of genes, the next step becomes to characterize patterns of interactions among more genes, or network motifs. In simpler experimental systems, a limited number of motifs have been observed to exist with varying degrees of frequency[206-208]. Diagrams of such subnetworks may enable the understanding of specific systems and their context dependence such as sets of genes underlying disease phenotypes. Are there motifs that are more highly represented in disease versus health? When such diagrams are assembled on a genome-wide scale, the model of how the genome functions as a whole shall begin to take shape.

Several analytical methods have been proposed to address the issue of trans-acting gene regulation and gene interactions. One method is based on decision trees (C4.5) involving the calculation of conditional entropy. Here, gene expression profiles are taken as an outcome variable and compared in every pairwise comparison against the DSO profiles (attributes) of every other gene. Information gain is calculated as the decrease in entropy, or amount of heterogeneity in the expression data that is explained by the other genes. The method may be expanded to incorporate compound attributes consisting of both DSO and expression profiles for genes against which any given gene's expression may be compared. Tree construction occurs by progressive splitting of the expression data with the most informative nodes located higher up on the tree. Other comparable methods for determining interactions between variables include classification and regression trees[209] (CART), random forests[210], and multifactorial dimensionality reduction[211] (MDR).

Another approach is to utilize gene expression profiles as a quantitive trait in a whole genome scan[212, 213] (eQTL mapping), and to examine patterns of inheritance. Owing to the number of traits in these expression profiles (>12000), this would be quite computationally intensive, not merely for single-gene associations but for gene-gene interactions. A typical study involving 1000 permutation tests over 12488 genes is calculated to take 100,000 CPU hours (over 8 days using a multi-processor machine with >500 processors)[201]. However, as computing resources increase exponentially, this analysis is rapidly becoming more accessible. An analysis of the RCS data using such an approach is currently in progress.

The concept of a gene network was first coined over 30 years ago[7, 214, 215] has only recently become testable experimentally, even in the broadest sense. Estimation of interaction networks in yeast has proceeded from several angles including chromatic immunoprecipitation assays[20], Bayesian network inference[27], and expression QTL mapping[28, 169, 216], using panels of single or double gene deletion strains. Studies of similar scale have yet to be realized in higher multicellular organisms, largely due to the increased cost and experimental complexity associated with such systems. The studies in yeast arrive at similar conclusions regarding the overall topology of the network; a minority of genes displays a simple connectivity, and a majority exhibits complex dependencies between multiple genes. If these observations reflect general principles obeyed by all gene networks, then we should see similar distributions in multicellular organisms.

An eventual goal of forming a model of gene regulation is to create dynamic simulations in order to predict effects of variants or perturbations on the network and the system as a whole. While an ambitious goal, the first steps in direction have begun in isolated experimental systems[217]. A major obstacle currently faced by such studies is the level of noise seen in biological systems at the molecular level. To circumvent some of these challenges, subnetworks of genes have been specifically engineered to allow measurement of transcription rates[218-221] in single cells. Obviously much remains to be discovered. However, the study of gene regulation dynamics inside a living cell remains an exciting area of study under development.

CONCLUSIONS

As complete genome sequences become available for more model organisms, similar high-throughput approaches shall be required to progressively map gene regulation throughout the genomes of many species. This shall be a gradual process. This thesis represents one step. The approach outlined could be expanded to catalogue cis-regulation on a genome-wide scale. This shall likely be completed for a number of model systems within the next 5 years, and a picture of trans-regulation is likely to follow shortly thereafter. Gradually, with time and luck, these and other such efforts shall populate the model of gene regulation throughout the genome. The process will most likely yield unexpected discoveries, perhaps exposing behaviors of the system that we have yet to recognize. Regardless, a more complete picture of gene regulation on a genome-wide scale shall form the basis for more efficient identification of genes underlying complex phenotypes, and the molecular mechanisms of disease. Combined with further high-

throughput approaches, we may eventually be able to forge the links between the layers of complexity that comprise the phenome[8]. Ultimately, the goal of this research is to enable novel ways of thinking about biology. Understanding genome-wide networks of regulation shall require new concepts that can accommodate the scale and complexity of observations provided by genomic technologies. It has been over 100 years since Mendel demonstrated the heritability of traits[2] and since Darwin postulated evolution[1]. It has been over 50 years since Watson and Crick elucidated the structure of DNA[3] and the deciphering of the genetic code[4]. The next principles shall likely pertain to the structure of gene networks across all biological systems as they pertain to complexity and self-organization[222].

It is perhaps useful to note that, should technological development continue along its path, there should be no decrease in our ability to generate data. At the start of this degree, the Human Genome Project[223] was approaching completion. Genome sequences for *Drosophila melanogaster*[224], *Caenorhabditis elegans*[225] and *Mus Musculus*[226] were also completed during this time. Other forms of functional data such as genome-wide expression profiles are now accumulating at rates similar to sequence databases. SNP and phenotype databases are growing dramatically with the increasing focus on complex traits and diseases. The need for adequate means to analyze and integrate this data shall continue to be an issue as these resources grow. So long as the tools for efficient access to information follow suit, the growth of biological databases promises many discoveries for generations to come.

With the increasing array of genomic technologies at our disposal, the complexity of biology may eventually yield its secrets. In the face of such technological development, the link between data and knowledge hinges critically on our ability to create conceptual models to structure our observations. The integration and assimilation of genomic data shall require a framework of sufficient scale and complexity. The models shall likely be imperfect, conceptual toys allowing us to form testable hypotheses. However, I believe that these approaches are a preliminary step towards understanding biology on a genome-wide scale. This thesis has provided a glimpse of what lies ahead and an approach for journeying into the unknown.

REFERENCES

1. Darwin, C., *On the origin of species*. 1859, Cambridge, Mass. ; London: Harvard University Press.
2. Mendel, G., Gregor Mendel's letters to Carl Nageli, 1866-1873. *Genetics*, 1950. **35**(5:2): p. 1-29.
3. Watson, J.D. and F.H. Crick, The structure of DNA. *Cold Spring Harb Symp Quant Biol*, 1953. **18**: p. 123-31.
4. Crick, F.H., et al., General nature of the genetic code for proteins. *Nature*, 1961. **192**: p. 1227-32.
5. Jacob, F. and J. Monod, Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 1961. **3**: p. 318-56.
6. Monod, J., *Chance and necessity : an essay on the natural philosophy of modern biology*. 1971, Glasgow: W. Collins. 187 p.
7. Kauffman, S., Gene regulation networks: a theory for their global structure and behaviors. *Curr Top Dev Biol*, 1971. **6**(6): p. 145-82.
8. Scriver, C.R., After the genome--the phenome? *J Inherit Metab Dis*, 2004. **27**(3): p. 305-17.
9. Glazier, A.M., J.H. Nadeau, and T.J. Aitman, Finding genes that underlie complex traits. *Science*, 2002. **298**(5602): p. 2345-9.
10. Nadeau, J.H., Modifier genes and protective alleles in humans and mice. *Curr Opin Genet Dev*, 2003. **13**(3): p. 290-5.
11. Cormier, R.T., et al., Secretory phospholipase Pla2g2a confers resistance to intestinal tumorigenesis. *Nat Genet*, 1997. **17**(1): p. 88-91.
12. Hong, K.H., et al., Deletion of cytosolic phospholipase A(2) suppresses Apc(Min)-induced tumorigenesis. *Proc Natl Acad Sci U S A*, 2001. **98**(7): p. 3935-9.

13. Long, A.D., et al., Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in Drosophila melanogaster. *Genetics*, 1996. **144**(4): p. 1497-510.
14. Mackay, T.F., The genetic architecture of quantitative traits. *Annu Rev Genet*, 2001. **35**: p. 303-39.
15. Fijneman, R.J., et al., Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat Genet*, 1996. **14**(4): p. 465-7.
16. van Wezel, T., et al., Gene interaction and single gene effects in colon tumour susceptibility in mice. *Nat Genet*, 1996. **14**(4): p. 468-70.
17. Ikeda, A., et al., Microtubule-associated protein 1A is a modifier of tubby hearing (moth1). *Nat Genet*, 2002. **30**(4): p. 401-5.
18. Groot, P.C., et al., The recombinant congenic strains for analysis of multigenic traits: genetic composition. *Faseb J*, 1992. **6**(10): p. 2826-35.
19. Nadeau, J.H., Modifier genes in mice and humans. *Nat Rev Genet*, 2001. **2**(3): p. 165-74.
20. Ren, B., et al., Genome-wide location and function of DNA binding proteins. *Science*, 2000. **290**(5500): p. 2306-9.
21. Yuh, C.H., H. Bolouri, and E.H. Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 1998. **279**(5358): p. 1896-902.
22. Lewis, E.B., Pseudoallelism and gene evolution. *Cold Spring Harb Symp Quant Biol*, 1951. **16**: p. 159-74.
23. Lewin, B., *Genes VII*. 2000, Oxford ; New York: Oxford University Press. xvii, 990 p.
24. Darnell, J.E., D. Baltimore, and H.F. Lodish, *Molecular cell biology*. 2nd ed. 1990, New York: Scientific American Books : Distributed by W.H. Freeman. xl, 1105 p.
25. Buckland, P.R., Allele-specific gene expression differences in humans. *Hum Mol Genet*, 2004. **13 Spec No 2**: p. R255-60.

26. Bray, N.J., et al., Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet*, 2003. **113**(2): p. 149-53.
27. Pe'er, D., et al., Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 2001. **17 Suppl 1**: p. S215-24.
28. Brem, R.B. and L. Kruglyak, The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 2005.
29. Cheung, V.G., et al., Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 2003. **33**(3): p. 422-5.
30. Morley, M., et al., Genetic analysis of genome-wide variation in human gene expression. *Nature*, 2004. **430**(7001): p. 743-7.
31. Enard, W., et al., Intra- and interspecific variation in primate gene expression patterns. *Science*, 2002. **296**(5566): p. 340-3.
32. Ferea, T.L., et al., Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci U S A*, 1999. **96**(17): p. 9721-6.
33. Oleksiak, M.F., J.L. Roach, and D.L. Crawford, Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus. *Nat Genet*, 2005. **37**(1): p. 67-72.
34. Oleksiak, M.F., G.A. Churchill, and D.L. Crawford, Variation in gene expression within and among natural populations. *Nat Genet*, 2002. **32**(2): p. 261-6.
35. Rifkin, S.A., J. Kim, and K.P. White, Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat Genet*, 2003. **33**(2): p. 138-44.
36. Segre, D., et al., Modular epistasis in yeast metabolism. *Nat Genet*, 2005. **37**(1): p. 77-83.
37. Lee, I., et al., A probabilistic functional network of yeast genes. *Science*, 2004. **306**(5701): p. 1555-8.
38. Tamayo, P., et al., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 1999. **96**(6): p. 2907-12.
39. Eisen, M.B., et al., Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 1998. **95**(25): p. 14863-8.

