

2 MATERIALS AND METHODS

In this chapter I will:

- Describe the population samples analysed in this thesis.
- Define the sets of SNPs and genes tested for association.
- Introduce the statistical tests used for association of SNP genotype with mRNA levels, as well methods for significance correction.
- Outline the particulars of the three studies making up this thesis:
 - Impact of eQTL-nsSNP interaction on gene expression in cis and trans (Chapter 3)
 - Fine-scale architecture of the cis regulatory landscape (Chapter 4)
 - Cell type specificity of eQTLs (Chapter 5)

2.1 THE SAMPLES

The population samples studied in this thesis belong to two resources that have been set up to explore human genetic variation: the HapMap Project and the GenCord Project. In the following sections I give a brief outline of these resources.

2.1.1 The HapMap Project

The International HapMap Project was launched in 2002 as a collaborative effort to identify and catalogue genetic similarities and differences in human populations (International HapMap Consortium 2003). The ultimate goal of HapMap was to provide a public resource for medical genetic research by developing a detailed haplotype map (HapMap) of the human genome that would describe common patterns of genetic variation. The core strategy of this project involved genotyping DNA from LCLs

generated from blood samples of individuals belonging to a diverse set of populations. The project is ongoing and is currently in its third phase. (International HapMap Consortium 2003; International HapMap Consortium 2005; International HapMap Consortium 2007).

The aim of HapMap Phase 1 was to genotype at least one common SNP per five kb across the euchromatic portion of the genome of 269 individuals from four geographically distinct populations. The individuals genotyped were: 30 mother–father–adult child trios of northern and western European ancestry living in Utah from the CEPH collection (abbreviated CEU), 45 unrelated Han Chinese individuals in Beijing, China (CHB), 44 unrelated Japanese individuals in Tokyo, Japan (JPT) and 30 trios from the Yoruba in Ibadan, Nigeria (YRI). Approximately 1.3 million SNPs were genotyped per population and a detailed description of this resource was published in 2005 (International HapMap Consortium 2005).

In Phase 2, a further 2.1 million SNPs were genotyped in each of 270 individuals (Phase 1 individuals and an additional sample from the JPT population). The resulting HapMap had a SNP density of approximately one SNP per kb and was estimated to contain approximately 25–35% of all common SNPs (9-10 million SNPs with a MAF \geq 0.05) in the assembled human genome. A description of this resource was published in 2007 (International HapMap Consortium 2007).

Phase 3 of the HapMap was ongoing at the time of writing and involved additional individuals from the four initial populations, as well as seven additional populations. Over 4 million SNPs were genotyped for 541 individuals of the four initial populations (CEU, CHB, JPT, YRI) and approximately 1.5 million SNPs were genotyped in 760 individuals of seven new populations (90 ASW: African ancestry in Southwest USA; 100 CHD: Chinese in Metropolitan Denver, Colorado, USA; 100 GIH: Gujarati Indians in Houston, Texas, USA; 100 LWK: Luhya in Webuye, Kenya; 90 MEX: Mexican ancestry in Los Angeles, California, USA; 180 MKK: Maasai in Kinyawa, Kenya; 100

TSI: Tuscans in Italy). At the time of writing, this study was in preparation for publication.

Table 2 summarizes SNPs and individuals assayed in each of the three phases of the project. Data analysed in this thesis include all four HapMap Phase 2 populations (210 unrelated individuals from CEU, CHB, JPT and YRI) in the study described in Chapter 3. Additional samples for those populations as well as four of the seven new HapMap Phase 3 populations (792 unrelated individuals from CEU, CHB, GWK, JPT, LWK, MEX, MKK and YRI) were analysed in the study described in Chapter 4.

		Populations										
		CEU	CHB	JPT	YRI	ASW	CHD	GIH	LWK	MEX	MKK	TSI
Phase 1	SNPs	1,105,063	1,087,365	1,087,365	1,076,442							
	Individuals	90	45	44	90							
Phase 2	SNPs	3,904,218	3,936,482	3,936,482	3,846,092							
	Individuals	90	45	45	90							
Phase 3	SNPs	4,030,562	4,052,129	4,052,216	3,984,146	1,561,382	1,306,152	1,407,818	1,529,693	1,410,231	1,537,561	1,419,861
	Individuals	180	90	91	180	90	100	100	100	90	180	100

Table 2. Summary of SNPs and individuals assayed in each of the three phases of HapMap. Population descriptors: CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; YRI: Yoruban in Ibadan, Nigeria; ASW: African ancestry in Southwest USA; CHD: Chinese in Metropolitan Denver, Colorado, USA; GIH: Gujarati Indians in Houston, Texas, USA; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California, USA; MKK: Maasai in Kinyawa, Kenya; TSI: Tuscans in Italy.

2.1.2 The GenCord Project

The GenCord project is a collection of cell lines derived from umbilical cords of 85 individuals of Western European origin, following appropriate consent and ethical approval (Dimas, Deutsch et al. 2009). The project was conceived as a resource for the identification of QTLs involved in the regulation of cellular phenotypes in primary

fibroblasts, LCLs and primary T-cells. Umbilical cord was chosen because it is readily available and allows the acquisition of multiple cell types for each individual. Sample collection was performed systematically on full term or near full term pregnancies to ensure homogeneity for sample age.

2.1.3 Using HapMap and GenCord to investigate regulatory variation

HapMap and GenCord were used to investigate the impact of genetic variation on expression levels within and across human populations, but also across cell types. Statistical methods were used to associate SNP genotypes with mRNA levels (see sections 2.4 and 2.5), and experimental methods were subsequently employed for the biological verification of a subset of predicted associations (see sections 2.6.5 and 2.8.3).

