The Brisbane Systems Genetics Study: Genetical Genomics Meets Complex Trait Genetics

Joseph E. Powell[1,2]*, **Anjali K. Henders**[2], **Allan F. McRae**[1,2], **Anthony Caracella**[2], **Sara Smith**[2], **Margaret J. Wright**[2], **John B. Whitfield**[2], **Emmanouil T. Dermitzakis**[3], **Nicholas G. Martin**[2], **Peter M. Visscher**[1,2,4]❡, **Grant W. Montgomery**[2]❡

1 University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia, **2** Queensland Institute of Medical Research, Herston, Brisbane, Australia, **3** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **4** The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

Abstract

There is growing evidence that genetic risk factors for common disease are caused by hereditary changes of gene regulation acting in complex pathways. Clearly understanding the molecular genetic relationships between genetic control of gene expression and its effect on complex diseases is essential. Here we describe the Brisbane Systems Genetics Study (BSGS), a family-based study that will be used to elucidate the genetic factors affecting gene expression and the role of gene regulation in mediating endophenotypes and complex diseases. BSGS comprises of a total of 962 individuals from 314 families, for which we have high-density genotype, gene expression and phenotypic data. Families consist of combinations of both monozygotic and dizygotic twin pairs, their siblings, and, for 72 families, both parents. A significant advantage of the inclusion of parents is improved power to disentangle environmental, additive genetic and non-additive genetic effects of gene expression and measured phenotypes. Furthermore, it allows for the estimation of parent-of-origin effects, something that has not previously been systematically investigated in human genetical genomics studies. Measured phenotypes available within the BSGS include blood phenotypes and biochemical traits measured from components of the tissue sample in which transcription levels are determined, providing an ideal test case for systems genetics approaches. We report results from an expression quantitative trait loci (eQTL) analysis using 862 individuals from BSGS to test for associations between expression levels of 17,926 probes and 528,509 SNP genotypes. At a study wide significance level approximately 15,000 associations were observed between expression levels and SNP genotypes. These associations corresponded to a total of 2,081 expression quantitative trait loci (eQTL) involving 1,503 probes. The majority of identified eQTL (87%) were located within *cis*-regions.

Citation: Powell JE, Henders AK, McRae AF, Caracella A, Smith S, et al. (2012) The Brisbane Systems Genetics Study: Genetical Genomics Meets Complex Trait Genetics. PLoS ONE 7(4): e35430. doi:10.1371/journal.pone.0035430

Editor: Dan E. Arking, Johns Hopkins University, United States of America

Received November 24, 2011; **Accepted** March 16, 2012; **Published** April 26, 2012

Copyright: © 2012 Powell et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: We acknowledge funding by Australian National Health and Medical Research Council (NHMRC) grants 241944, 339462, 389927, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 552485, 552498 and Australian Research Council grants A7960034, A79906588, A79801419, DP0770096, DP0212016, DP0343921 for building and maintaining the adolescent twin family resource through which samples were collected. This research was supported by NHMRC grants 389892, 496667 and 613601. Dr. Montgomery and Dr. Visscher are supported by the NHMRC Fellowship Scheme. Dr. Dermitzakis acknowledges funds from the Louis-Jeantet Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: joseph.powell@uq.edu.au

❡ These authors contributed equally to this work.

Introduction

Dissection of the genetic architecture underlying quantitative traits and complex disease is essential to our understanding of the aetiology of complex diseases that cause most of the disease burden in society. Genome-wide association studies (GWAS) have proven to be an effective tool for identifying common causal loci of moderate to large effect size associated with human diseases and traits [1], [2]. However, as has been well discussed, GWAS have struggled to discover loci that collectively explain large proportions of the heritability of most complex traits [3], [4]. It has been proposed [5–7] that the nature of genetic variance for complex disease may be different to that of Mendelian disease that are caused by protein coding mutations, in that it may result from hereditary changes in gene regulation rather than gene variants that alter protein function. A number of recently published studies support this hypothesis [8–10]. More specifically, genetic differences between individuals in quantitative traits, endophenotypes (phenotypes that are risk factors for disease) and susceptibility to common diseases may be caused by differences in gene expression at a number of interacting loci [11]. Therefore, understanding the genetic basis of gene expression is likely to lead to a better understanding of genetic variation of quantitative traits and risk factors for common diseases.

Transcript abundance is a proximal endophenotype affected by genetic factors and has already facilitated the identification of candidate susceptibility genes for metabolic disease traits [12], psychiatric disorders [13], [14], asthma [15] and Crohn's disease [16]. This has mostly been possible when the tissue of expression was relevant to the interrogated complex trait, as disease phenotypes tend to manifest themselves only in certain tissues. Common disease is an endpoint of a complex pathway and there are likely to be many possible perturbations, or combinations thereof, in the pathways that lead to disease. Understanding the pathway as a whole, including the genetic determination of major hubs of gene regulation and how they relate to endophenotypes, will therefore be a major step in understanding the ultimate disease outcome and may provide information on the optimal choice of where to intervene in the pathway to prevent or treat disease.

Here we outline a major family-based system genetics study initiated to enhance our knowledge about common trait susceptibility by providing genome-wide expression and genotype data for 962 extensively phenotyped individuals in a family based design. This study builds on existing genotype and phenotype data collected at Queensland Institute of Medical Research (QIMR) through the collection of gene expression levels in whole blood (WB) (on all individuals) and lymphoblastoid cell lines (LCL) (on a subset of 50 MZ pairs). This study will help illuminate the genetic parameters underlying regulation of gene expression and to understand the genetic basis of complex traits through the relationship between gene expression and endophenotypes. We present a detailed description of the BSGS and report results on the association of expression levels with SNP genotypes from 862 BSGS individuals.

BSGS resource

BSGS comprises genotypic, expression and phenotype data that can be used to elucidate the genetic basis of gene expression and the relationship between gene expression and complex phenotypes. BSGS and its rationale are summarised in figure 1.