40. DeRisi, J.L., V.R. Iyer, and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997. **278**(5338): p. 680-6.
41. Fodor, S.P., et al., Light-directed, spatially addressable parallel chemical synthesis. *Science*, 1991. **251**(4995): p. 767-773.
42. Lockhart, D.J., et al., Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 1996. **14**(13): p. 1675-80.
43. Rockett, J.C. and G.M. Hellmann, Confirming microarray data--is it really necessary? *Genomics*, 2004. **83**(4): p. 541-9.
44. Silver, L.M., *Mouse genetics : concepts and applications*. 1995, New York: Oxford University Press. xiii, 362 p.
45. Kozak, L.P. and M. Rossmeisl, Adiposity and the development of diabetes in mouse genetic models. *Ann N Y Acad Sci*, 2002. **967**: p. 80-7.
46. Hill, A.V., Genetics of infectious disease resistance. *Curr Opin Genet Dev*, 1996. **6**(3): p. 348-53.
47. Chu, G., K. Haghighi, and E.G. Kranias, From mouse to man: understanding heart failure through genetically altered mouse models. *J Card Fail*, 2002. **8**(6 Suppl): p. S432-49.
48. Hirst, G.L. and A. Balmain, Forty years of cancer modelling in the mouse. *Eur J Cancer*, 2004. **40**(13): p. 1974-80.
49. Bauer, A.K., A.M. Malkinson, and S.R. Kleeberger, Susceptibility to neoplastic and non-neoplastic pulmonary diseases in mice: genetic similarities. *Am J Physiol Lung Cell Mol Physiol*, 2004. **287**(4): p. L685-703.
50. Fleischman, R.A., et al., Deletion of the c-kit protooncogene in the human developmental defect piebald trait. *Proc Natl Acad Sci U S A*, 1991. **88**(23): p. 10885-9.
51. Bogue, M.A. and S.C. Grubb, The Mouse Phenome Project. *Genetica*, 2004. **122**(1): p. 71-4.
52. Beck, J.A., et al., Genealogies of mouse inbred strains. *Nat Genet*, 2000. **24**(1): p. 23-5.

53. Mural, R.J., et al., A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 2002. **296**(5573): p. 1661-71.
54. Hudson, T.J., et al., A radiation hybrid map of mouse genes. *Nat Genet*, 2001. **29**(2): p. 201-205.
55. Pletcher, M.T., et al., Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol*, 2004. **2**(12): p. e393.
56. Wiltshire, T., et al., Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc Natl Acad Sci U S A*, 2003. **100**(6): p. 3380-5.
57. Yalcin, B., et al., Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc Natl Acad Sci U S A*, 2004. **101**(26): p. 9734-9.
58. Frazer, K.A., et al., Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res*, 2004. **14**(8): p. 1493-500.
59. Wang, X., et al., Haplotype analysis in multiple crosses to identify a QTL gene. *Genome Res*, 2004. **14**(9): p. 1767-72.
60. Festing, M.F.W., *Inbred strains in biomedical research.* 1979, London: Macmillan. xii, 483 p.
61. Van Etten, W.J., et al., Radiation hybrid map of the mouse genome. *Nat Genet*, 1999. **22**(4): p. 384-7.
62. Fortin, A., et al., Recombinant congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of complex traits. *Genomics*, 2001. **74**(1): p. 21-35.
63. Wheeler, D.L., et al., Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*, 2004. **32**: p. D35-40.
64. Darvasi, A. and A. Pisante-Shalom, Complexities in the genetic dissection of quantitative trait loci. *Trends Genet*, 2002. **18**(10): p. 489-91.
65. Hitzemann, R., et al., On the integration of alcohol-related quantitative trait loci and gene expression analyses. *Alcohol Clin Exp Res*, 2004. **28**(10): p. 1437-48.
66. Lan, H., et al., Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, 2003. **164**(4): p. 1607-14.

67. Eaves, I.A., et al., Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res*, 2002. **12**(2): p. 232-43.

68. Darvasi, A., Genomics: Gene expression meets genetics. *Nature*, 2003. **422**(6929): p. 269-70.

69. Phillips, T.J. and J.K. Belknap, Complex-trait genetics: emergence of multivariate strategies. *Nat Rev Neurosci*, 2002. **3**(6): p. 478-85.

70. Mootha, V.K., et al., Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A*, 2003. **100**(2): p. 605-10.

71. Chu, S., et al., The transcriptional program of sporulation in budding yeast. *Science*, 1998. **282**(5389): p. 699-705.

72. Ivell, R., A question of faith--or the philosophy of RNA controls. *J Endocrinol*, 1998. **159**(2): p. 197-200.

73. Lander, E.S., Array of hope. *Nat Genet*, 1999. **21**(1 Suppl): p. 3-4.

74. Young, R.A., Biomedical discovery with DNA arrays. *Cell*, 2000. **102**(1): p. 9-15.

75. Khan, J., et al., Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res*, 1998. **58**(22): p. 5009-5013.

76. Beger, C., et al., Identification of Id4 as a regulator of BRCA1 expression by using a ribozyme-library-based inverse genomics approach. *Proc Natl Acad Sci U S A*, 2001. **98**(1): p. 130-135.

77. Butte, A.J. and I.S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 2000: p. 418-29.

78. Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999. **286**(5439): p. 531-7.

79. Warrington, J.A., et al., Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics*, 2000. **2**(3): p. 143-147.

80. Novak, J.P., R. Sladek, and T.J. Hudson, Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, 2002. **79**(1): p. 104-13.

81. Oliveira, J.G., et al., The housekeeping gene glyceraldehyde-3-phosphate dehydrogenase is inappropriate as internal control in comparative studies between skin tissue and cultured skin fibroblasts using Northern blot analysis. *Arch Dermatol Res*, 1999. **291**(12): p. 659-61.

82. Suzuki, T., P.J. Higgins, and D.R. Crawford, Control selection for RNA quantitation. *Biotechniques*, 2000. **29**(2): p. 332-337.

83. Souaze, F., et al., Quantitative RT-PCR: limits and accuracy. *Biotechniques*, 1996. **21**(2): p. 280-285.

84. Wu, Y.Y. and J.L. Rees, Variation in epidermal housekeeping gene expression in different pathological states. *Acta Derm Venereol*, 2000. **80**(1): p. 2-3.

85. Serazin-Leroy, V., et al., Semi-quantitative RT-PCR for comparison of mRNAs in cells with different amounts of housekeeping gene transcripts. *Mol Cell Probes*, 1998. **12**(5): p. 283-291.

86. Thellin, O., et al., Housekeeping genes as internal standards: use and limits. *J Biotechnol*, 1999. **75**(2-3): p. 291-295.

87. Savonet, V., et al., Pitfalls in the use of several "housekeeping" genes as standards for quantitation of mRNA: the example of thyroid cells. *Anal Biochem*, 1997. **247**(1): p. 165-167.

88. Staunton, J.E., et al., Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, 2001. **98**(19): p. 10787-92.

89. Gurdon, J.B., *The control of gene expression in animal development.* 1974, Cambridge, Mass.,: Harvard University Press. 160 p. illus.

90. Watson, P.M., et al., Differential regulation of leptin expression and function in A/J vs. C57BL/6J mice during diet-induced obesity. *Am J Physiol Endocrinol Metab*, 2000. **279**(2): p. E356-65.

91. Purcell, M.K., et al., Fine mapping of Ath6, a quantitative trait locus for atherosclerosis in mice. *Mamm Genome*, 2001. **12**(7): p. 495-500.

92. Mu, J.L., et al., Quantitative trait loci analysis for the differences in susceptibility to atherosclerosis and diabetes between inbred mouse strains C57BL/6J and C57BLKS/J. *J Lipid Res*, 1999. **40**(7): p. 1328-1335.

93. Beckers, M.C., et al., Natural resistance to infection with Legionella pneumophila: chromosomal localization of the Lgn1 susceptibility gene. *Mamm Genome*, 1995. **6**(8): p. 540-5.

94. Forbes, C.A., et al., The Cmv1 host resistance locus is closely linked to the Ly49 multigene family within the natural killer cell gene complex on mouse chromosome 6. *Genomics*, 1997. **41**(3): p. 406-13.

95. Min-Oo, G., et al., Pyruvate kinase deficiency in mice protects against malaria. *Nat Genet*, 2003. **35**(4): p. 357-62.

96. Stevenson, M.M., et al., Macrophage activation during Plasmodium chabaudi AS infection in resistant C57BL/6 and susceptible A/J mice. *Infect Immun*, 1992. **60**(3): p. 1193-1201.

97. Manenti, G., et al., Genetic mapping of cancer susceptibility/resistance loci in the mouse. *Recent Results Cancer Res*, 1998. **154**: p. 292-7.

98. Malkinson, A.M., M.N. Nesbitt, and E. Skamene, Susceptibility to urethan-induced pulmonary adenomas between A/J and C57BL/6J mice: use of AXB and BXA recombinant inbred lines indicating a three-locus genetic model. *J Natl Cancer Inst*, 1985. **75**(5): p. 971-974.

99. Ding, Y. and D. Wilkins, The effect of normalization on microarray data analysis. *DNA Cell Biol*, 2004. **23**(10): p. 635-42.

100. Wade, C.M., et al., The mosaic structure of variation in the laboratory mouse genome. *Nature*, 2002. **420**(6915): p. 574-8.

101. Wayne, M.L. and L.M. McIntyre, Combining mapping and arraying: An approach to candidate gene identification. *Proc Natl Acad Sci U S A*, 2002. **99**(23): p. 14903-6.

102. Walker, J.R., et al., Applications of a rat multiple tissue gene expression data set. *Genome Res*, 2004. **14**(4): p. 742-9.

103. Wittenburg, H., et al., Interacting QTLs for cholesterol gallstones and gallbladder mucin in AKR and SWR strains of mice. *Physiol Genomics*, 2002. **8**(1): p. 67-77.

104. Monks, S.A., et al., Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, 2004. **75**(6): p. 1094-105.

105. Sampson, S.B., et al., An edited linkage map for the AXB and BXA recombinant inbred mouse strains. *Mamm Genome*, 1998. **9**(9): p. 688-94.

106. Lee, P.D., et al., Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res*, 2002. **12**(2): p. 292-7.

107. Pritchard, C.C., et al., Project normal: defining normal variance in mouse gene expression. *Proc Natl Acad Sci U S A*, 2001. **98**(23): p. 13266-71.

108. Novak, J.P., R. Sladek, and T.J. Hudson, Characterization of variability in large-scale gene expression data: implications for study design. 2001.

109. Irizarry, R.A., et al., Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 2003. **31**(4): p. e15.

110. Li, C. and W.H. Wong, Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 2001. **98**(1): p. 31-6.

111. Kent, W.J., et al., The human genome browser at UCSC. *Genome Res*, 2002. **12**(6): p. 996-1006.

112. Giles, P.J. and D. Kipling, Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 2003. **19**(17): p. 2254-62.

113. Storey, J.D. and R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 2003. **100**(16): p. 9440-5.

114. Kim, B.S., et al., Spearman's footrule as a measure of cDNA microarray reproducibility. *Genomics*, 2004. **84**(2): p. 441-8.