For expression association studies using the HapMap populations, publicly available genotype data were combined with expression data generated by our group at the Wellcome Trust Sanger Institute (WTSI) for the same set of individuals. This analytical set up made it possible to explore how genetic variation shapes gene expression differences within and across populations, chiefly as a consequence of allele frequency differences. The HapMap Project was launched in 2002 and there have been a number of data release stages over the past few years. As a consequence, analyses performed in this thesis used two different releases of HapMap data: a) Phase 2 data were used to explore the impact of eQTL-nsSNP interactions on gene expression in cis and trans (Chapter 3) and b) Phase 3 data were used to investigate the fine-scale architecture of the cis regulatory landscape (Chapter 4). HapMap genotype data are publicly available at <http://www.hapmap.org> and expression data generated by our group for these populations are available at: <ftp://ftp.sanger.ac.uk/pub/genevar/>.

The GenCord study design involves expression quantification in each of three cell types separately, and genotyping using DNA from a single cell type (LCLs). This

analytical set up made it possible to address how genetic variation shapes gene expression differences within a population and across cell types, as a consequence of the cell type-dependent action of genetic variation (Chapter 5). The chief advantage of GenCord is that it allows direct comparisons to be made across cell types, as samples were collected and processed in a systematic way. GenCord was also used to explore the fine-scale architecture of the cis regulatory landscape in a cell type-specific context (Chapter 5). Expression data are available at <ftp://ftp.sanger.ac.uk/pub/genevar/> and in the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/> accession number GSE17080).

The datasets used in each of the three studies are outlined in Table 3. In the following sections I describe how SNP genotype (section 2.2) and gene expression (section 2.3) information was obtained or generated, I present the general statistical methods used for detection of variants associated with gene expression (sections 2.4 and 2.5) and outline the specific analyses carried out for each study.

		HapMap Phase 2	HapMap Phase 3	GenCord
eQTL- nsSNP interaction	(Chapter 3)	X		
eQTL fine-scale architecture	(Chapter 4)		X	X
eQTL cell type specificity	(Chapter 5)			X

Table 3. Overview of datasets analysed in each of the three studies presented in this thesis. (Note that the eQTL fine-scale architecture study is outlined in Chapter 4 for HapMap, but results using the same strategy are also presented in Chapter 5 for GenCord).

2.2 THE SNPs

HapMap SNP genotypes were generated by the International HapMap Consortium and are publicly available at www.hapmap.org. GenCord SNP genotypes were generated

by our collaborators in the Department of Genetic Medicine and Development, at the University of Geneva Medical School (UGMS).

2.2.1 HapMap Phase 2

Phase 2 of the HapMap involved genotyping of nearly four million SNPs in each of 270 individuals from CEU, CHB, JPT and YRI populations. HapMap version 21 (NCBI Build 35) SNPs were used to interrogate the interaction between functional variants, namely that between cis eQTLs (i.e. tags of regulatory variants) and nsSNPs (protein-coding SNPs). This study is described in Chapter 3.

2.2.1.1 *Cis eQTLs*

Cis eQTLs were identified in a genome-wide association study (GWAS) by Stranger et al (2007) as Phase 2 HapMap SNPs (mapping in a 2 Mb window centred on the expression probe midpoint) that showed a statistically significant association with mRNA levels at the 0.01 permutation threshold (see section 2.6.1).

2.2.1.2 *nsSNPs*

nsSNPs are protein-coding variants that result in a single amino acid substitution in the protein product. The strategy used to select nsSNPs for this study is summarized in Figure 4 and involved the following steps: rsIDs and coordinates for all known nsSNPs were downloaded from Biomart (<http://www.biomart.org/biomart/martview>), dbSNP125 (NCBI Build 36), and HapMart (version 21 NCBI Build 35) (<http://hapmart.hapmap.org/BioMart/martview>). The distribution of nsSNPs in genes was interrogated for three gene collections created using different annotation methods (Brent 2005; Flicek 2007): CCDS, Ensembl and RefSeq genes. CCDS genes are those genes for which structure has been agreed upon by NCBI, Ensembl, and UCSC. Ensembl genes are to a certain extent annotated automatically, whereas Refseq gene

annotation is largely manual. The Refseq collection was chosen as it represents a set of genes with explicitly linked nucleotide and protein sequence and a large enough number of genes to work with. Refseq gene IDs and coordinates were downloaded from the UCSC genome browser (<http://genome.ucsc.edu/>) and genes mapping on chromosomes X and Y, as well as those without coordinate information were removed. nsSNPs of all frequencies were subsequently mapped on Refseq genes using nsSNP and gene coordinates.