Gene expression data were generated for a total of 962 individuals from 314 families, for which we have genotype and phenotype data. Of the 962 individuals there are 128 MZ pairs (68 female and 60 male) and 206 DZ pairs (51 female, 53 male and 102 opposite sex). For 72 families expression data were collected on both parents as well as their offspring, allowing us to test parent-of-origin effects for gene expression. The 314 families comprise a variety of family structures and sizes, details of which are given in Figure 2. Details of the pairwise relationships are given in Table 1.

BSGS comprises the following data;

Expression. Gene expression data were collected in two stages: **Stage I** - involving 50 MZ pairs, on whom expression data were collected for both WB and LCL; **Stage II** - the remaining 862 individuals, where expression data were collected in WB only. A full description of the data collection for the stage I study is given in Powell et al. [17], and so for the sake of brevity, we focus on the description of stage II here. Transcript expression levels in WB were measured with over 47,000 genome-wide probes using the Illumina HT-12 v4.0 microarray chip (see materials and methods). Of the probes on the HT-12 v4.0 chip, approximately 29,000 are well characterized and annotated coding transcripts, 11,000 are coding transcript but poorly annotated, 4,000 are non-coding and 3,000 are experimentally confirmed mRNA sequences that align to EST clusters.

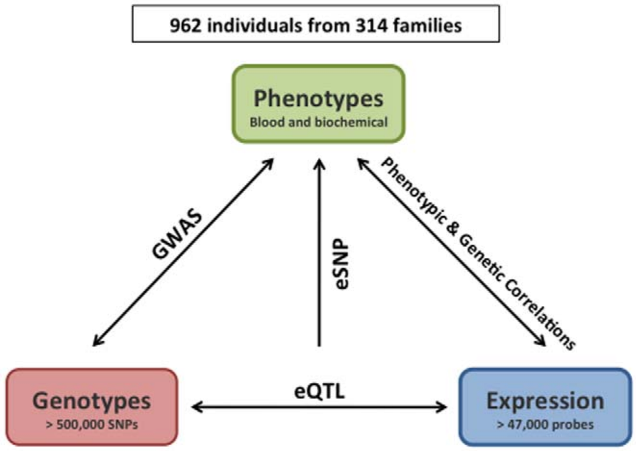

Figure 1. Summary of BSGS study design. The structure of the study design allows us to investigate fundamental questions about the genetic basis of gene expression and their correlation with phenotypes that are known risk factors for disease.
doi:10.1371/journal.pone.0035430.g001

Genotypic. Individuals are genotyped using the Illumina 610K chip [18–20]. After quality control procedures (see materials and methods) over 500 k genome-wide SNPs are available.

Phenotypes. A large number of hematology and blood biochemical phenotypes were measured at time of sample collection and later studies [19], [20–25]. In many cases these phenotypes have been analysed as part of a larger QIMR twin study focusing on linkage, association and heritability analyses. A brief summary of published results and phenotypes is given in table 2. Of particular relevance are the phenotypes measured in blood samples, including biochemical measures, because of their close relationship to the tissue in which expression levels have been measured. Such a relationship provides an ideal test case for the systems genetics approach that we are taking.

Results

Here we present results on expression data quality and an eQTL analysis using expression data from WB for the 862 (Stage II) individuals in BSGS. Details of expression quality for stage I individuals are given in Powell et al. [17].

Stage II expression data quality

Gene expression levels were generated for the 862 individuals from WB samples using the PAXgene^{TM} tube system. Expression levels of extracted mRNA were measured using Illumina HumanHT-12 v4.0 whole genome chip. For use as a research resource the quality of the expression data is important, although quality control procedures are likely to differ depending on specific analyses. Original data consists of raw, un-normalised, expression levels with chip background levels subtracted for 47,323 probes. Illumina software GenomeStudio also calculates a detection p-value (see methods), which represents the probability that a given transcript is expressed above the chip background level. A value below a p-value threshold (i.e. 0.05) indicates a gene is detected and thus, the number of detected transcripts is a good overall indicator of sample expression quality, with all samples on a given chip expected to have similar numbers of transcripts detected. All 862 samples show high numbers of probes detected with $p<0.05$ (mean = 17,310 and SD = 1,512) with similar numbers within and between chips (Figure S1). These levels are similar or slightly higher than other published reports using Illumina chips (for example in Monocytes [26], Lymphoblastoid cell lines (LCL) [27], [28], Fibroblasts and T-cells [28]).

Gene expression is known to vary between individuals due to differences in environmental [29], [30] and genetic [31] factors. Additionally, differences in the genetic control of gene expression between tissues [17], [32], coupled with specific environments of cell types, lead to certain genes being expressed in some tissues and not others at the time of sample collection [33]. A consequence is that we expect many probes not to be expressed in a proportion of the individuals in our sample. Studies of gene expression data use a