115. Champy, M.F., et al., Mouse functional genomics requires standardization of mouse handling and housing conditions. *Mamm Genome*, 2004. **15**(10): p. 768-783.

116. Storch, K.F., et al., Extensive and divergent circadian gene expression in liver and heart. *Nature*, 2002. **417**(6884): p. 78-83.

117. Jin, W., et al., The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat Genet*, 2001. **29**(4): p. 389-395.
118. Fortin, A., et al., Identification of a new malaria susceptibility locus (Char4) in recombinant congenic strains of mice. *Proc Natl Acad Sci U S A*, 2001. **98**(19): p. 10793-8.
119. Bauer, P.R., et al., Regulation of endothelial cell adhesion molecule expression in an experimental model of cerebral malaria. *Microcirculation*, 2002. **9**(6): p. 463-70.
120. Prows, D.R., et al., Genetic analysis of ozone-induced acute lung injury in sensitive and resistant strains of mice. *Nat Genet*, 1997. **17**(4): p. 471-4.
121. Rozenberg, O., et al., Paraoxonase (PON1) deficiency is associated with increased macrophage oxidative stress: studies in PON1-knockout mice. *Free Radic Biol Med*, 2003. **34**(6): p. 774-84.
122. Hegele, R.A., Paraoxonase genes and disease. *Ann Med*, 1999. **31**(3): p. 217-24.
123. Devereux, T.R., et al., Assignment of a locus for mouse lung tumor susceptibility to proximal chromosome 19. *Mamm Genome*, 1994. **5**(12): p. 749-55.
124. Manenti, G., et al., Haplotype sharing suggests that a genomic segment containing six genes accounts for the pulmonary adenoma susceptibility 1 (Pas1) locus activity in mice. *Oncogene*, 2004. **23**(25): p. 4495-504.
125. Houtman, R., et al., Lung proteome alterations in a mouse model for nonallergic asthma. *Proteomics*, 2003. **3**(10): p. 2008-18.
126. Katsunuma, T., et al., Analysis of gene expressions of T cells from children with acute exacerbations of asthma. *Int Arch Allergy Immunol*, 2004. **134**(1): p. 29-33.
127. Al-Rabia, M.W., et al., Membrane receptor-mediated apoptosis and caspase activation in the differentiated EoL-1 eosinophilic cell line. *J Leukoc Biol*, 2004. **75**(6): p. 1045-55.
128. Laprise, C., et al., Functional classes of bronchial mucosa genes that are differentially expressed in asthma. *BMC Genomics*, 2004. **5**(1): p. 21.
129. Iizawa, Y., R.D. Wagner, and C.J. Czuprynski, Analysis of cytokine mRNA expression in Listeria-resistant C57BL/6 and Listeria-susceptible A/J mice during Listeria monocytogenes infection. *Infect Immun*, 1993. **61**(9): p. 3739-44.

130. Lemon, W.J., et al., Identification of candidate lung cancer susceptibility genes in mouse using oligonucleotide arrays. *J Med Genet*, 2002. **39**(9): p. 644-55.
131. Schadt, E.E., et al., Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 2003. **422**(6929): p. 297-302.
132. Steinmetz, L.M., et al., Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 2002. **416**(6878): p. 326-30.
133. Callow, M.J., et al., Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res*, 2000. **10**(12): p. 2022-9.
134. Gu, W., et al., Gene expression between a congenic strain that contains a quantitative trait locus of high bone density from CAST/EiJ and its wild-type strain C57BL/6J. *Funct Integr Genomics*, 2002. **1**(6): p. 375-86.
135. Cowles, C.R., et al., Detection of regulatory variation in mouse genes. *Nat Genet*, 2002. **32**(3): p. 432-7.
136. Sandberg, R., et al., Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc Natl Acad Sci U S A*, 2000. **97**(20): p. 11038-43.
137. Aronow, B.J., et al., Divergent transcriptional responses to independent genetic causes of cardiac hypertrophy. *Physiol Genomics*, 2001. **6**(1): p. 19-28.
138. Actor, J.K., et al., Relationship of survival, organism containment, and granuloma formation in acute murine tuberculosis. *J Interferon Cytokine Res*, 1999. **19**(10): p. 1183-93.
139. Boyle, A.E. and K. Gill, Sensitivity of AXB/BXA recombinant inbred lines of mice to the locomotor activating effects of cocaine: a quantitative trait loci analysis. *Pharmacogenetics*, 2001. **11**(3): p. 255-64.
140. De Sanctis, G.T., et al., Quantitative locus analysis of airway hyperresponsiveness in A/J and C57BL/6J mice. *Nat Genet*, 1995. **11**(2): p. 150-4.
141. Dietrich, W.F., et al., Lgn1, a gene that determines susceptibility to Legionella pneumophila, maps to mouse chromosome 13. *Genomics*, 1995. **26**(3): p. 443-50.
142. Dwyer-Nield, L.D., et al., Quantitative trait locus mapping of genes regulating pulmonary PKC activity and PKC-alpha content. *Am J Physiol Lung Cell Mol Physiol*, 2000. **279**(2): p. L326-32.

143. Festing, M.F., A. Yang, and A.M. Malkinson, At least four genes and sex are associated with susceptibility to urethane-induced pulmonary adenomas in mice. *Genet Res*, 1994. **64**(2): p. 99-106.

144. Gershenfeld, H.K. and S.M. Paul, Mapping quantitative trait loci for fear-like behaviors in mice. *Genomics*, 1997. **46**(1): p. 1-8.

145. Gill, K., et al., Alcohol-induced locomotor activation in C57BL/6J, A/J, and AXB/BXA recombinant inbred mice: strain distribution patterns and quantitative trait loci analysis. *Psychopharmacology (Berl)*, 2000. **150**(4): p. 412-21.

146. Gill, K., et al., Alcohol preference in AXB/BXA recombinant inbred mice: gender differences and gender-specific quantitative trait loci. *Mamm Genome*, 1998. **9**(12): p. 929-35.

147. Ihrig, M., M.D. Schrenzel, and J.G. Fox, Differential susceptibility to hepatic inflammation and proliferation in AXB recombinant inbred mice chronically infected with Helicobacter hepaticus. *Am J Pathol*, 1999. **155**(2): p. 571-82.

148. Johnson, K.R., Q.Y. Zheng, and L.C. Erway, A major gene affecting age-related hearing loss is common to at least ten inbred strains of mice. *Genomics*, 2000. **70**(2): p. 171-80.

149. Koza, R.A., et al., Synergistic gene interactions control the induction of the mitochondrial uncoupling protein (Ucp1) gene in white fat tissue. *J Biol Chem*, 2000. **275**(44): p. 34486-92.

150. Lu, X., E. Skamene, and P.M. Richardson, Studies of axonal regeneration in C57BL/6J and A/J mice. *Brain Res*, 1994. **652**(1): p. 174-6.

151. Manenti, G., et al., Pas1 is a common lung cancer susceptibility locus in three mouse strains. *Mamm Genome*, 1997. **8**(11): p. 801-4.

152. Matesic, L.E., A. De Maio, and R.H. Reeves, Mapping lipopolysaccharide response loci in mice using recombinant inbred and congenic strains. *Genomics*, 1999. **62**(1): p. 34-41.

153. Matesic, L.E., et al., Quantitative trait loci modulate neutrophil infiltration in the liver during LPS-induced inflammation. *FASEB J*, 2000. **14**(14): p. 2247-54.

154. Mathis, C., S.M. Paul, and J.N. Crawley, Characterization of benzodiazepine-sensitive behaviors in the A/J and C57BL/6J inbred strains of mice. *Behav Genet*, 1994. **24**(2): p. 171-80.

155. Mogil, J.S., C.A. Lichtensteiger, and S.G. Wilson, The effect of genotype on sensitivity to inflammatory nociception: characterization of resistant (A/J) and sensitive (C57BL/6J) inbred mouse strains. *Pain*, 1998. **76**(1-2): p. 115-25.

156. O'Malley, J., et al., Comparison of acute endotoxin-induced lesions in A/J and C57BL/6J mice. *J Hered*, 1998. **89**(6): p. 525-30.

157. Prows, D.R., et al., Ozone-induced acute lung injury: genetic analysis of F(2) mice generated from A/J and C57BL/6J strains. *Am J Physiol*, 1999. **277**(2 Pt 1): p. L372-80.

158. Prows, D.R. and G.D. Leikauf, Quantitative trait analysis of nickel-induced acute lung injury in mice. *Am J Respir Cell Mol Biol*, 2001. **24**(6): p. 740-6.

159. Spearow, J.L., et al., Mapping genes that control hormone-induced ovulation rate in mice. *Biol Reprod*, 1999. **61**(4): p. 857-72.

160. Sugiyama, F., et al., Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics*, 2001. **71**(1): p. 70-7.

161. Surwit, R.S., et al., Differential effects of fat and sucrose on the development of obesity and diabetes in C57BL/6J and A/J mice. *Metabolism*, 1995. **44**(5): p. 645-51.

162. Wu-Hsieh, B., Relative susceptibilities of inbred mouse strains C57BL/6 and A/J to infection with Histoplasma capsulatum. *Infect Immun*, 1989. **57**(12): p. 3788-92.

163. De Sanctis, G.T., et al., Quantitative trait locus mapping of airway responsiveness to chromosomes 6 and 7 in inbred mice. *Am J Physiol*, 1999. **277**(6 Pt 1): p. L1118-23.

164. Prows, D.R., et al., Genetic susceptibility to nickel-induced acute lung injury. *Chemosphere*, 2003. **51**(10): p. 1139-48.

165. Manenti, G. and T.A. Dragani, Pas1 haplotype-dependent genetic predisposition to lung tumorigenesis in rodents: a meta-analysis. *Carcinogenesis*, 2004.

166. Khaitovich, P., et al., Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, 2004. **14**(8): p. 1462-73.
167. Rockman, M.V. and G.A. Wray, Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol*, 2002. **19**(11): p. 1991-2004.
168. Wittkopp, P.J., B.K. Haerum, and A.G. Clark, Evolutionary changes in cis and trans gene regulation. *Nature*, 2004. **430**(6995): p. 85-8.
169. Brem, R.B., et al., Genetic dissection of transcriptional regulation in budding yeast. *Science*, 2002. **296**(5568): p. 752-5.
170. Pastinen, T. and T.J. Hudson, Cis-acting regulatory variation in the human genome. *Science*, 2004. **306**(5696): p. 647-50.
171. Buchner, D.A., M. Trudeau, and M.H. Meisler, SCNM1, a putative RNA splicing factor that modifies disease severity in mice. *Science*, 2003. **301**(5635): p. 967-9.
172. Pastinen, T., et al., A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics*, 2004. **16**(2): p. 184-93.
173. Demant, P. and A.A. Hart, Recombinant congenic strains--a new tool for analyzing genetic traits determined by more than one gene. *Immunogenetics*, 1986. **24**(6): p. 416-22.
174. Moen, C.J., et al., The recombinant congenic strains--a novel genetic tool applied to the study of colon tumor development in the mouse. *Mamm Genome*, 1991. **1**(4): p. 217-27.
175. Stassen, A.P., et al., Genetic composition of the recombinant congenic strains. *Mamm Genome*, 1996. **7**(1): p. 55-8.
176. van Zutphen, L.F., et al., Segregation of genes from donor strain during the production of recombinant congenic strains. *Lab Anim*, 1991. **25**(3): p. 193-7.
177. Tripodis, N., et al., Complexity of lung cancer modifiers: mapping of thirty genes and twenty-five interactions in half of the mouse genome. *J Natl Cancer Inst*, 2001. **93**(19): p. 1484-91.
178. Kent, W.J., BLAT--the BLAST-like alignment tool. *Genome Res*, 2002. **12**(4): p. 656-64.
179. Cheung, V.G. and R.S. Spielman, The genetics of variation in gene expression. *Nat Genet*, 2002. **32 Suppl**: p. 522-5.