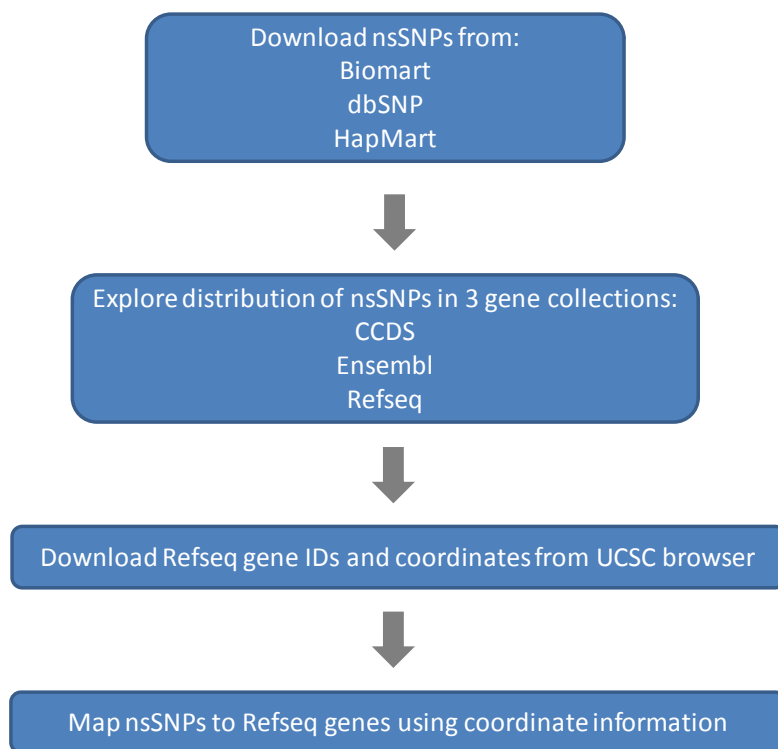


Figure 4. Strategy employed to select nsSNPs for the eQTL-nsSNP interaction study described in Chapter 3.

2.2.2 HapMap Phase 3

In Phase 3 of the HapMap, over 4 million SNPs were genotyped in the initial four populations and 1.5 million SNPs were genotyped in the seven additional populations.

Out of a total of 1,301 individuals genotyped, I used HapMap version 27 (NCBI Build 36) SNPs from 792 individuals from the CEU, CHB, GWK, JPT, LWK, MEX, MKK, and YRI populations to investigate the fine-scale architecture of the regulatory regions around genes, in the study described in Chapter 4.

2.2.3 GenCord

Approximately half a million SNPs were genotyped in the 85 individuals of GenCord. DNA samples were extracted from cord tissue LCLs with the Puregene cell kit (Gentra-Qiagen, Venlo, The Netherlands). Genotyping was performed using the illumina 550K SNP array (illumina, San Diego, California, USA) following the instructions of the manufacturers (Figure 5). This work was carried out by Samuel Deutsch at the UGMS. Principal component analysis (PCA) was performed on the genotype data to detect potential outliers. This analysis was carried out by Stephen Montgomery at the WTSI.

2.3 THE GENES

Transcript levels in HapMap LCLs and in the three cell types of GenCord were quantified using gene expression arrays at the WTSI. All data generated are publicly available at <http://www.sanger.ac.uk/Software/Genevar>. GenCord data are also available on the GEO (section 2.1.3).