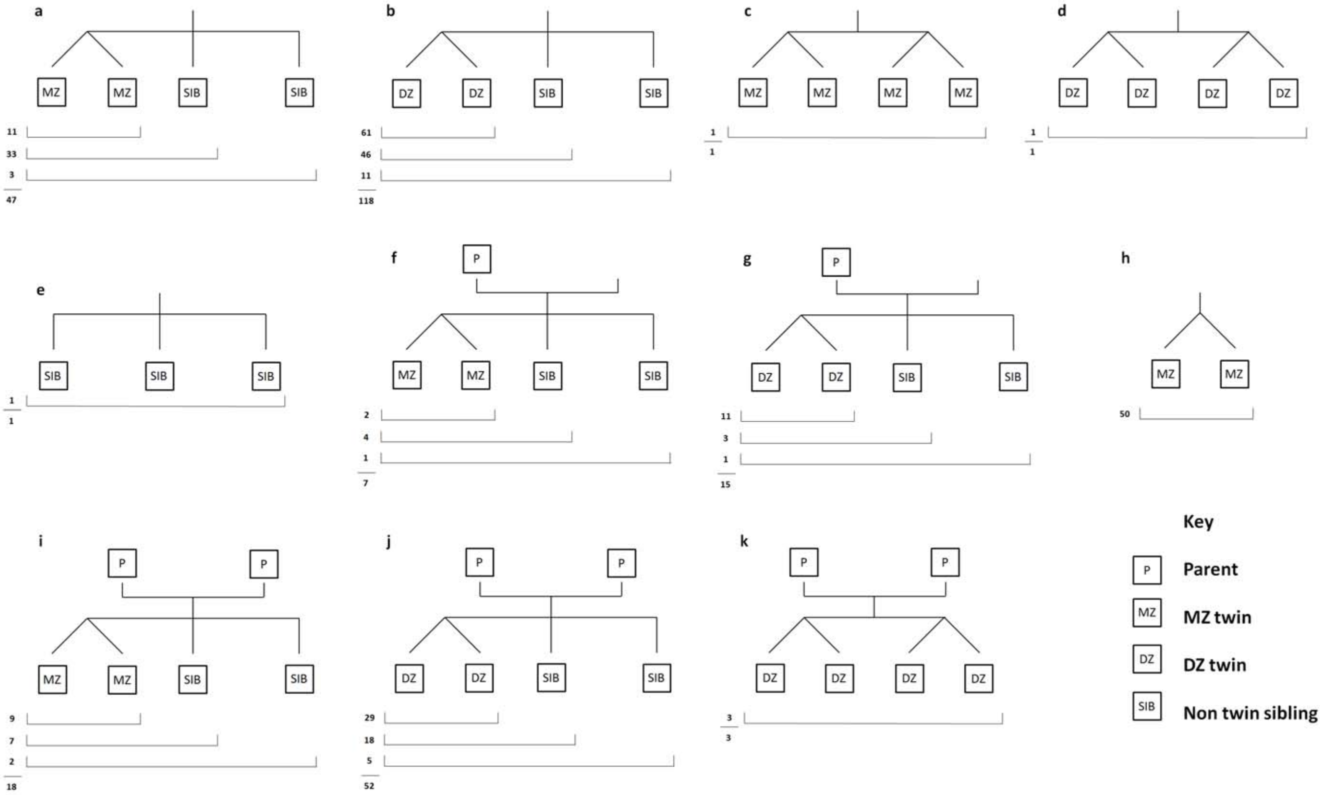

Figure 2. Samples collected in BSGS comprise of a number of different families. Family structure **h** represents the 50 MZ pairs comprising the stage I study. The remaining family structures are from stage II. The numbers of each family structure are given below the pedigree diagram. By utilising expression information contained between and within twin pairs, siblings and between progeny and parents we are able to estimate genetic and non-genetic variance components using linear mixed models.
doi:10.1371/journal.pone.0035430.g002

variety of criteria to select probes for inclusion in analyses. A commonly used criterion is that a given gene must be detected as expressed in a certain proportion of the samples, with the particular proportion decided based on requirements and constraints of the analyses. Of the 47,323 probes 5,364 (11.3%) are not detected as expressed in any of the 862 individuals, whilst 6,281 (13.3%) are detected as expressed in all individuals (Figure S2).

Identification of eQTL

Of the 47,323 probes whose expression levels were measured on the chip, only probes that were well characterised and detected as expressed in 10% of the sample were carried forward, leaving a total of 17,926 probes (corresponding to 13,533 RefSeq genes) whose expression was tested for association with the 528,509 SNPs. At the study wide significance threshold (5.25×10^{-12}) a total of 14,916 associations were identified involving 10,421 SNPs (eSNPs) and 1,503 probes. In situations where a probe had multiple significant SNP associations, independent eQTL were called if the distance between significant SNPs was greater than 2MB. Using this criterion a total of 2,081 eQTL were called for the 1,503 probes with 1,357 (90%) of probes having a single eQTL and 19 probes with greater than five eQTL. The median number of significant SNPs within an eQTL region was 4, although 351 (16.8%) of eQTL contained over 10 significant SNPs and 27 (1.3%) over 50. Of the 10,421 eSNPs the majority (73.6%) were associated with a single probe whilst the remaining 26.4% have

Table 1. Relationship pairing between 962 individuals in BSGS.

Relationship pairs	Code	N	Notes
Monozygotic twins	MZ	128	68 female pairs; 60 male pairs; 50 MZ pairs form stage I, where we have expression data from both WB and LCL RNA sources
Dizygotic twins	DZ	206	51 female pairs; 53 male pairs; 102 mixed sex pairs
Siblings	SIB	343	81 female pairs; 82 male pairs; 180 mixed sex pairs
Parent – Offspring	PO	425	98 father – daughter pairs; 103 father – son pairs; 113 mother – daughter pairs; 111 mother – son pairs
Parent – Parent	PP	71	

doi:10.1371/journal.pone.0035430.t001

Table 2. Central phenotypes in BSGS and a brief summary of previous studies identifying genetic parameters and association signals.

Phenotypes	Summary	Reference
Hemoglobin concentration	Hemoglobin phenotypes have been associated with SNPs in *TMPRSS6, LRRC16A, HK1* and *HK2*.	[19], [20]
Red blood cell count	Association with SNPs close to *IRX6*.	[20]
Platelet count	Platelet count has suggestive association with SNPs in *KCNIP*.	[20]
White blood cell count	White blood cell count has suggestive association with SNPs in *MACF1*.	[20]
Monocytes	Monocyte count is associated with SNPs close to *BAG4* and *ITGA4*.	[20]
Eosinophils	Eosinophil count has suggestive association with SNPs in *ITPR1*.	[20]
CD4+/CD8+ T-cell ratio	Collectively, these phenotypes are associated with SNPs in the MHC and the Schlafen family of genes. They are also endophenotypes for Type 1 Diabetes, HIV-1 immune control and autoimmune diseases.	[25]
Plasma Cholesterol (HDL and LDL) and Triglyceride concentrations	A known endophenotypes for cardiovascular disease. Data comprising part of BSGS have shown strong associations between Cholesterol and genes on chromosome 19 and between Triglyceride and genes on chromosome 7.	[18], [22]
Blood pressure	An important endophenotype for hypertension. Data comprising part of BSGS have shown strong associations between blood pressure and genes on chromosomes 4,5,14 and 17.	[23], [24]
Iron, Ferritin and Transferrin levels	Collectively, these phenotypes show association with SNPs in *TMPRSS6, HFE, PGM1* and *TF*.	[19], [61], [62]

doi:10.1371/journal.pone.0035430.t002

significant associations with two or more probes, with the maximum being eight.