180. Knight, J.C., Regulatory polymorphisms underlying complex disease traits. *J Mol Med*, 2004.

181. Diez, E., et al., The neuronal apoptosis inhibitory protein (Naip) is expressed in macrophages and is modulated after phagocytosis and during intracellular infection with Legionella pneumophila. *J Immunol*, 2000. **164**(3): p. 1470-7.

182. Knight, J.C., Allele-specific gene expression uncovered. *Trends Genet*, 2004. **20**(3): p. 113-6.

183. Lee, P.D., et al., Tissue-specific differences in basal gene expression between A/J and C57BL/6J inbred mouse strains. *Physiol Genomics*, 2005. **Submitted**.

184. Sladek, R., *Personal communication*, 2005.

185. Thanaraj, T.A., et al., ASD: the Alternative Splicing Database. *Nucleic Acids Res*, 2004. **32 Database issue**: p. D64-9.

186. Karolchik, D., et al., The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 2004. **32 Database issue**: p. D493-6.

187. Rosen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.*, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology.*, S.A. Krawetz and S. Misener, Editors. 2000, Humana Press: Totowa, NJ. p. 365-386.

188. Bray, N.J., et al., Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Hum Mol Genet*, 2004. **13**(22): p. 2885-92.

189. Ge, B., *Personal communication*, 2005.

190. Lander, E.S. and D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 1989. **121**(1): p. 185-99.

191. Lynch, M. and B. Walsh, *Genetics and analysis of quantitative traits*. 1998, Sunderland, Ma.: Sinauer. xvi, 980 p.

192. Lander, E.S. and D. Botstein, Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb Symp Quant Biol*, 1986. **51 Pt 1**: p. 49-62.

193. Wilkins, J.F. and D. Haig, What good is genomic imprinting: the function of parent-specific gene expression. *Nat Rev Genet*, 2003. **4**(5): p. 359-68.

194. Morison, I.M., C.J. Paton, and S.D. Cleverley, The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res*, 2001. **29**(1): p. 275-6.
195. Pastinen, T., *Personal communication*, 2005.
196. Naef, F., et al., DNA hybridization to mismatched templates: a chip study. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2002. **65**(4 Pt 1): p. 040902.
197. Irizarry, R.A., et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003. **4**(2): p. 249-64.
198. Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003. **19**(2): p. 185-93.
199. Whitney, A.R., et al., Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A*, 2003. **100**(4): p. 1896-901.
200. Risch, N.J., Searching for genetic determinants in the new millennium. *Nature*, 2000. **405**(6788): p. 847-56.
201. Carlborg, O. and L. Andersson, Use of randomization testing to detect multiple epistatic QTLs. *Genet Res*, 2002. **79**(2): p. 175-84.
202. Kauffman, S. and A.D. Ellington, Thinking combinatorially. *Curr Opin Chem Biol*, 1999. **3**(3): p. 256-9.
203. Davidson, E.H., D.R. McClay, and L. Hood, Regulatory gene networks and the properties of the developmental process. *Proc Natl Acad Sci U S A*, 2003. **100**(4): p. 1475-80.
204. Anholt, R.R., et al., The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome. *Nat Genet*, 2003. **35**(2): p. 180-4.
205. Geley, S., et al., Genes mediating glucocorticoid effects and mechanisms of their regulation. *Rev Physiol Biochem Pharmacol*, 1996. **128**: p. 1-97.
206. Socolar, J.E. and S.A. Kauffman, Scaling in ordered and critical random boolean networks. *Phys Rev Lett*, 2003. **90**(6): p. 068702.
207. Shen-Orr, S.S., et al., Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 2002. **31**(1): p. 64-8.

208. Yeger-Lotem, E., et al., Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 2004. **101**(16): p. 5934-9.

209. Breiman, L., *Classification and regression trees.* Wadsworth statistics/probability series. 1984, Belmont, Calif.: Wadsworth International Group. x, 358 p.

210. Pavlov, Y.L., *Random forests.* 2000, Utrecht: VSP. 122 p.

211. Ritchie, M.D., L.W. Hahn, and J.H. Moore, Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol*, 2003. **24**(2): p. 150-7.

212. Bystrykh, L., et al., Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet*, 2005. **37**(3): p. 225-32.

213. Jansen, R.C. and J.P. Nap, Genetical genomics: the added value from segregation. *Trends Genet*, 2001. **17**(7): p. 388-91.

214. Glass, L. and S.A. Kauffman, The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*, 1973. **39**(1): p. 103-29.

215. Glass, L. and S.A. Kauffman, Co-operative components, spatial localization and oscillatory cellular dynamics. *J Theor Biol*, 1972. **34**(2): p. 219-37.

216. Yvert, G., et al., Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 2003. **35**(1): p. 57-64.

217. Lahav, G., et al., Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nat Genet*, 2004. **36**(2): p. 147-50.

218. Gardner, T.S., C.R. Cantor, and J.J. Collins, Construction of a genetic toggle switch in Escherichia coli. *Nature*, 2000. **403**(6767): p. 339-42.

219. Gardner, T.S., et al., Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 2003. **301**(5629): p. 102-5.

220. Hasty, J., D. McMillen, and J.J. Collins, Engineered gene circuits. *Nature*, 2002. **420**(6912): p. 224-30.

221. Kaern, M., W.J. Blake, and J.J. Collins, The engineering of gene regulatory networks. *Annu Rev Biomed Eng*, 2003. **5**: p. 179-206.

222. Kauffman, S.A., *At home in the universe : the search for laws of self-organization and complexity*. 1995, New York: Oxford University Press. viii, 321 p.

223. Lander, E.S., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. **409**(6822): p. 860-921.

224. Adams, M.D., et al., The genome sequence of Drosophila melanogaster. *Science*, 2000. **287**(5461): p. 2185-95.

225. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 1998. **282**(5396): p. 2012-8.

226. Waterston, R.H., et al., Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002. **420**(6915): p. 520-62.

APPENDIX A. REVIEW ARTICLE - LA PUCE À ADN EN MÉDECINE ET EN SCIENCE.

DNA microarrays in medicine and science

PD Lee and TJ Hudson

Originally published in Médecine Sciences, January 16, 2000, 16:43-49.

m/s2000

médecine/sciences 2000 ; 16 : 43-9

La puce à ADN en médecine et en science

Peter Lee
Thomas J. Hudson

P. Lee, T.J. Hudson : Centre génomique de Montréal, Université McGill, 1650, Cedar avenue, Montréal, Québec, H3G 1A4 Canada.

▶ La révolution génomique offre de nouveaux outils pour l'étude de processus biologiques complexes à l'échelle pangénomique. Cet article passe en revue les différents principes et applications de l'une de ces technologies émergentes, celle des puces à ADN. Il existe actuellement deux procédés majeurs de fabrication de puces à ADN : (1) le dépôt direct de molécules d'ADNc ; (2) la synthèse *in situ* d'oligonucléotides sur une surface solide. Ces deux types de procédés, bien que présentant des différences qualitatives, offrent tous deux la possibilité d'un grand nombre d'applications nouvelles, à la fois fondamentales et cliniques, en permettant d'étudier simultanément plusieurs milliers de gènes et de découvrir rapidement de nombreux polymorphismes fonctionnels au niveau génomique. ◀

Le projet de séquençage du génome humain est en plein développement, et une première version préliminaire sera probablement disponible dès le printemps de cette année (*voir* l'article de Jean Weissenbach et Marcel Salanoubat, *p. 10 de ce numéro*). L'avènement de ce programme de séquençage, et des projets similaires concernant d'autres organismes, révolutionne la recherche en biomédecine, tant par l'élaboration de nouvelles technologies d'analyses de l'ADN que par la création d'immenses banques de données informatiques. Ces ressources extraordinaires amènent la communauté scientifique à se poser de nouvelles questions et permettront vraisemblablement d'élucider des mécanismes moléculaires complexes.

Traditionnellement, les biologistes ont utilisé des approches réductionnistes afin de disséquer un problème. La portée de chacune des questions posées est ainsi bien souvent limitée, tous les efforts étant tournés vers la résolution de points très spécifiques. Nous analysons ainsi de façon minutieuse, à l'échelle moléculaire, chaque composante d'un processus biologique de base. Cependant, les techniques de biologie moléculaire classique trouvent leurs limites lorsqu'elles sont appliquées à l'élucidation de processus complexes. Les fruits des divers projets de séquençage de génomes offrent de nouveaux outils afin d'étudier ces processus à l'échelle du génome. Dans cet article, nous examinerons l'un de ces outils, dont l'utilisation est de plus en plus répandue, la puce à ADN ou *microarray*.

Le principe de l'hybridation moléculaire permettant de détecter la présence d'acides nucléiques est maintenant bien établi. Les *Southern blots* et les *Northern blots* font depuis longtemps partie des techniques de base de tous les laboratoires de recherche en biologie. Le jumelage de plusieurs technologies a permis la miniaturisation de ces techniques d'hybridation, permettant ainsi de déceler des milliers de molécules d'acide nucléique de façon simultanée sur des matrices solides mesurant quelques centimètres carrés. Deux procédés majeurs de fabrication de puces à ADN sont couramment utilisés : (1) le dépôt direct d'ADNc sur lamelle de verre activée ; ou (2) la synthèse *in situ* d'oligonucléotides par photolithographie *(figure 1)*.