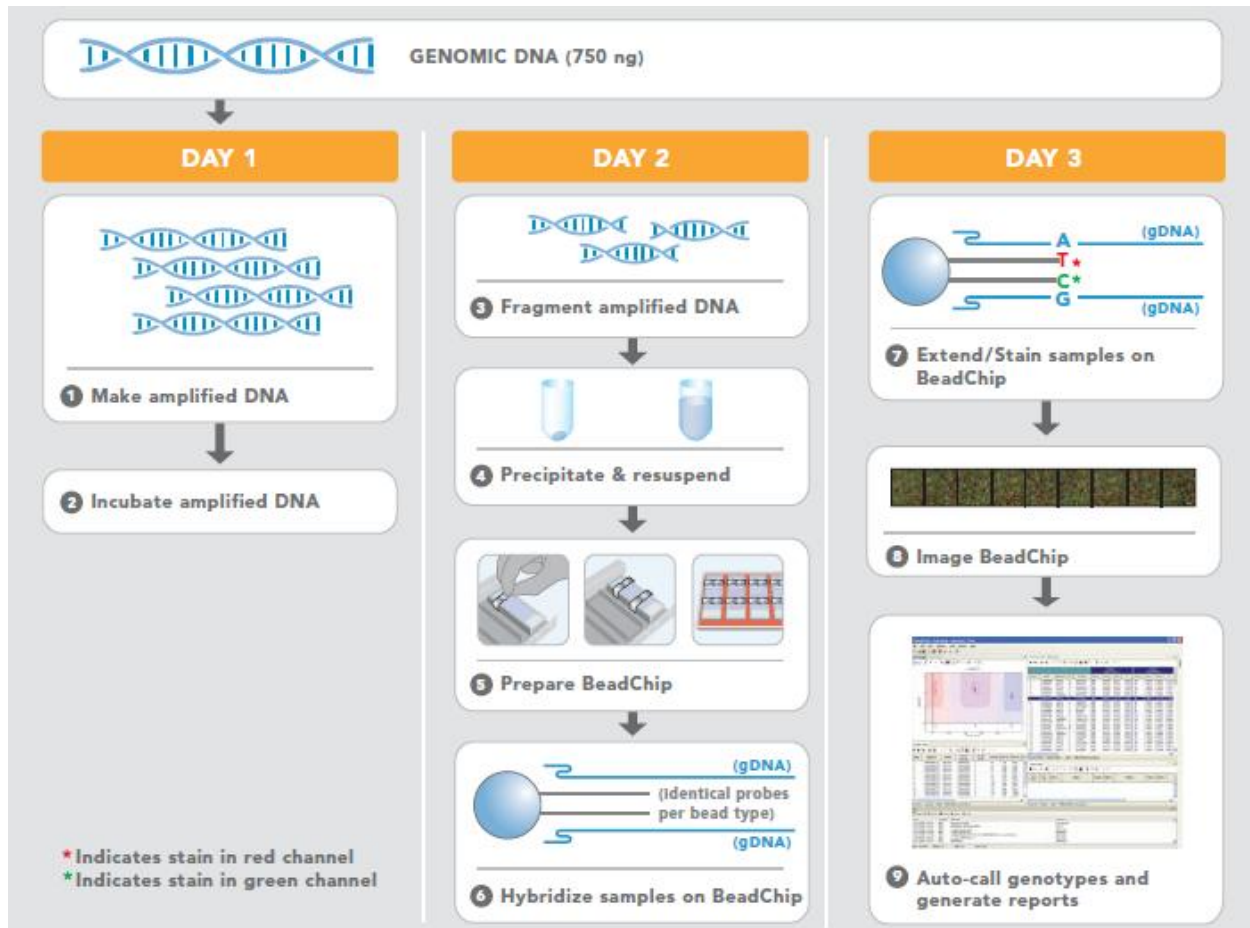


Figure 5. Infinium SNP genotyping assay. The DNA sample used for this assay is isothermally amplified overnight (**Steps 1 and 2**). This amplification has no appreciable allelic partiality. Approximately 750 ng of DNA is used to assay 500,000 SNP loci and the amplified product is fragmented by a controlled enzymatic process (**Step 3**). After alcohol precipitation and resuspension of the DNA (**Step 4**), the BeadChip is prepared for hybridization in the capillary flow-through chamber (**Step 5**); samples are applied to BeadChips and incubated overnight. The amplified and fragmented DNA samples anneal to locus-specific 50-mers (covalently linked to one of over 500,000 bead-types) during the hybridization step (**Step 6**). One bead type corresponds to each allele per SNP locus. After hybridization, allelic specificity is conferred by enzymatic base extension. Products are subsequently fluorescently stained (**Step 7**). The intensities of the beads' fluorescence are detected by the illumina BeadArray Reader (**Step 8**), and are in turn analysed using illumina's software for automated genotype calling (**Step 9**). Figure and assay description from <http://www.illumina.com/>.

2.3.1 HapMap Phase 2

2.3.1.1 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from LCLs of the 210 unrelated individuals of the HapMap Phase 2 (Coriell, Camden, New Jersey, USA). For each RNA extraction, two one-quarter scale Message Amp II reactions (in vitro transcription reactions or IVTs) (Ambion, Austin, Texas, USA) were performed using 200 ng of total RNA, to produce cRNA. To assay transcript levels, 1.5 μ g of the cRNA was hybridized to illumina's commercial whole genome expression array, Sentrix Human-6 v1 Expression BeadChip (Kuhn, Baker et al. 2004). These arrays utilize a bead pool containing ~48,000 unique bead types (one for each of 47,294 transcripts, plus controls), each with several hundred thousand gene-specific 50mer probes attached (Figure 6). Six arrays were run in parallel on a single BeadChip. Each bead type (probe) is present on a single array on average 30 times. Each of the two IVT reactions from the 210 samples was hybridized to two arrays each, so that each cell line had four replicate hybridizations. cRNA was hybridized to arrays, labelled with Cy3-streptavidin (Amersham Biosciences, Little Chalfont, UK) and scanned with a Bead Station (illumina). This work was carried out by Catherine Ingle at the WTSI.

With the illumina bead technology, a single hybridization of RNA from one cell line to an array produced approximately 30 intensity values for each of 47,294 bead types. These background-corrected values for a single bead type were summarized by illumina software and output to the user as a set of 47,294 intensity values for each individual hybridization. In this experiment, each cell line was hybridized to four arrays, resulting in four reported intensity values (as averages of the values from the 30 beads per probe) for each of the 47,294 bead types. To combine information from replicate hybridizations, raw data were read using the *Beadarray R* package (Dunning, Smith et al. 2007) and normalized on a \log_2 scale using a quantile normalization method

(Bolstad, Irizarry et al. 2003) across replicates of a single individual, followed by a median normalization method across individuals of a single population. These normalized values (for each probe, across replicates for each individual) were used in subsequent analyses. Normalization was carried out by Mark Dunning and Simon Tavaré at the Cancer Research UK Cambridge Research Institute (CRI).

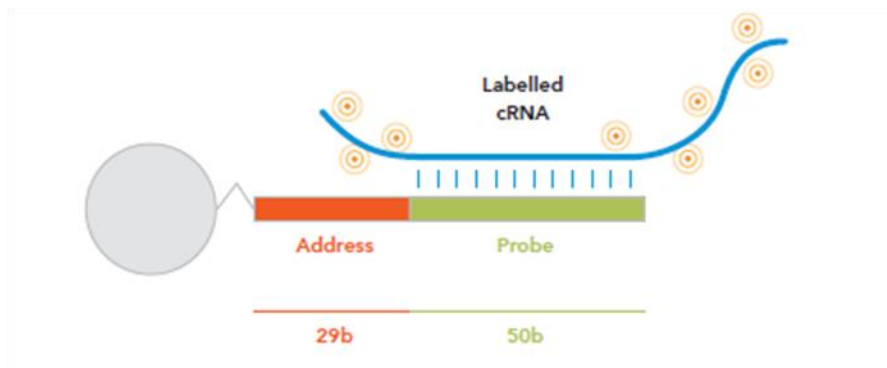


Figure 6. Gene expression probe. Gene expression probes are attached to beads, which are then assembled into arrays. For simplicity, this figure shows only one oligomer attached to the bead; actual beads have hundreds of thousands of copies of the same sequence attached. Figure and description from <http://www.illumina.com/>.