Among the 2,081 eQTL the SNP with the smallest *p*-value had R^2 (proportion of expression variability explained) from 4.6% to >80% with a median of 12.1% (Figure 3). A total of 328 eQTLs had R^2 greater than 25% and 61 greater than 50%. It is worth noting that these estimates will be biased upwards because hypothesis testing and estimation were performed on the same data ("winner's curse").

Of the 2,081 eQTL, 1,810 (87%) are *cis*-acting, being located within a 2MB region either side of the 3′ or 5′ end of the Transcription Start Site (TSS). The remaining 271 eQTL are *trans*-acting, although the majority of these (232) are located on the same chromosome as their probes TSS. Twelve probes had both *cis*-and *trans*-acting eQTL. To explore the location of *cis*-eQTLs, the distance of the most significant SNP for each *cis*-eQTL per probe was mapped relative to the TSS. In agreement with previous studies [28], [34–35] a strong signal was found close to the TSS, with no discernable trend in a 3′ or 5′ direction (Figure 4). This symmetrical trend is likely to reflect variation in core regulatory sequences such as promoter elements. Details of the top twelve *cis*-eQTL (based on eSNP *p*-values) are given in table 3.

Regulatory control for gene expression, particularly *trans*-acting, is not evenly distributed across the genome [36]. To provide an outline of the positions of regulatory control for gene expression across the genome, we produced a Manhattan plot showing the number of eSNP (each eSNP represents a single eQTL) within each 1MB region (Figure 5). *Cis*-acting eSNPs are distributed across the genome, as has been reported by multiple studies [26–27], [37]. We observe that *trans*-acting eSNPs are often found in close proximity to one another, supporting the concept of *trans*-regulatory hotspots [38].

Many studies mapping eQTL consider just associations within the *cis*-region: the main benefit being a considerably reduced burden of multiple testing compared to studies including *trans*-acting SNPs. Given correlations exist between both SNP genotypes and probe expression levels, it is likely that our study-wide significance threshold (corrected for all SNP by probe tests) results in true associations being missed. By reducing the significance threshold to less conservative levels, the number of *trans*-eQTL considerably increased, as expected by chance (Table 4). However, the number of *cis*-eQTL only moderately increased, indicating that conservative threshold used did not result in a considerable under estimation of the true number of *cis*-eQTL (Table 4).

Discussion

Through understanding the genetic control of gene expression and its relationship with indicators of common disease we can gain a better understanding of disease itself. Common disease is an endpoint of a complex pathway of genetic and environmental factors and there are likely to be many possible perturbations in the pathways that lead to disease. Gaining insights into the pathway as a whole, including the genetic control of major hubs of gene regulation and how they relate to endophenotypes, is a major step in understanding the ultimate outcome of (disease) and may provide information on the optimum choice of where to intervene in the pathway to prevent or treat disease. Recent success at identifying genetic determinants for networks of regulatory control affecting metabolic disease [39] has highlighted the application of systems genetic approaches in understanding the aetiology of complex phenotypes. We aim to use BSGS for the identification of polymorphisms, or combinations thereof, that affect gene regulation in one or more genes which in turn cause differences between individuals in endophenotypes or susceptibility to disease and may lead to drug discovery or other treatments to decrease the burden of common disease in society.

The cellular origin of many biochemical and blood phenotypes is often complex with some traits originating in other tissues but displayed in blood. For example, insulin originates from β-cells in the pancreas but is carried in the blood where is performs its role of regulating glucose levels. However, insulin levels are regulated

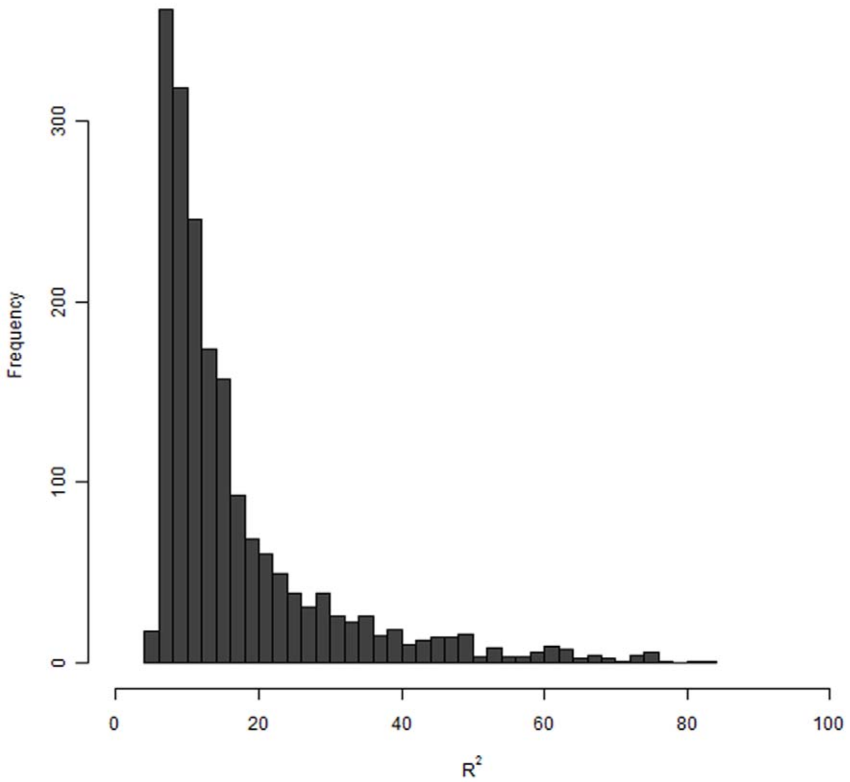

Figure 3. Distribution of the *R^2* observed for the best eSNP from the 1,885 eQTLs.
doi:10.1371/journal.pone.0035430.g003