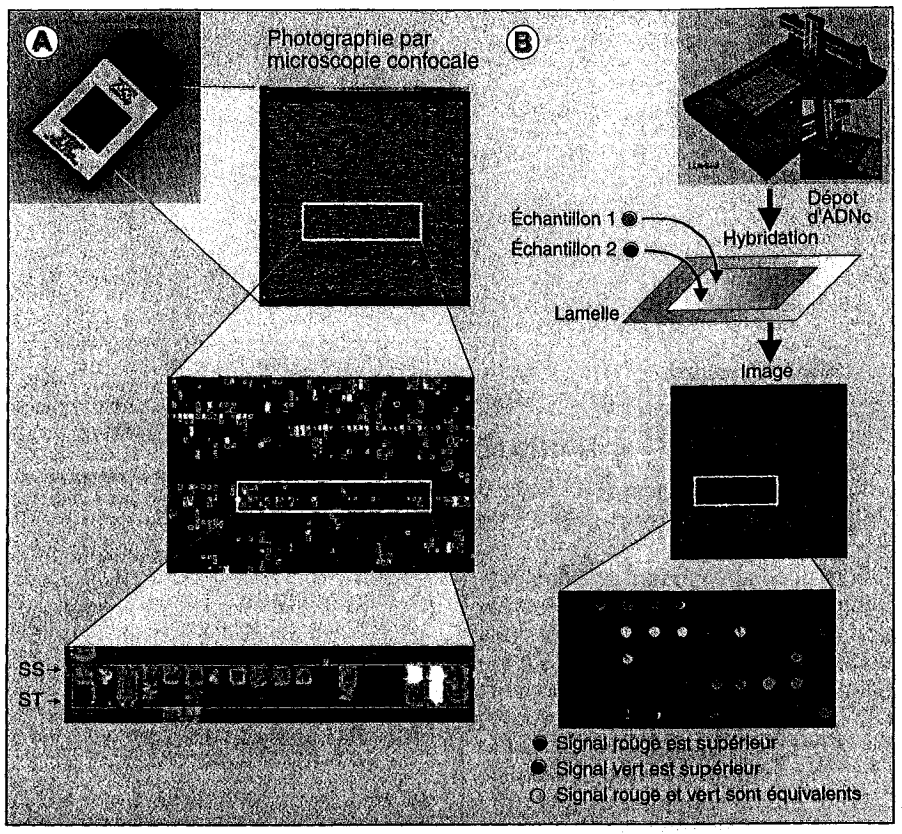


Figure 1. ***Puces à ADN. A.*** *Puce à ADN produite par Affymetrix. La puce à ADN est contenue dans une plaquette de plastique contenant une chambre d'hybridation. Après l'hybridation d'un échantillon d'ADN ou d'ARN marqué par un fluorophore, la surface contenant les 60 000 à 400 000 oligonucléotides est analysée par microscopie confocale et photographiée. Le grossissement de l'image permet de visualiser chaque espèce d'oligonucléotide retrouvé sur une surface de 20 µm x 20 µm. Dans l'exemple du bas, un gène est représenté par une série d'oligonucléotides de 25 nucléotides dérivés de la séquence du gène (SS: sonde spécifique). Chaque oligonucléotide possède son propre contrôle d'hybridation, obtenu par la synthèse d'un second oligonucléotide dont la séquence varie d'un seul nucléotide en position centrale (ST: sonde témoin). La concentration de l'ARN est mesurée par la moyenne des différences des oligonucléotides SS et SC.* ***B.*** *Puces à ADN préparées en parallèle à l'aide d'un micropipetteur robotisé qui dépose des ADNc sur la surface de la puce. Deux échantillons d'ARN provenant de différents tissus ou traitements sont marqués par des fluorophores différents (Cy-3 vert et Cy-5 rouge). La quantité relative de chaque gène est déterminée par le rapport d'émission de chaque fluorophore à des longueurs d'ondes différentes. La partie B est adaptée du site internet (<http://cmgm.Stanford.EDU/pbrown/>http://cmgm.Stanford.EDU/pbrown/) avec la permission de Joseph DeRisi.*

Dépôt de sondes sur puce à ADN

Le premier type de puce à ADN consiste en une lamelle de verre (identique à celle utilisée en microscopie traditionnelle) sur laquelle des milliers d'ADNc sont déposés à l'aide d'un micropipetteur robotisé. Grâce à cette technique, chacun des gènes (de fonction connue ou inconnue) est représenté par un seul point sur la lamelle. En général, deux échantillons d'ARN (sous forme d'ADNc obtenus par transcription inverse) sont co-hybridés sur la puce à ADNc. Les deux échantillons marqués par un fluorophore différent (Cy-3 vert ou Cy-5 rouge) s'hybrident simultanément avec les molécules complémentaires sur la puce. L'intensité du signal lumineux mesurée aux deux longueurs d'ondes correspondant aux différents fluorophores est alors mesurée à l'aide d'un microscope confocal. Le rapport de fluorescence rouge/vert est ainsi déterminé et permet de comparer les taux d'expression relatifs de chacun des gènes pour les deux échantillons d'ADNc. Un excès du gène *X* dans l'échantillon marqué en rouge produira un signal rouge ; un excès du gène *Y* dans l'échantillon marqué en vert produira un signal vert ; enfin, une expression équivalente du gène *Z* dans les deux échantillons produira un signal jaune. L'un des avantages de cette analyse comparée repose sur le fait que le rapport rouge/vert n'est pas influencé par la qualité de la goutte déposée par le pipetteur robotisé.

Synthèse d'oligonucléotides sur puce à ADN

Le second type de puce à ADN, proposé par la société Affymetrix, est constitué d'oligonucléotides synthétisés directement sur un substrat solide par photolithographie. Dans ce procédé, une lumière dirigée sur des sites spécifiques de la puce active la réaction d'oligo-synthèse [1, 2]. La synthèse d'un oligonucléotide de 25 paires de bases occupe un carré de 20 µm x 20 µm et contient plus de 10^7 copies de cette molécule. La surface d'une puce est d'environ 1,28 cm^2, et peut contenir 400 000 oligonucléotides différents ! Une puce à ADN destinée à des études d'expression contient pour chaque gène un ensemble d'oligonucléotides mimant la séquence du gène, souvent choisis dans sa région 3', réduisant ainsi les risques d'hybridations croisées avec des séquences homologues de ce gène. Des oligonucléotides, dont la séquence varie pour une seule base, sont également ajoutés, ce qui permet de confirmer que le signal obtenu pour chacun des gènes est bien spécifique. Contrairement à la puce à ADN décrite plus haut, celle produite par ce procédé permet l'hybridation d'un seul échantillon marqué à la fois. L'intensité de l'hybridation est également mesurée par microscopie confocale.

Comparaison des puces à ADN construites par ces deux procédés

Les puces à ADN produites par micropipetteurs offrent une grande

souplesse d'utilisation car il est facile pour le chercheur d'en modifier le contenu. En outre, elles sont relativement peu coûteuses, et, pour les amateurs, les modalités de construction d'un système robotisé sont disponibles sur Internet *(Tableau I)*. La préparation et l'optimisation de ces puces à ADN ne sont cependant pas simples, l'assemblage de milliers de gènes sur une puce nécessitant la validation et la purification de nombreux ADNc. Les puces à ADN produites par Affymetrix permettent d'étudier plus de 45 000 gènes humains, plus de 30 000 gènes murins et environ 6 000 gènes de levures. Elles ont cependant le désavantage d'être très coûteuses et leur contenu n'est pas modulable. Il est à souhaiter qu'avec le temps, elles deviennent plus fonctionnelles.
Ces deux technologies sont d'ores et déjà très prometteuses. D'autres types de puces, comme les biopuces électroniques (Nanogen), dans lesquelles des circuits électriques miniatures sont utilisés afin de diriger les tests moléculaires à la surface de la puce, sont maintenant en développement. Cette technologie permettra entre autres la séparation de cellules par affinité ainsi que le développement d'autres tests moléculaires.

Applications des puces à ADN

Analyses d'expression de gènes

Les premières puces ont servi à évaluer l'expression simultanée de milliers de gènes dans des systèmes biologiques bien connus, tels que celui du métabolisme respiratoire et de fermentation chez la levure [3], le cycle cellulaire de la levure *(figure 2)* [4, 5], la sporulation [6] et la stimulation de fibroblastes par le sérum [7]. Ces premiers travaux ont permis de valider la technologie. La comparaison des résultats obtenus par les puces à ADN avec ceux préalablement obtenus par d'autres approches a démontré une concordance pour les gènes dont l'expression était déjà connue dans ces systèmes biologiques. De plus, pour les milliers de nouveaux résultats obtenus, il existe une conformité « intuitive » avec des connaissances provenant d'autres processus cellulaires.

Analyses de voies biochimiques

Les puces à ADN permettent l'analyse de voies métaboliques spécifiques. Ainsi, Fambrough *et al.* [8] ont étudié la cascade de signalisation des récepteurs à activité tyrosine kinase dans les cellules NIH-3T3. Cette famille de récepteurs membranaires est conservée de la levure jusqu'à l'homme et joue un rôle fondamental dans la transmission de signaux de la membrane au noyau cellulaire. L'une des analyses utilisant des puces à ADN a été réalisée avec des cellules exprimant un récepteur muté du PDGF *(platelet derived growth factor)*, sur lequel les résidus tyrosine potentiellement spécifiques de certaines cascades de signalisation ont été modifiés. Or, cette analyse a révélé une redondance fonctionnelle considérable entre les différents mutants, qui n'avait pu être anticipée par les seuls modèles exploités anté-

Tableau I

RESSOURCES INFORMATIQUES SUR INTERNET

URL	Description	Référence
http://cmgm.stanford.edu/pbrown/explore/	Cycle métabolique de la levure	[3]
http://genome-www.stanford.edu/Saccharomyces/	Base de données sur le cycle cellulaire de la levure	[4]
http://genome-www.stanford.edu/cellcycle/	Analyse du cycle cellulaire de la levure	[5]
http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/	Projet de micropuces au NHGRI. Distribution de logiciels ArrayDB et ArrayViewer	[19]
http://web.wi.mit.edu/young/expression/	Données d'expression par génomique	[20]
http://cmgm.stanford.edu/pbrown/sporulation/index.html	Programme transcriptionnel de sporulation	[6]
http://genome-www.stanford.edu/serum/	Base de données d'expression sur la stimulation par le sérum	[7]
http://rana.stanford.edu/clustering/	Logiciel Cluster	[16]
http://waldo.wi.mit.edu/MPR/	Logiciel GeneCluster	[17]
http://www.affymetrix.com/technology/index.html	Aperçus des puces à ADN Affymetrix	
http://cmgm.stanford.edu/pbrown/mguide/index.html	Guide de fabrication de robot pour puces à ADN	