2.3.1.2 Selection of variable probes

To ensure variability in the gene expression phenotype, the intersection of the top 18,000 most variable probes in each of the four populations was selected from the 47,294 probes, resulting in a set of 13,797 probes. An additional set of probes with large differences in rank variability between populations was also selected by ranking all transcripts by variability within each population and making all pairwise comparisons between populations to quantify difference in rank between population pairs. The top 1% of transcripts with largest absolute value rank difference from each population pair comparison were selected. The union of these lists provided an additional 2,021 probes. Probes mapping to chromosomes X and Y, as well as those mapping to the mitochondrion genome were discarded. Probes with no match in the human genome

Build 35 were also removed, as were 469 probes that contained SNPs in their sequence. This resulted in a subset of 14,456 probes (mapping to 13,643 unique autosomal genes) that were highly variable within and between populations and were used for association analysis (Chapter 3). Variable probe selection was carried out by Barbara Stranger and Manolis Dermitzakis at the WTSI.

2.3.2 HapMap Phase 3

2.3.2.1 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from LCLs of the 792 unrelated individuals of the HapMap Phase 3 (Coriell). Gene expression (mRNA levels) was quantified using illumina's commercial whole genome expression array, Sentrix Human-6 Expression BeadChip version 2 (~48,000 transcripts interrogated; illumina) as described previously (in this case only two IVTs were performed). This work was carried out by Catherine Ingle, James Nisbett, and Magdalena Sekowska at the WTSI.

Hybridization intensity values were normalized on a \log_2 scale using a quantile normalization method (Bolstad, Irizarry et al. 2003) across all replicates of a single individual followed by a median normalization method across individuals of a single population. GIH, LWK, MEX and MKK populations were normalized for admixture using a customized version of *Eigenstrat* which outputs principal component adjustments for expression data (Price, Patterson et al. 2006). Expression values were adjusted using ten primary axes of variation from intra-population PCA and these normalized expression values were used as input for the association analysis. Normalization and PCA correction were performed by Stephen Montgomery at the WTSI.

2.3.2.2 Probe selection

illumina's Sentrix Human-6 Expression BeadChip version 2 array covers over 24,000 unique, curated genes from the Refseq collection, as well as genes for which annotation is less well-established. In this case, probes were not filtered for expression variability, but were restricted to those corresponding to Refseq genes. SSAHA (Sequence Search and Alignment by Hashing Algorithm) (Ning, Cox et al. 2001), an algorithm for very fast matching and alignment of DNA sequences, was used to map probes on the Ensembl genes using the Ensembl Application Programme Interface (API) (<http://www.ensembl.org/info/data/api.html> Ensembl 49 NCBI Build 36). It was found that 22,512 probes mapped to 19,862 Ensembl genes and depending on the number of transcripts, some genes were covered by multiple probes. Conversely, a subset of probes mapped to more than one Ensembl gene and were discarded, as were probes mapping on chromosomes X and Y. Following filtering, a non-redundant total of 21,800 probes (corresponding to 18,226 Ensembl genes) was used for association analysis (Chapter 4). Mapping of probes on Ensembl genes using the Ensembl API was carried out with the help of Nathan Johnson at the European Bioinformatics Institute (EBI).

2.3.3 GenCord

2.3.3.1 GenCord sample collection

Umbilical cords were collected from 85 newborns of Western European origin born at the maternity ward of the University of Geneva Hospital, for which pregnancies were full term or near full term (38-41 weeks). For each sample, informed consent was obtained after an interview of the mother with a trained nurse and the project was approved by the University of Geneva Hospital Ethics Committee. From each umbilical cord three cell types were derived: 1) primary fibroblasts, 2) LCLs and 3) phytohemagglutinin (PHA) stimulated primary T-cells. In addition, total buffy coat was

frozen in RPMI medium (Invitrogen, Carlsbad, California, USA) 10% DMSO (Sigma, St. Louis, Missouri, USA), 20% FCS (Invitrogen) for future studies. This work was carried out by Samuel Deutsch at the UGMS.

2.3.3.2 *GenCord cell line preparation*

Cord blood was collected in 50 ml falcon tubes containing 10 ml of anti-coagulants (Sodium citrate and EDTA, Sigma) and kept at 4°C for less than 24 hours prior to treatment. For separation, cord blood was diluted two-fold in PBS (Invitrogen), layered on Ficoll-Paque (GE Healthcare Lifesciences, Chalfont St. Giles, UK) and centrifuged for 30 minutes at 800g. The mononuclear cell layer was removed, washed twice in 40 ml of PBS and re-suspended in 1 ml of RPMI 20% FCS, 1% antibiotics (Amimed, Basel, Switzerland).

For fibroblast preparation, cord tissue was finely cut under sterile conditions in 1 ml DMEM 10% FCS, 1% antibiotics (Amimed), transferred to a T25 flask and cultured upside-down for 12 hours to allow cells to attach to the surface of the flask. Flasks were turned around and left for approximately one week until fibroblast clusters appeared. Fibroblasts were then expanded with standard procedures. For preparation of LCLs, 300 µl of re-suspended cells and 100 µl of EBV were transferred to a 24-well plate well and cultured in an incubator at 37°C, 5 % CO₂. Fresh medium was added and replaced every 2-3 days. Cells were kept in culture for no less than 21 days prior to freezing. For PHA stimulated T-cell preparation, re-suspended mononuclear cells were diluted to a concentration of 1 x 10⁶ cells/ml in RPMI (Invitrogen) with 5 µg/ml of PHA (Sigma), and cultured for five days with 2/3 medium replacement after 2.5 days. A subset of samples was characterized by flow cytometric analysis for expression of CD3, CD25 and CD69 (Becton Dickinson, Franklin Lakes, New Jersey, USA) revealing a homogenous activated T-cell population.