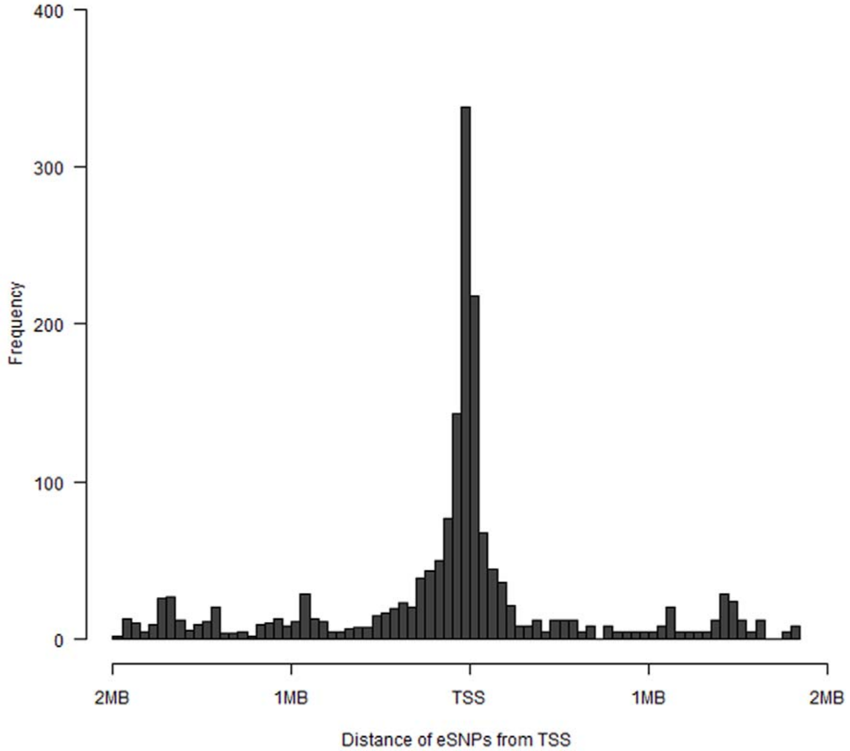

Figure 4. The distribution of *cis*-eSNPs distance from the Transcription Start Site (TSS). The distances of eSNPs from the TSS were divided into 50KB bins across the *cis*-region.
doi:10.1371/journal.pone.0035430.g004

Table 3. Top *cis*-eQTL results.

Probe ID	Gene	Probe Chromosome	TSS bp	Top SNP	Top SNP location	-log10 *p*-value	R^2
ILMN_1715169	HLA-DRB1	6	32654825	rs9271170	6 - 32577889	131.5	73.5
ILMN_1743145	ERAP2	5	96274820	rs10051637	5 - 96279490	131.0	81.6
ILMN_1798177	CHURC1	14	64471249	rs7143432	14 - 65379146	130.4	83.0
ILMN_2209027	RPS26	12	54722494	rs10876864	12 - 56401085	122.0	74.6
ILMN_2403228	CLEC12A	12	10029119	rs7313235	12 - 10132283	121.3	75.8
ILMN_2352023	RIPK5	1	203378372	rs12139373	1 - 205054879	114.4	70.2
ILMN_2312606	IRF5	7	127973722	rs6965542	7 - 128655918	112.6	74.8
ILMN_1791511	TMEM176A	7	150133038	rs7806458	7 - 150476888	107.5	66.4
ILMN_2038775	TUBB2A	6	3154070	Rs9392465	6 - 3162378	107.2	66.7
ILMN_3298167	ZSWIM7	17	15879944	rs1045599	17 - 15879910	105.5	65.9
ILMN_1661266	HLA-DQB1	6	32736001	rs9273349	6 - 32625869	102.0	61.2
ILMN_2313901	PAM	5	102340879	rs28092	5 - 102149795	99.7	66.2

The chromosome and base pair position of the probe transcription start site (TSS) are given for each probe. R^2 is the proportion of transcript level variance explained by the SNP with the strongest association.

doi:10.1371/journal.pone.0035430.t003

by multiple factors, many relating to blood components and so it is reasonable to expect that the expression of some genes in blood will have an effect on insulin levels. Thus, using a tissue closely related to phenotypes we are able to relate transcript abundance to the blood and biochemical traits and hopefully expound the role of gene regulation in mediating endophenotypes and complex disease. Nevertheless, findings should be analysed with caution as we and others [17], [32], have recently shown that the genetic basis for gene expression is often different between tissues.

We hypothesise that the indirect route from genotype to phenotype via gene expression is more informative to elucidate the nature of complex trait variation than the direct route from genotype to phenotype. The expression data collected from families allows us to estimate quantitative genetic variance components through the use of linear mixed models and estimation methods such as restricted maximum likelihood (REML) [40]. By utilising information contained between and within twin pairs, siblings and between progeny and parents we are able to estimate additive and non-additive genetic effects of gene expression and can explicitly test a variance partition model that includes parent-of-origin effects.

Here we have presented results of an eQTL study involving ~18 k probes, each tested against >500 k SNPs. Despite numerous large scale studies reporting eQTL [26–28], [34], [37], [41–42], generating eQTL results for a particular study is important as differences in the genetic control of gene expression between populations [43–45] and between tissues [17], [32] is well known. Furthermore, here, and elsewhere [26], [41–42] large sample sizes have permitted the investigation of *trans*-acting eSNP often ignored by smaller studies [28], [34].