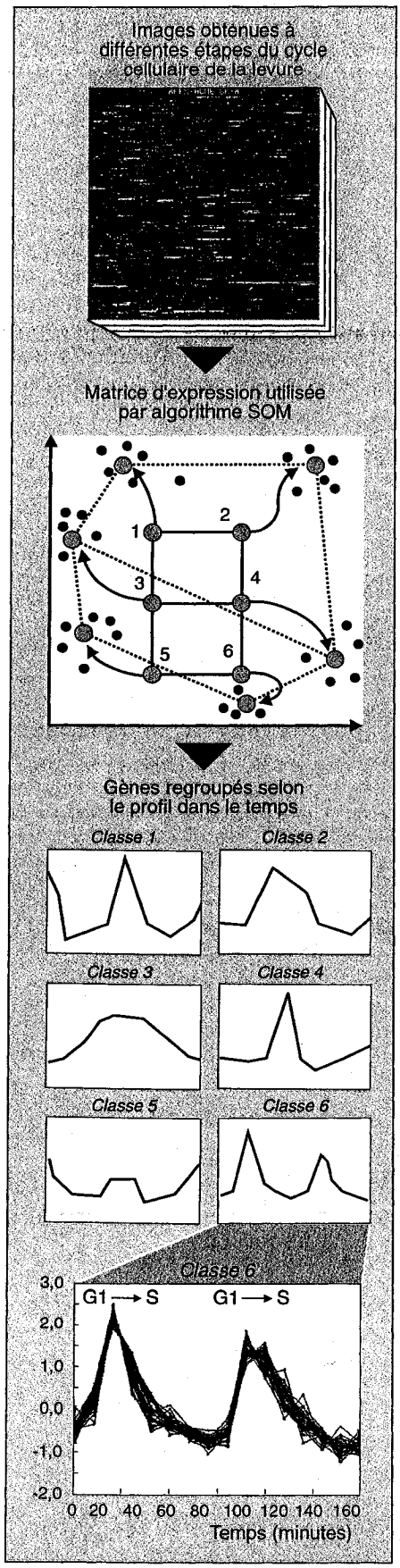


Figure 2. ***Analyse d'expression du cycle cellulaire de la levure.*** *Les données provenant de multiples mesures d'expression par puces à ADN à différentes étapes du cycle cellulaire de la levure engendrent un déluge d'informations. Un algorithme nommé* Self-Organizing Maps *(SOM), inventé par Kohonen [21], a été utilisé par Tamayo [17] pour l'étude des données génomiques. Cette méthode permet de regrouper les gènes ayant un profil d'expression semblable. Brièvement, les données sont d'abord localisées dans un espace (ou matrice d'expression) et sont ensuite groupées selon leur proximité dans cet espace, signifiant qu'elles ont des profils d'expression semblables. Dans l'exemple provenant du cycle cellulaire, plusieurs classes de gènes ont été décelées (de fonctions connues et inconnues), ayant des profils d'expression démontrant une périodicité semblable pendant le cycle cellulaire. L'exemple donné à la partie inférieure de la figure représente les gènes ayant une expression maximale durant la transition de la phase G1 à la phase S du cycle cellulaire de la levure. (Adapté de l'article de Tamayo* et al. *[17] avec la permission de* Proc Natl Acad Sci USA.*)*

rieurement. L'étude a aussi permis de démontrer que ces récepteurs peuvent remplir d'autres fonctions, telles que la stimulation de la production d'interleukine(s) pour l'un des récepteurs mutés. Il est possible que certaines de ces réponses soient réprimées dans le cas des récepteurs normaux, et l'utilisation de puces à ADN aurait alors permis de mettre en évidence l'existence de voies biochimiques alternatives.

Validation de mécanismes d'action de médicaments

Les puces à ADN peuvent également être utilisées pour étudier le mécanisme d'action d'un médicament. En principe, un médicament qui agit par inhibition spécifique d'un seul gène ou de son produit devrait engendrer un effet identique à celui résultant de l'inactivation de ce gène par délétion ou par mutation. Marton *et al.* [9] ont utilisé une puce à ADN contenant l'ensemble des gènes de la levure afin de démontrer l'existence d'une corrélation significative entre le profil obtenu lors d'une stimulation médicamenteuse antimicrobienne et le profil d'expression d'une levure portant un gène muté et impliqué dans le métabolisme d'action de ce médicament. Ce principe peut être exploité pour la création d'une base de données contenant un grand nombre de profils d'expression, provenant à la fois de cellules stimulées par des médicaments et de souches contenant différentes mutations. Ces données offrent un moyen de « décoder » les profils complexes d'expression de groupes de gènes modulés par différentes classes de médicaments.

Classification phénotypique et prédiction

En clinique, la distinction entre une leucémie myléoïde aiguë (LMA) et une leucémie lymphoïde aiguë (LLA) est d'une importance cruciale afin de bien orienter le traitement thérapeutique. Néanmoins, les techniques usuelles de cytopathologie sont complexes et requièrent une très grande expertise. Golub *et al.* [10] ont tenté de définir différentes classes de leucémies sur la base du profil d'expression de près de 6 000 gènes humains. L'analyse d'une soixantaine de leucémies présentant des cytopathologies connues a permis de distinguer plus d'une centaine de gènes ayant un profil différent dans les cas de LMA et de LLA. Bien qu'aucun de ces gènes (telle la myéloperoxidase) n'ait un mode d'expression identifique pour toutes les leucémies d'une même classe, l'analyse du groupe composé des 50 à 100 gènes les plus différents entre les deux classes de leucémies a une valeur prédictive significative (pratiquement de 100 % dans les cohortes de validation ultérieure, lorsque le taux de confiance est jugé satisfaisant). De plus, ces analyses ont permis d'identifier un troisième type de leucémies non identifiables par la seule approche cytopathologique. Les analyses par les puces à ADN ont donc une application immédiate en clinique, et présentent en outre la possibilité d'identifier de nouvelles classes de phénotypes pathologiques.

Analyses d'ADN génomique

L'étude des variations génomiques est d'une grande utilité en recherche bio-

médicale. Une grande partie de la variabilité interindividuelle observée au sein d'une même espèce, en particulier la susceptibilité à certaines maladies, est due à des différences (ou polymorphismes) existant au niveau de la séquence de l'ADN génomique. En raison de leur potentiel et de la rapidité d'analyse qu'elles présentent, les puces à ADN offrent un avantage considérable sur les techniques déjà existantes pour aborder ces variations.

Études génomiques comparées

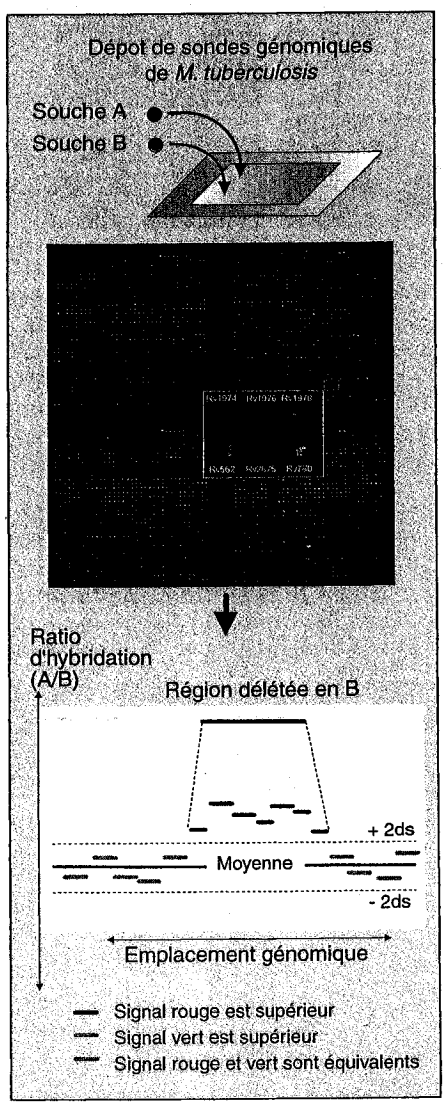
Behr *et al.* [11] ont étudié plusieurs souches du bacille de Calmette et Guérin (BCG) originaires de l'Institut Pasteur au cours du XX^{e} siècle *(figure 3)*. A partir de la séquence complète de *M. tuberculosis*, un assemblage de sondes génomiques couvrant l'ensemble de ce génome a été déposé sur une puce à ADN. Le génome de chaque souche de BCG étudiée a été amplifié, marqué par un fluorochrome et hybridé sur la puce à ADN. Cela a permis d'identifier des divergences au niveau génomique et de les corréler de façon temporelle avec l'historique de la dissémination mondiale de ces souches. Cette étude suggère l'existence d'un lien entre l'efficacité du vaccin contre le BCG et la prééminence des souches classifiées par ces marqueurs génomiques. A l'avenir, ce type d'analyse sera sûrement appliquée à d'autres pathogènes.

Reséquençage : détection de SNP

Grâce aux puces à ADN, une séquence connue peut être séquencée de nouveau dans l'intention de découvrir des polymorphismes affectant un seul nucléotide (SNP, *single nucleotide polymorphisms*). Cette technologie utilise le principe de séquençage par hybridation (SPH) [12]. Brièvement, une séquence connue peut être caractérisée par un assemblage d'oligonucléotides chevauchants. Pour cette application particulière, il faut donc créer une puce à ADN contenant tous les 25 mers chevauchants, définissant la séquence à interroger ainsi que trois amorces contenant les trois permutations possibles pour le nucléotide central de l'amorce (c'est-à-dire qu'un T sera remplacé par un A, un C et un G). Cette méthode a été utilisée par Wang *et al.* [13] pour reséquencer 2,3 Mb du génome chez 7 individus et a ainsi permis d'identifier 3 241 SNP humains. Cette approche a également été utilisée afin d'identifier des polymorphismes dans un gène de susceptibilité au cancer du sein *(BRCA1)* [14] ainsi que pour plus d'une centaine d'autres gènes impliqués dans des processus vasculaires, métaboliques et endocriniens [15]. Enfin, le reséquençage de génomes de pathogènes (tel celui du VIH) permet d'identifier des mutations fonctionnelles, ouvrant ainsi la voie vers des thérapies spécifiques et déterminées en fonction des polymorphismes génomiques du pathogène.

Génotypage

La détermination du génotype pour un SNP donné exploite la haute spé-



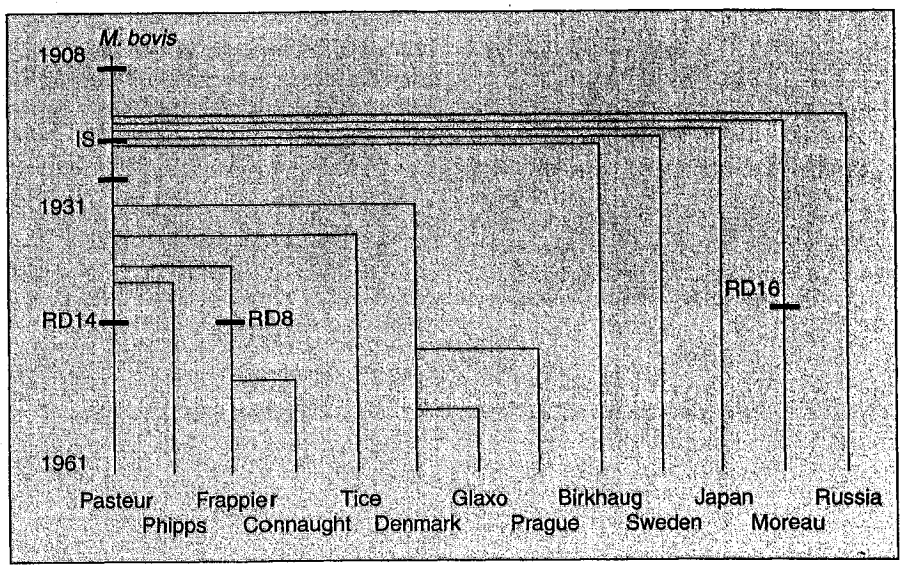


Figure 3. ***Analyse génomique de souches BCG.*** *Une puce à ADN contenant des sondes génomiques de* M. tuberculosis *a été utilisée pour interroger le génome de plusieurs souches de BCG [11]. Dans cet exemple, les comparaisons entre deux souches ont permis d'identifier de petites régions génomiques affichant un excès de couleur rouge, signifiant une délétion au sein du génome de la souche B marquée en vert. L'analyse complète de plusieurs souches a mis en évidence plusieurs microdélétions (contenant parfois un ou plusieurs gènes). (Adapté de l'article de Behr* et al. *[11] avec la permission de* Science.*)*

cificité de la puce à ADN à discriminer une complémentarité parfaite *versus* une complémentarité imparfaite causée par une seule paire de bases non appariées. La puce à ADN de génotypage *(figure 4)* contient donc des amorces spécifiques de chacun des allèles, outre des amorces témoins [13]. Une puce à ADN a la capacité de contenir des milliers de polymorphismes. Les applications possibles pour ces puces à ADN sont nombreuses : criblage génomique pour études de liaison génétique, études d'association avec des milliers de gènes candidats, caractérisation d'anomalies cytogénétiques, etc.