RNA from each cell type was prepared with RNeasy columns with on-column DNase treatment (Qiagen, Venlo, The Netherlands), quantified with NanoDrop (Thermo Scientific, Waltham, Massachusetts, USA) and analyzed with a 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). This work was carried out by Samuel Deutsch at the UGMS.

2.3.3.3 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from fibroblasts, LCLs, and T-cells of the 85 unrelated individuals of the GenCord as described above. Two one-quarter scale Message Amp II reactions (Ambion) were performed for each RNA extraction with 200 ng of total RNA. 1.5 μ g of cRNA was hybridized to illumina's WG-6 v3 Expression BeadChip array to quantify transcript abundance as described previously. In total there were two technical replicates (labelling and hybridization) for each RNA sample. This work was carried out by Catherine Ingle, James Nisbett, and Magdalena Sekowska at the WTSI. Intensity values were \log_2 transformed and normalized independently for each cell type using quantile normalization for sample replicates, and median normalization across all individuals. Each cell type was renormalized using the mean of the medians of each cell type expression values. Normalization was carried out by Stephen Montgomery at the WTSI.

2.3.3.4 Probe selection

The illumina WG-6 v3 Expression BeadChip array covers over 27,000 unique coding transcripts belonging to the Refseq collection. For the majority of these transcripts annotation is well-established, with approximately 7,000 transcripts having provisional annotation. This array also covers non-coding transcripts, as well as experimentally confirmed mRNA sequences aligning to EST clusters. Only probes corresponding to transcripts with good or provisional annotation (Refseq genes) were selected for

association testing. A total of 36,156 probes with Refseq IDs were queried for their corresponding Ensembl gene IDs in Biomart (Ensembl 50, NCBI Build 36). Of these, 23,805 probes had a corresponding Ensembl gene ID and after discarding probes mapping to chromosomes X and Y, as well as those that mapped to more than one Ensembl genes, 22,651 probes (corresponding to 17,945 RefSeq genes and 15,596 Ensembl genes) were used for subsequent analysis (Chapter 5).

2.4 ASSOCIATION TESTS

Additive linear regression (LR) and Spearman rank correlation (SRC) were used to test for association in cis between SNP genotypes and expression levels of genes. For each gene, variants mapping in a 2 Mb window centred on the TSS were tested for association (cis eQTLs used in the eQTL-nsSNP study described in Chapter 3 were identified in a previous study (see 2.2.1.1) that defined cis eQTLs as variants mapping in a 2 Mb window centred on the probe midpoint). This 2 Mb window defines the genomic region tested for cis association with gene expression. Particular tests and analyses conducted for each of the three studies are discussed in the relevant sections of this chapter.

2.4.1 Additive linear regression

A main effects additive LR model was used to test for association between SNP genotype and probe expression levels. The additive effect of a SNP genotype was tested by coding the genotypes at each locus as 0, 1 and 2 corresponding to counts of alphabetically sorted alleles in each genotype (e.g. counting the number of G alleles for a A/G SNP: AA = 0, AG = 1, GG = 2). Normalized \log_2 expression was regressed on SNP genotypes for each gene, and the following additive model was fitted: the genotype X_i

of individual i at the given SNP may be classified as one of three states $X_i = 0, 1$ or 2 . The linear regression fitted was:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

Where Y_i is the normalized \log_2 expression levels of the probe for individual i , $i=1\dots n$ and ε_i are independent normally distributed random variables with mean 0 and constant variance (Stranger, Forrest et al. 2005). The nominal parametric p-value of the test of no association (i.e. $b_1 = 0$), the slope, and r^2 for each SNP-probe pair were reported (Figure 7). LR however is sensitive to outlier effects and for this reason association tests were also carried out using SRC. SRC performs at a level equivalent to LR, detecting 77% - 86% of the associations uncovered by LR (Stranger, Nica et al. 2007).