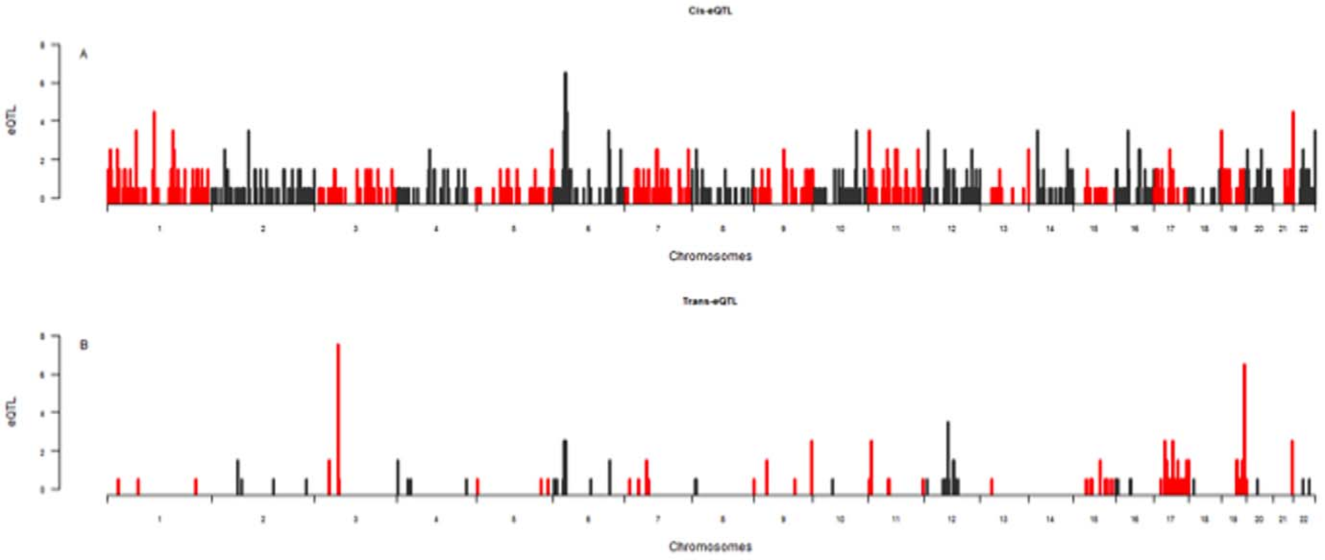

Figure 5. Positions of *cis* (A) and *trans* (defined as greater than 2MB from the transcription start site) (B) eSNP across the genome. The number of eSNP within 1MB bins is shown. A single eSNP represents a unique eQTL.

doi:10.1371/journal.pone.0035430.g005

Table 4. Number of associations and eQTL at various levels of significance threshold.

Significance threshold	Total associations	Expected number of associations	Total significant SNPs	Probes with 1+ significant SNP	eQTL	Cis-eQTL	Trans-eQTL
$10e^{-8}$	31,032	951	20,068	3,969	5,679	2,673	3,006
$10e^{-10}$	20,049	9.5	13,319	1,933	2,512	1,953	559
$5.25e^{-12}$	**14,210**	**0.05**	**10,160**	**1,503**	**1,885**	**1,529**	**256**
$10e^{-15}$	10,519	0	7,468	1,126	1,507	1,256	251
$10e^{-25}$	4,362	0	3,382	572	731	639	92
$10e^{-50}$	896	0	764	185	199	172	27
$>10e^{-100}$	60	0	55	27	27	22	5

The study-wide significance threshold employed for our eQTL analysis is highlighted. The expected number of associations is the number of association that are expected to be observed under the null hypothesis of no associations between probe expression levels and SNPs.
doi:10.1371/journal.pone.0035430.t004

Methods

Sample Collection

Individuals present in this study were recruited as part of the Brisbane Twin Nevus and cognition studies (known as BTN and MAPS respectively). This study was approved by the Queensland Institute for Medical Research-Human Research Ethics Committee. All participants gave informed written consent. As described in detail elsewhere [21], [46–50], adolescent MZ and DZ twins, their siblings, and their parents have been recruited over a 16 year period into an ongoing study of the genetic and environmental factors influencing pigmented nevi and the associated risk of developing skin cancer and cognition. The sample is of northern European origin (mainly Anglo-Celtic). A Principal Component Analysis (PCA) comparing individuals in this study to HapMap3 [51] and GenomEUtwin [52] populations shows the close similarity of ancestry to northern European populations (Figure S3). All participants gave informed consent, and the study protocol was approved by the appropriate institutional review boards.

Analysis of Stage II data

Genotyping. DNA samples from each individual were genotyped by the Scientific Services Division at deCODE Genetics, Iceland, using the Illumina 610-Quad Beadchip. Genotypes were called with the Illumina BeadStudio software. Full details of genotyping procedures are given in Medland et al. [53]. Standard QC filters were applied so that further analysis was on samples and SNPs with high data quality. We first applied filters to SNP data before evaluating genotyping quality per individual. After removing SNPs with minor allele frequencies (MAF) <1% with a mean BeadStudio GenCall <0.7 a total of 528,509 remained for further analysis.

Gene expression quantification. Whole blood for expression profiling was collected directly into PAXgene^{TM} tube (QIAGEN, Valencia, CA). Total RNA was extracted from PAXgene^{TM} tubes using the WB gene RNA purification kit (QIAGEN, Valencia, CA). RNA from all samples was run on an Agilent Bioanalyzer to assess quality and to estimate RNA concentrations RNA was converted to cDNA, amplified and purified using the Ambion Illumina TotalPrep RNA Amplification Kit (Ambion).

Expression profiles were generated by hybridising 750 ng of cRNA to Illumina HumanHT-12 v4.0 Beadchip according to Illumina whole-genome gene expression direct hybridization assay guide (Illumina Inc, San Diego, USA). Briefly, 500 ng of total RNA were used to generate biotinylated cRNA, which was fragmented and hybridised to an Illumina whole genome expression chip, HumanHT-12 v4.0 for 18 h at 58°C. Beadchips were then washed and stained and subsequently scanned to obtain fluorescence intensities. Samples were scanned using an Illumina Bead Array Reader. Samples were randomised across chips and chip positions, with check for balance across families, sex and generation.

eQTL analysis

Normalisation and processing. The following normalisation procedures were applied to the raw expression data for the eQTL analysis. Pre-processing of data generated by the Illumina Bead Array Reader was done using Illumina software, GenomeStudio (Illumina Inc., San Diego). Pre-processing included; correction for chip background effects, removal of outlier beads, computation of average bead signal and calculation of detection *p*-values using negative controls present on the array. Removal of chip background effects can lead to negative expression levels for transcripts with low levels of measured expression. To avoid problems with further normalisation procedures, negative values were changed to missing data identifiers. Thus, in subsequent normalisation procedures and analyses probes coded as missing are ignored.