Informatique

La gestion du déluge d'informations produites par ces technologies représente un défi énorme pour les biologistes. L'un des facteurs limitants est donc la capacité d'analyse de l'information, pour laquelle de nouvelles méthodes sont nécessaires. Une nouvelle génération d'outils informatiques est actuellement en développement : ainsi, pour réaliser les études d'expression, il existe des algorithmes capables d'identifier des groupes de gènes partageant des profils d'expression semblables *(cluster* [16], *gene cluster* [17]*)*. D'autres méthodes sont destinées à trouver des relations entre des gènes ayant des expressions identiques, comme la recherche de motifs semblables au niveau des séquences localisées en 5' ou 3' du gène (par exemple, Yeast Toolset AlignACE pour la levure [18]). La création de ces nouvelles bases de données est indispensable afin de pouvoir regrouper, visualiser et partager les profils d'expression obtenus par les puces à ADN (ArrayViewer, ArrayDB [19]). Le *Tableau I* dresse la liste des ressources informatiques disponibles sur Internet et qui concernent les puces à ADN.

Ainsi, moins de cinquante ans après la découverte de la structure de l'ADN, nous entrons résolument dans l'ère post-génomique. Ces technologies fascinantes permettront d'explorer l'univers cellulaire dans toute sa complexité. Ayant gardé longtemps les apparences d'un roman de science-fiction, la révolution génomique est désormais une réalité captivante ■

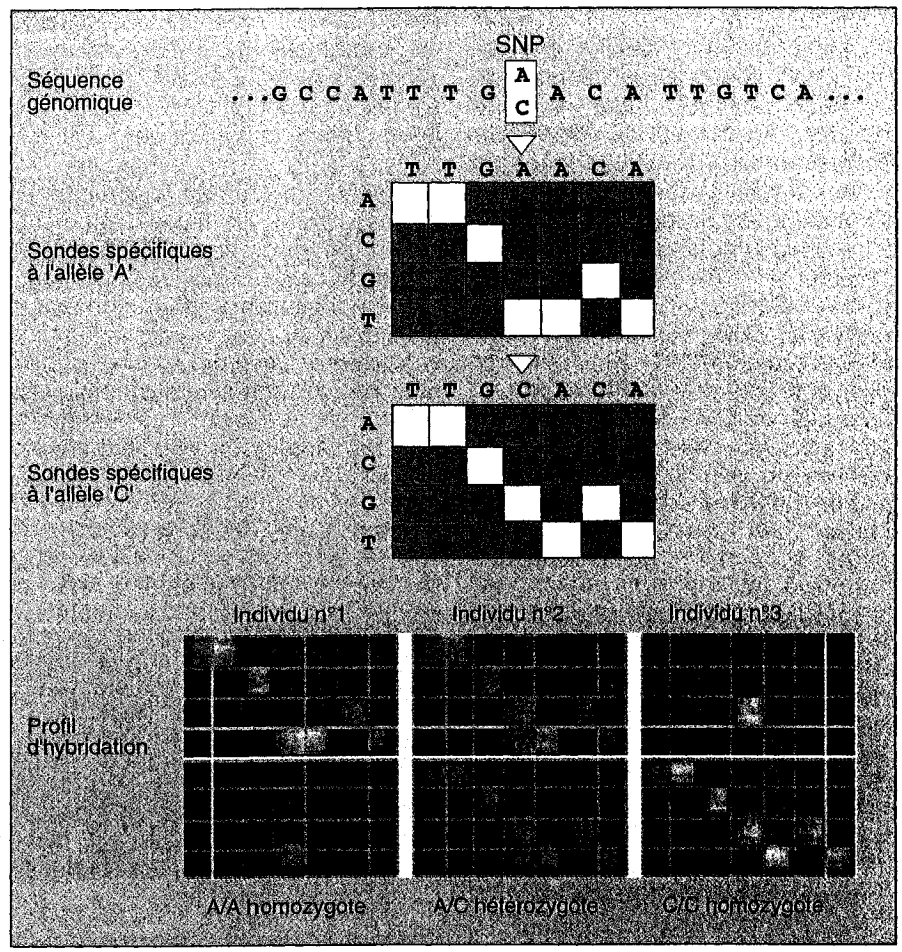


Figure 4. *Génotypage de SNP.* ***Pour réaliser le génotypage d'un SNP, tel que celui qui apparaît en haut de cette figure, deux séries d'oligonucléotides doivent être présents sur la puce, chaque série représentant un allèle différent. Chaque colonne contient des oligonucléotides successifs qui sont complémentaires de la séquence interrogée et dont la base centrale est substituée par un A, C, G, ou T dans les quatre rangées. Les sondes spécifiques aux allèles A et C apparaissent en blanc, tandis que les sondes témoins apparaissent en gris. Le génotype d'un individu peut être déterminé par l'analyse de la variation du signal d'hybridation provenant d'un produit PCR marqué contenant le SNP. Le signal d'hybridation de trois individus avec les génotypes AA, AC et CC est présenté à la partie inférieure de la figure. Une micropuce peut d'ailleurs interroger des milliers de SNP en parallèle. (Adapté de l'article de Wang*** et al. ***[13] avec la permission de*** Science.***)***

Remerciements

Les auteurs remercient Marcel Behr, Jean-Paul Comet, Marie-Claude Vohl et Claire Goguen pour leurs précieux commentaires.

RÉFÉRENCES

1. Chee M, Yang R, Hubbell E, *et al.* Accessing genetic information with high-density DNA arrays. *Science* 1996 ; 274 : 610-4.

2. Lockhart DJ, Dong H, Byrne MC, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996 ; 14 : 1675-80.

3. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997 ; 278 : 680-6.

4. Cho RJ, Campbell MJ, Winzeler EA, *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998 ; 2 : 65-73.

5. Spellman PT, Sherlock G, Zhang MQ, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998 ; 9 : 3273-97.

6. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. The transcriptional program of sporulation in budding yeast. *Science* 1998 ; 282 : 699-705.

RÉFÉRENCES

7. Iyer VR, Eisen MB, Ross DT, *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* 1999 ; 283 : 83-7.

8. Fambrough D, McClure K, Kazlauskas A, Lander ES. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* 1999 ; 97 : 727-41.

9. Marton MJ, DeRisi JL, Bennett HA, *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 1998 ; 4 : 1293-301.

10. Golub TR, Slonim D, Tamayo P, *et al.* Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science* 2000 (sous presse).

11. Behr MA, Wilson MA, Gill WP, *et al.* Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 1999 ; 284 : 1520-3.

12. Drmanac S, Kita D, Labat I, *et al.* Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat Biotechnol* 1998 ; 16 : 54-8.

13. Wang DG, Fan JB, Siao CJ, *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998 ; 280 : 1077-82.

14. Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996 ; 14 : 441-7.

15. Cargill M, Altshuler D, Ireland J, *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999 ; 22 : 231-8.

16. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998 ; 95 : 14863-8.

17. Tamayo P, Slonim D, Mesirov J, *et al.* Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999 ; 96 : 2907-12.

18. Van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998 ; 281 : 827-42.

19. Ermolaeva O, Rastogi M, Pruitt KD, *et al.* Data management and analysis for gene expression arrays. *Nat Genet* 1998 ; 20 : 19-23.

20. Holstege FC, Jennings EG, Wyrick JJ, *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 1998 ; 95 : 717-28.

21. Kohonen T. *Self-organizing maps.* Berlin : Springer Verlag, 1997.

TIRÉS À PART

T.J. Hudson.

m/s2000

Summary

DNA chips in medicine and science

The Human Genome Project is changing our conception of modern biology. Recent advances in technology are now enabling us to observe complex processes on a genome-wide scale. This review examines the emerging technology of DNA microarrays. Notwithstanding the differences related to manufacture characteristics and properties of the two major technologies used today, DNA microarrays offer the potential to simultaneously investigate thousands of genes. Expression DNA chips containing gene probes rely on the expression profile of collections of genes to investigate complex biochemical pathways, validate drug targets, and classify cell phenotype. Microarrays may be used to detect variations in DNA sequences and correlate these with phenotypes – as in genome scans for linkage studies, mutations detection, large-scale association studies, and analyses of drug responses. Numerous applications related to modern medicine in the areas of diagnostics and drug management are rapidly emerging.