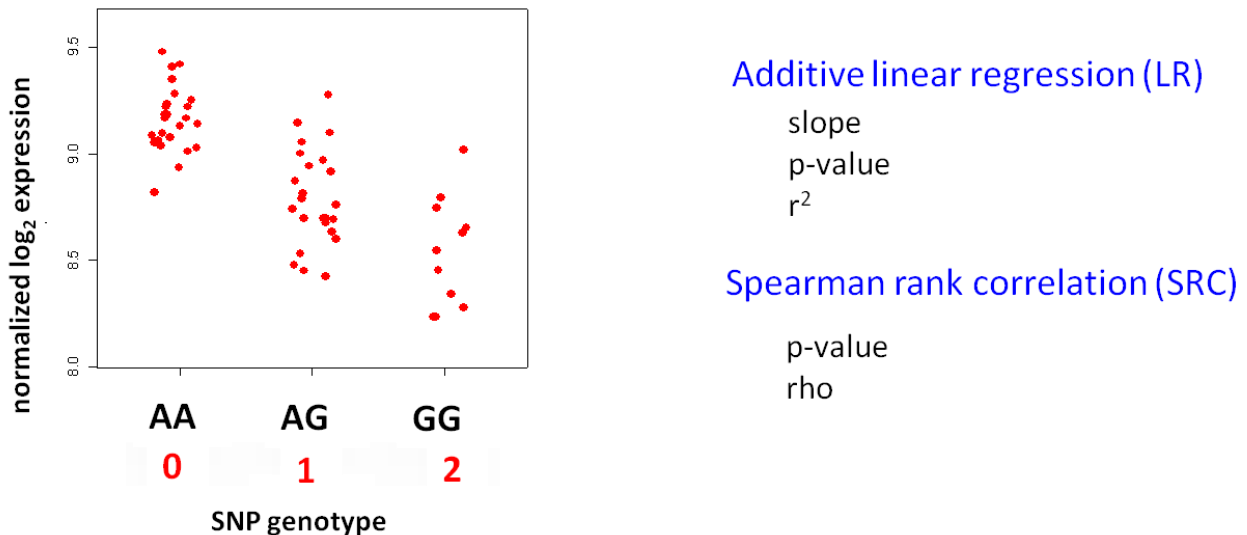


Figure 7. Statistical tests employed to associate SNP genotype with normalized \log_2 mRNA intensity levels. SNP genotypes were coded as 0, 1 or 2 (in this case corresponding to AA, AG and GG respectively). Linear regression (LR) was used to test the additive effects of SNP genotype on mRNA intensity (expression levels). The slope, the p-value and r^2 of the test were reported. To avoid outlier effects that affect LR output, Spearman rank correlation (SRC), a non-parametric test, was also used. The p-value and the correlation coefficient (rho) were reported.

2.4.2 Spearman rank correlation

SRC was also used to test for association between SNP genotypes and probe expression levels. SRC is a non-parametric measure of correlation that assesses how well an arbitrary monotonic function describes the relationship between two rank-ordered variables, without making any other assumptions about the particular nature of the relationship between these variables. Variables are initially converted into ranks (in this case the lower expression values are assigned lower ranks) and a correlation analysis is performed. When two observations are equal (tied) the average rank is used. SRC yields a statement of the degree of interdependence of the scores of the two variables, the Spearman correlation coefficient or rho. Rho describes the strength and direction of the correlation. The nominal p-value for the test of no association and rho were reported.

2.5 MULTIPLE TEST CORRECTION

To assess significance of association between SNP genotype and probe expression levels, 10,000 permutations of each expression phenotype relative to the genotypes were performed for each gene (Churchill and Doerge 1994; Doerge and Churchill 1996; Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007) (Figure 8). For each round of permutations, the minimal permuted p-value was reported and a distribution of 10,000 minimum permuted p-values was generated. An association to gene expression was considered significant if the nominal p-value from the association test (observed p-value) was lower than the 0.5, 0.01, 0.001 and 0.0001 tail of the distribution of the minimal permuted p-values, defining four permutation significance thresholds. For each gene, the most stringent p-value was retained.

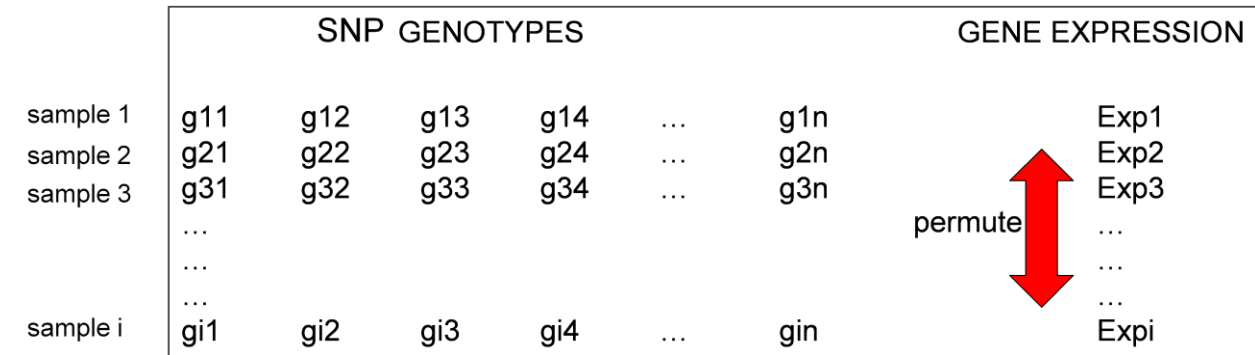


Figure 8. Significance levels for each gene were determined through permutations. 10,000 permutations of expression levels (Exp1, Exp2, Exp3...Expi) relative to genotypes (g11, g21, g31...gi1) were performed for each gene and for all individuals (samples) in the single population analysis. The minimal permuted p-value obtained for each round of permutations was used to generate a distribution of minimal permuted p-values for each gene. Four significance thresholds were defined at the 0.5, 0.01, 0.001 and 0.0001 tails of the distribution. Significant associations of SNP genotype to expression levels were those for which the association test p-value (observed p-value) was lower than the selected permutation significance threshold.

2.6 EQTL-nsSNP INTERACTION STUDY (CHAPTER 3)

2.6.1 The interaction model

HapMap Phase 2 data were used to explore interactions between regulatory and protein-coding variants and their impact on gene expression in cis and trans. The regulatory variants tested for interaction were cis eQTLs (MAF \geq 0.05) identified in a previous study (Stranger, Nica et al. 2007) located within a 2 Mb window centred on the probe midpoint. The protein-coding variants tested were nsSNPs mapping in Refseq genes.