The Illumina HT-12 v4.0 chip contains 47,323 probes, although some probes are not assigned to RefSeq genes. To avoid spurious associations, we removed 678 probes for which at least one HapMap 3 (NCBI build 36, dbSNP b126) SNP was within the probe sequence [54]. For the eQTL analysis we removed any probes where less than 10% of samples had a detection *p*-value<0.05. Of the 24,317 probes retained, the mean of the proportion of samples with *p*-values<0.05 was 97%, implying that relatively little missing data remained within the expression dataset. After removing 6,322 putative and/or not well-characterised genes i.e. probe names starting with HS ($n = 1,841$), KIAA ($n = 158$) and LOC ($n = 4,323$), 17,926 well-characterised detected probes remained for analysis, which corresponds to 13,486 RefSeq genes.

To minimise the influence of overall signal levels, which may reflect RNA quantity and quality rather than a true biological difference between individuals, the following standardisation procedures were applied to the 17,926 probes. Adjusted expression levels for each probe were transformed using a quantile transformation [55], [56] to achieve a stabilized distribution across average expression levels. Further normalisation was performed to allow expression levels to be compared across chips and genes. This was achieved fitting a linear mixed model;

$$y_{ijk} = \mu + C_j + P_k + \varepsilon_{ijk} \qquad (1)$$

Where y_{ijk} is the log-transformed expression level for individual i on chip j. The variable μ represents the mean expression level across all individuals and C_j and P_k are random effects removing variation in the data due to chip j and chip position k respectively, and ε_{ijk} is the residual. The between chip variance is expected to be small due to the scaling that was performed during the pre-processing of the data. The residuals from this model were used in all further analyses. To increase robustness, the distribution of normalised expression levels for each probes were tested for deviation from normality using the Shapiro-Wilk test. All 17,926 probes had normally distributed ($p<0.05$) expression levels.

Testing for association. We tested for association between the 528,509 genotyped SNPs and the normalised expression levels of the 17,926 probes using the FASTASSOC component of MERLIN [57], [58]. The FASTASSOC option fits a simple linear regression model to estimate an additive effect for each probe and SNP combination, with SNP genotypes coded as the number of copies of the minor allele (0, 1 or 2) carried by each individual. We used the Lander-Green algorithm [58], [59], implemented in Merlin, to estimate expected genotype scores for individuals with missing genotype data. Covariates of sex and generation were included in the model, where generation denotes either the parental or the adolescent generation. Previous analysis has shown (not published) that generation is a useful substitute for age without the burden of additional degrees of freedom. The model applies a variance component approach to account for the correlations between different expression levels within each family. The model fit is evaluated using a score test, which substantially reduces computational time compared to maximum-likelihood methods, at the expense of a slight loss of power [58].

Conditional regression analysis was used to address the potential to miss secondary eQTL in linkage disequilibrium (LD) with other eQTL. For each probe with an identified eQTL we corrected for the main effects of the top eSNP (SNP with the largest R^2) by regressing its genotypes against the expression levels. Residuals from this analysis were then used for second round of eQTL mapping, allowing us to detect independent eQTL. If additional eQTL were identified from this second round of analysis, the process was repeated, correcting for the main effects of the top eSNP from the first and second eQTL using multivariate regression.

Associations were evaluated in two categories depending on the location of the SNP relative to the transcription start site (TSS). Cis-eQTL were defined as associations between SNPs within 2MB of either the 3′ or 5′ end of the TSS. We defined *trans*-associations as associations involving SNPs elsewhere in the genome. To correct for multiple testing, we used a study-wide significance level of 0.05, corrected for the number of SNP by probe associations tested, corresponding to a p-value threshold of 5.25×10^{-12}.

We tested for the effects of population structure and cryptic relatedness between individuals by applying the method 'genomic control' [60] to results of the association analysis. We derived a coefficient of 1.002, indicating negligible population stratification.

Supporting Information

Figure S1 GenomeStudio provides a *p*-value for each transcript in each sample. For a given sample the number of transcripts with *p*-values below a given threshold provides an indication of its quality. Figure S1 shows the number of transcripts with $p<0.05$ for each sample with each colour representing a single chip.
(DOCX)

Figure S2 The distribution of the number of probes detected as expressed in stage II. A total of 47,323 are measured on the Illumina HT-12 v4.0 chip, of these 5,364 (11.3%) are not detected as expressed in any of the individuals, whilst 6,281 (13.3%) are detected as expressed in all individuals.
(DOCX)

Figure S3 A Principle Component Analysis (PCA) of 16 global populations and the individuals collected in this study. Principal Component one (PC1) and two (PC2) values were derived from approximately 280,000 autosomal markers. Populations samples marked with *1 were collected as part of the HapHap3 project [1] and *2 as part of the GenomEUtwin project [2]. **ASW***1 African Americans, **CEU***1 European Americans, **CHB***1 Han Chinese, **CHD***1 Chinese, **GIH***1 Guajarati-Indians, **JPT***1 Japanese, **LWK***1 Luhya Kenyans, **MEX***1 Mexicans, **MKK***1 Maasai Kenyans, **TSI***1 Italians, **TRI***1 Yorubans Nigeria, **DEN***2 Danish, **FIN***2 Finish, **NET***2 Dutch, **SWE***2 Swedish, **UK***2 British.
(DOCX)

Acknowledgments

We gratefully acknowledge the participation of the individuals sampled in this work. We would like to thank Lisa Bowdler and Steven Crooks for their technical assistance with the microarray hybridizations, Alison Mackenzie, Marlene Grace and Ann Eldridge for data collection and Dale Nyholt and Scott Gordon for data management.

Author Contributions

Conceived and designed the experiments: JEP AFM ETD PMV GWM. Performed the experiments: AKH AC SS. Analyzed the data: JEP. Contributed reagents/materials/analysis tools: JBW MJW NGN. Wrote the paper: JEP PMV GWM.

References

1. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728–731.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
3. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
5. Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? Nature Genet 26: 151–157.
6. Tan HY, Callicott JH, Weinberger DR (2008) Intermediate phenotypes in schizophrenia genetics redux: is it a no brainer? Mol Psychiatr 13: 233–238.
7. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet 6: e1000888.
8. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37: 710–717.
9. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423-U422.
10. Naukkarinen J, Surakka I, Pietilainen KH, Rissanen A, Salomaa V, et al. (2010) Use of Genome-Wide Expression Data to Mine the "Gray Zone" of GWA Studies Leads to Novel Candidate Obesity Genes. PLoS Genet 6: e1000976.
11. Khor B, Gardet A, Xavier RJ (2011) Genetics and pathogenesis of inflammatory bowel disease. Nature 474: 307–317.