CONFÉRENCES DE PHILOSOPHIE ET HISTOIRE DE LA MÉDECINE - Année universitaire 1999/2000

Ces conférences auront lieu les **mardis de 17 h à 19 h**

au Centre de documentation d'Histoire de la Médecine - 15, rue de l'École de Médecine - 75006 Paris - Pavillon 4

Renseignements : Professeur Robert Zittoun ✆ 01 42 34 69 48/01 42 34 69 65

Dates	Philosophie médicale 17 h 00-18 h 00	Histoire de la médecine 18 h 00-19 h 00
14 décembre 1999	La décision entre probabilités et incertitudes Robert Zittoun	A propos de quelques victoires médicales (1920-1950) François Dagognet
4 janvier 2000	Éthique : Les fondements de l'éthique médicale Daniel François-Wachter	Naissance de l'anesthésie Patrick Conan
11 janvier 2000	Éthique médicale contemporaine Daniel François-Wachter	Les révolutions biologiques Jean-Claude Ameisen
18 janvier 2000	Éthique de fin de vie Paula La Marne	Histoire de la gériatrie Alain Lellouch
25 janvier 2000	Brevets biologiques et éthique du clonage Henri Atlan	Médicalisation de la procréation Pierre Jouannet
1er février 2000	Engagement et responsabilité médicale Marie-Sylvie Richard	Jacques Monod Patrice Debré
22 février 2000	Ontologie : Être humain ou l'essence d'une vie Bernard-Marie Dupont	Histoire de l'autopsie Jacques Diebold
29 février 2000	Temps du SIDA Catherine Zittoun	Jonathan Mann, santé publique et droits de l'homme Emmanuel Hirsch
7 mars 2000	Psychologie : L'angoisse Jean Carpentier	La psychiatrie à Vienne autour de 1900 Christine Lévy
14 mars 2000	La relation soignant-soigné Martine Ruszniewski	Histoire de l'architecture hospitalière Jacques-Louis Binet
21 mars 2000	Psychopathologie de l'enfant malade Danièle Brun	Korczak ou comment aimer les enfants Stanislas Tomkiewicz
18 avril 2000	Anthropologie et sociologie : Anthropologie et sociologie médicale Marie-Frédérique Baqué	Anthropologie raciale et national-socialisme Benoît Massin
25 avril 2000	Le risque iatrogène François Bwald	Pratiques alimentaires et émergences de nouvelles maladies Olivier Robain
2 mai 2000	Économie de santé et gestion du soin : justification et enjeux Anne-Laurence Le Faou	L'émergence d'une culture de la santé publique Julien Faure

APPENDIX B. LABORATORY AND ANIMAL USE ETHICS APPROVALS

McGill University
Animal Use Protocol

ACTION	✓	DATE
P.I.	✓	
FACC	✓	
ACC	✓	
VET	✓	
INVEST		
Photo		

Project #	1654
Investigator #	547
Approval End Date	June 30, 2001
Facility Committee	MEH

☐ **New Application**

☒ **Renewal of Project #** 1654-[illegible]

1. Investigator Information

Principal Investigator: Emil Skamene — **Telephone:** 934-8038

Department: Medicine — **Fax:** 933-7146

Address: Room A6.149 Montreal General Hospital 1650 Cedar Avenue

E-mail: md68@musica.mcgill.ca

Animal Use: Research [x] Teaching [ ] Specify Course number:

Project Title: Complex Traits Analysis in Mice

2. Funding Source

External [x] **Internal** [ ]

Source(s): Network Centers of Excellence The Canadian Genetic Diseases Network (MGH 2594, 2671, 2706, 6084)

Peer Reviewed source: Yes [x] No [ ] *If no, append required Peer Review Forms

Awarded [x] Pending [ ]

Funding Period: From: 01/03/1997 **To:** 30/09/2005

Animal Use Period: Start: 01/03/1997 End: 30/09/2005

3. Emergency: Person(s) designated to handle emergencies (2 emergency telephone numbers must be indicated)

Name:	Emil Skamene	Phone #: Work:	934-8038	Alternative #:	948-2686
Name:	Danuta Radzioch	Phone #: Work:	937-8011 x 4517	Alternative #:	331-1284

Certification:

The information in this application is exact and complete. I agree to follow the policies and procedures set forth by the Facility Animal Care Committee and McGill University, as well as those described in the "Guide to the Care and Use of Experimental Animals" prepared by the Canadian Council on Animal Care. I shall request the Animal Care Committee's approval prior to any deviations from the procedures described within.

Principal Investigator/Course Director — Date: June 25, 2000

Approval:

Chairperson, Facility Animal Care Committee		Date	July 13, 2000
AFCC Veterinarian		Date	July 31, 2000
Chairperson, Ethics Subcommittee (Q level or Teaching Protocols Only)		Date	
Approved period for animal use	Beginning: July 1, 2000	Ending:	June 30, 2001

☒ **NOTE REVIEWER'S MODIFICATION(S) ON PAGE 2**

AUG 29 2000

C level

McGill University
Animal Use Protocol

FOR OFFICE USE ONLY

Project #	1654
Investigator #	547
Approval End Date	JUNE 30, 2002
Facility Committee	MCH

☐ New Application

☒ Renewal of Project # 1654

1. Investigator Information

Principal Investigator: Emil Skamene, MD, PhD — Telephone: 934-8038

Department: Medicine — Fax: 933-7146

Address: 1650 Cedar Avenue Montreal H3G 1A4 Room A6 149

E-mail: md88@musica.mcgill.ca

Animal Use: Research ☐ Teaching ☒ Specify Course number:

Project Title: Complex Traits Analysis in Mice - Core Facility

2. Funding Source

External ☒ Internal ☐ a/c 2671

Source(s): Networks of Centres of Excellence - Canadian Genetic Diseases Network

ACTION	✓	DATE
PI	✓	Aug 31 '01
FACC	✓	✓
RGO	✓	✓
VET	✓	✓
DB		
approved		

Peer Reviewed source: Yes ☒ No ☐ *If no, see instructions - section 2

Awarded ☒ Pending ☐

Funding Period: From: March 1 1997 To: June 30, 2005

Proposed Start Date of Research: 01/03/97
(Day/Month/Year)

Expected Date of Completion: 30/06/05
(Day/Month/Year)

3. Emergency: Person(s) designated to handle emergencies (2 emergency telephone numbers must be indicated)

Name: Emil Skamene — Phone #: Work: 934-8038 — Alternative #: 946-2686

Name: Francine Gervais — Phone #: Work: 937-6011 x 4511 — Alternative #: 334-6975

Certification:

The information in this application is exact and complete. I agree to follow the policies and procedures set forth by the Facility Animal Care Committees and McGill University, as well as those described in the "Guide to the Care and Use of Experimental Animals" prepared by the Canadian Council on Animal Care. I shall request the Animal Care Committee's approval prior to any deviations from the procedures described within.

Principal Investigator/Course Director — Date: July 27 2001

Approval:

Chairperson, Facility Animal Care Committee		Date	Aug 02/2001
University Animal Care Officer		Date	7/28/01
roved period for animal use	Beginning July 1, 2001	Ending	JUNE 30, 2002

☐ NOTE REVIEWER'S MODIFICATION(S) ON PAGE 2

Revised 02/97

FEB 02 '05 11:49 FR RVH RESEARCH INST 514 843 1687 TO 9-9850012 P.02/14

Renewed

Guidelines for completing the form are available at www.mcgill.ca/rgo/animal

McGill University
Animal Use Protocol – Research

For office use only
Protocol #: 1654
Investigator #: 547
Approval End Date: June 30, 2003
Facility Committee: M G H

Title: Complex Traits Analysis in Mice GENETIC DISSECTION OF COMPLEX TRAITS USING PHENOTYPIC AND EXPRESSION ANALYSIS OF RECOMBINA CONGENIC MOUSE STRAINS
(must match the title of the funding source application)

New Application: Renewal of Protocol: # 1654 Pilot: Category *(see section 11)*: C renewed

1. Investigator Data:

Principal Investigator: Emil Skamene Phone #: 934-8038
Department: Medicine Fax#: 933-7146
Address: Room A6.149 1650 Cedar Ave Montreal Email: Emil.skamene@muhc.mcgill.ca

2. Emergency Contacts: Two people must be designated to handle emergencies.

Name: Emil Skamene Work #: 934-8038 Emergency #: 946-2686
Name: Danuta Radzioch Work #: 937-6011 x 44517 Emergency #: 331-1284

3. Funding Source:

External: X $155
Source (s) : NCE + GENOME QUEBEC + GENOME CANADA
Peer Reviewed: YES X NO**
Status : Awarded X Pending
Funding period: 2002-2003

Internal:
Source (s) :
Peer Reviewed: YES NO**
Status: Awarded Pending
Funding period:

For Office Use Only:

ACTION	✓	DATE
CCS	✓	Sept 4/02
DB	✓	"
APPROVED		

** All projects that have not been peer reviewed for scientific merit by the funding source require 2 Peer Review Forms to be completed e.g. Projects funded from industrial sources. Peer Review Forms are available at www.mcgill.ca/rgo/animal

Proposed Start Date of Animal Use (d/m/y): or ongoing: X

Expected Date of Completion of Animal Use (d/m/y): or ongoing: X

Investigator's Statement: The information in this application is exact and complete. I assure that all care and use of animals in this proposal will be in accordance with the guidelines and policies of the Canadian Council on Animal Care and those of McGill University. I shall request the Animal Care Committee's approval prior to any deviations from this protocol as approved. I understand that this approval is valid for one year and must be approved on an annual basis.

Principal Investigator's signature: [signature] Date: June 17 2002

Approved by:

Chair, Facility Animal Care Committee:	[signature]	Date: July 4/02
University Veterinarian:	[signature]	Date: 8/28/02
Chair, Ethics Subcommittee (as per UACC policy):	—	Date: —
Approved Animal Use	Beginning: July 1, 2002	Ending: June 30, 2003

X This protocol has been approved with the modifications noted in Section 13.

May 2002

AUG 26 2002

398-1586

Guidelines for completing the form are available at: www.mcgill.ca/rgo/animal

For office use only
Protocol #: 1654
Investigator #: 542
Approval End Date: June 30, 2004
Facility Committee: MGH

McGill University
Animal Use Protocol – Research

Title: Complex Traits Analysis in Mice
(must match the title of the funding source application)

New Application: Renewal of Protocol: # 1654 Pilot: Category *(see section II)*: C

1. Investigator Data:

Principal Investigator: Emil Skamene Phone #: 934-8038
Department: Medicine Fax#: 933-7146
Address: Room A6.149 1650 Cedar Ave Montreal Email: Emil.skamene@mubic.mcgill.ca

2. Emergency Contacts: Two people must be designated to handle emergencies.

Name: Emil Skamene Work #: 934-8038 Emergency #: 288-3303
Name: Daniel Houle Work #: 934-1934 x 44534 Emergency #: 514-381-9667

3. Funding Source: Canadian Genetic Diseases Network (NCE-CGDN)

External: X a/c 3700	Internal:
Source (s) : NCE-CGDN	Source (s) :
Peer Reviewed: YES X NO**	Peer Reviewed: YES NO**
Status : Awarded X Pending	Status: Awarded Pending
Funding period: 2003-2005	Funding period:

For Office Use Only:

ACTION	✓	DATE
CCs		
DB	GP	Sept 25/03
APPROVED		

** All projects that have not been peer reviewed for scientific merit by the funding source require 2 Peer Review Forms to be completed e.g. Projects funded from industrial sources. Peer Review Forms are available at www.mcgill.ca/rgo/animal

Proposed Start Date of Animal Use (d/m/y): or ongoing: X

Expected Date of Completion of Animal Use (d/m/y): or ongoing: X

Investigator's Statement: The information in this application is exact and complete. I assure that all care and use of animals in this proposal will be in accordance with the guidelines and policies of the Canadian Council on Animal Care and those of McGill University. I shall request the Animal Care Committee's approval prior to any deviations from this protocol as approved. I understand that this approval is valid for one year and must be approved on an annual basis.

Principal Investigator's signature: Date: July 21 2003

Approved by:

Chair, Facility Animal Care Committee:		Date: Sept. 3/03
University Veterinarian: Read		Date: Sept 16, 2003
Chair, Ethics Subcommittee (as per UACC policy):		Date:
Approved Animal Use	Beginning: July 1, 2003	Ending: June 30, 2004

This protocol has been approved with the modifications noted in Section 13.

4. Research Personnel and Qualifications

May 2002

ENTERED JAN 0 8 2004

SEP 1 2 2003