12. Chen YQ, Zhu J, Lum PY, Yang X, Pinto S, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452: 429–435.
13. Iwamoto K, Bundo M, Washizuka S, Kakiuchi C, Kato T (2004) Expression of HSPF1 and LIM in the lymphoblastoid cells derived from patients with bipolar disorder and schizophrenia. J Hum Genet 49: 227–231.
14. Kato T, Iwayama Y, Kakiuchi C, Iwamoto K, Yamada K, et al. (2005) Gene expression and association analyses of LIM (PDLIM5) in bipolar disorder and schizophrenia. Mol Psychiatry 10: 1045–1055.
15. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.
16. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40: 1107–1112.
17. Powell JE, Henders AK, McRae AF, Wright MJ, Martin NG, et al. (2012) Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res 22: 456–466.
18. Middelberg RP, Martin NG, Montgomery GW, Whitfield JB (2006) Genome-wide linkage scan for loci influencing plasma triglycerides. Clin Chim Acta 374: 87–92.
19. Benyamin B, Ferreira MA, Willemsen G, Gordon S, Middelberg RP, et al. (2009a) Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. Nat Genet 41: 1173–1175.
20. Ferreira MA, Hottenga JJ, Warrington NM, Medland SE, Willemsen G, et al. (2009) Sequence variants in three loci influence monocyte counts and erythrocyte volume. Am J Hum Genet 85: 745–749.
21. Wright M, De Geus E, Ando J, Luciano M, Posthuma D, et al. (2001) Genetics of cognition: outline of a collaborative twin study. Twin Res 4: 48–56.
22. Beekman M, Heijmans BT, Martin NG, Whitfield JB, Pedersen NL, et al. (2003) Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population. Eur J Hum Genet 11: 845–850.
23. Hottenga JJ, Whitfield JB, Posthuma D, Willemsen G, de Geus EJ, et al. (2007) Genome-wide scan for blood pressure in Australian and Dutch subjects suggests linkage at 5P, 14Q, and 17P. Hypertension 49: 832–838.
24. O'Connor DT, Zhu G, Rao F, Taupenot L, Fung MM, et al. (2008) Heritability and genome-wide linkage in US and australian twins identify novel genomic regions controlling chromogranin a: implications for secretion and blood pressure. Circulation 118: 247–257.
25. Ferreira MA, Mangino M, Brumme CJ, Zhao ZZ, Medland SE, et al. (2010) Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. Am J Hum Genet 86: 88–92.
26. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. PLoS One 5: e10693.
27. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nature Genet 39: 1217–1224.
28. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. Science 325: 1246–1250.
29. Choi JK, Kim SC (2007) Environmental effects on gene expression phenotype have regional biases in the human genome. Genetics 175: 1607–1613.
30. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, et al. (2008) Genetic Analysis of Human Traits In Vitro: Drug Response and Gene Expression in Lymphoblastoid Cell Lines. PLoS Genet 4: 16.
31. Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet 10: 595–604.
32. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, et al. (2011) Single-Tissue and ross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. PLoS Genet 7: e1001317.
33. Jongeneel CV, Delorenzi M, Iseli C, Zhou DX, Haudenschild CD, et al. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res 15: 1007–1014.
34. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
35. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet 7: e1002003.
36. Montgomery SB, Dermitzakis ET (2011) From expression QTLs to personalized transcriptomics. Nat Rev Genet 12: 277–282.
37. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nature Genet 39: 1202–1207.
38. Breitling R, Li Y, Tesson BM, Fu J, Wu C, et al. (2008) Genetical genomics: spotlight on QTL hotspots. PLoS Genet 4: e1000232.
39. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, et al. (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nature Genet 43: 561-U590.
40. Lynch M, Walsh B (1998) Genetics and the analysis of quantitative traits. Sinauer associates, Sunderland MA.
41. Myers AJ, Gibbs JR, Awebster J, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nature Genet 39: 1494–1499.
42. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: 1020–1032.
43. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, et al. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nature Genet 39: 226–231.
44. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, et al. (2007) Gene-expression variation within and among human populations. Am J Hum Genet 80: 502–509.
45. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, et al. (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nature Genet 42: 62–U79.
46. Aitken JF, Green AC, MacLennan R, Youl P, Martin NG (1996) The Queensland Familial Melanoma Project: study design and characteristics of participants. Melanoma Res 6: 155–165.
47. McGregor B, Pfitzner J, Zhu G, Grace M, Eldridge A, et al. (1999) Genetic and environmental contributions to size, color, shape, and other characteristics of melanocytic naevi in a sample of adolescent twins. Genet Epidemiol 16: 40–53.
48. Wright MJ, Martin NG (2004) Brisbane Adolescent Twin Study: Outline of study methods and research projects. Australian Journal of Psychology 56: 65–78.
49. Zhu G, Montgomery GW, James MR, Trent JM, Hayward NK, et al. (2007) A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. Eur J Hum Genet 15: 94–102.
50. Falchi M, Bataille V, Hayward NK, Duffy DL, Bishop JAN, et al. (2009) Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. Nat Genet 41: 915–919.
51. International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
52. Peltonen L (2003) GenomEUtwin: a strategy to identify genetic influences on health and disease. Twin Res 6: 354–360.
53. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, et al. (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. Am J Hum Genet 85: 750–755.
54. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, et al. (2007) Sequence polymorphisms cause many false cis eQTLs. PLoS ONE 2: e622.
55. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
56. Smyth GK, Speed TP (2003) Normalization of cDNA microarray data. Methods 31: 265–273.
57. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
58. Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81: 913–926.
59. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84: 2363–2367.
60. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology 60: 155–166.
61. Benyamin B, McRae AF, Zhu G, Gordon S, Henders AK, et al. (2009b) Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. Am J Hum Genet 84: 60–65.
62. Kutalik Z, Benyamin B, Bergmann S, Mooser V, Waeber G, et al. (2011) Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. Hum Mol Genet 20: 3710–3117.