The model of interaction brings together quantitative and qualitative variation as follows: a gene for which a cis eQTL has been detected will be expressed at different quantities among individuals in the population (Pastinen and Hudson 2004; Stranger, Nica et al. 2007) (Figure 9 a). On the other hand, genes containing nsSNPs give rise to

protein products that differ in quality by a single amino acid (Figure 9 b). In the case where a gene with an identified cis eQTL also contains an nsSNP, the resulting protein products will differ not only in quantity, but also in quality (amino acid sequence) among individuals (Figure 9 c, also see Figure 14). The co-existence of these two variant types may have cis and trans effects on gene expression. In cis, the eQTL (or rather the regulatory element tagged by the eQTL) can modify (magnify or mask) the functional effect of the nsSNP. This is a cis modification effect and nsSNPs harboured in genes with varying expression levels are hereon termed DE (differentially expressed). In trans, the different protein ratios arising from modification in cis may affect their downstream targets, leaving an imprint on genome-wide expression levels. This impact in trans is a true epistatic effect and can be explored using the specific and testable biological model presented.

The proposed model is centred on the concept of DE nsSNPs, and two strategies were employed to detect these variants. The first strategy involved scanning all genes with cis eQTLs for nsSNPs. The second strategy involved direct association testing of nsSNP genotype with expression levels of the gene it is harboured in. In this second case, the nsSNP can act as an eQTL for its own gene's expression levels. To summarize, an nsSNP is DE if: 1) it maps in a gene for which at least one cis eQTL has been identified or 2) it shows a significant association with its own gene's expression levels. The nsSNP shown in Figure 9 c is DE.

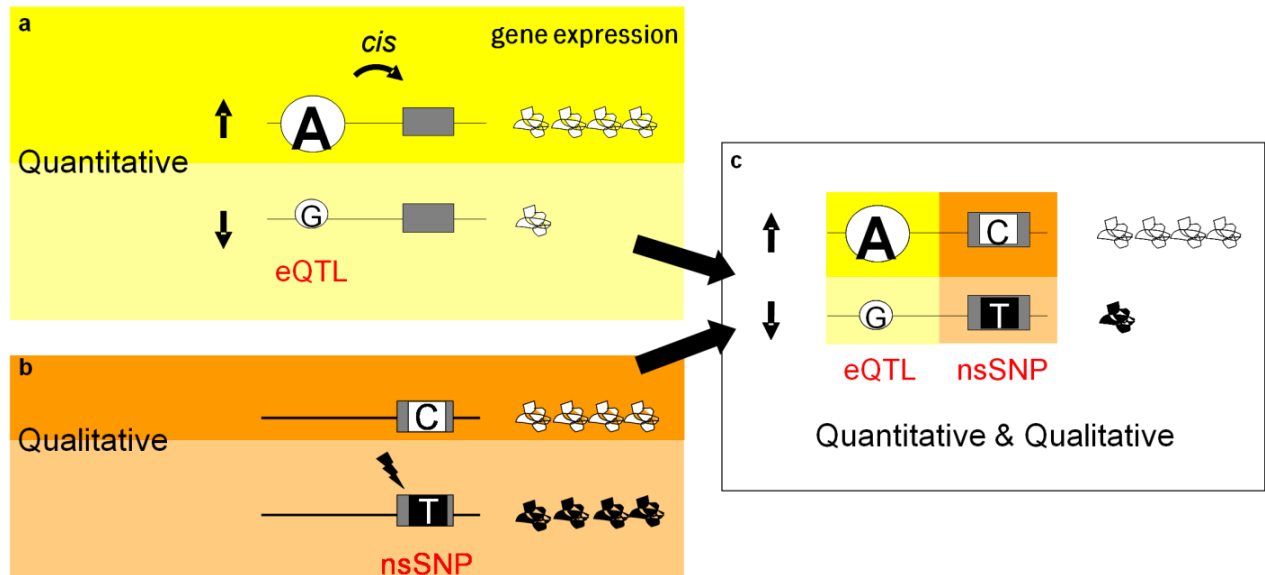


Figure 9. eQTL-nsSNP interaction. **a)** Genes with identified cis eQTLs have quantitative differences in their expression (high vs. low expression levels). **b)** Genes that possess an nsSNP give rise to protein products that differ qualitatively by a single amino acid (white vs. black protein product). **c)** If a gene possesses both a cis eQTL and an nsSNP, the resulting protein products will differ in quantity and quality. This is an example of an interaction in cis, where the functional effect of the nsSNP is modified by the cis eQTL. Furthermore, if this gene has downstream targets, their expression may be influenced through a trans effect on gene expression. See also Figure 14 in Chapter 3.

2.6.2 Single population nsSNP association test

One way to determine whether an nsSNP is DE is to perform a direct association test of nsSNP genotype and expression levels of the gene harbouring it. LR was used to test for association in cis between: 1) nsSNP genotypes (for nsSNPs with MAF ≥ 0.05) for the unrelated individuals of each HapMap Phase 2 population (60 CEU, 45 CHB, 45 JPT, and 60 YRI) and 2) normalized \log_2 quantitative gene expression measurements for the same individuals. Association testing was performed for each population separately and significance thresholds for each gene were assigned through permutations of expression values relative to genotypes. An association with gene expression was

considered significant if the nominal p-value from LR was lower than the 0.01 tail of the distribution of minimal permuted p-values. Correction for false positives was carried out by calculating the ratio of expected false positives at a given threshold over the number of significant associations at the same threshold (this is an approximation of the false discovery rate (FDR)).

2.6.3 Multiple population nsSNP association test

To increase power of association detection I combined data (SNP genotypes and normalized expression values) for unrelated individuals of multiple populations and repeated association testing. Three different multiple population comparison panels were compiled: 1) CEU-CHB-JPT-YRI, 2) CEU-CHB-JPT and 3) CHB-JPT. Association tests were carried out for each population panel separately using LR. In this case, correction for significance was through conditional permutations (Figure 10) whereby the correlated structure of gene expression values within each population was retained by randomizing data within each population (Stranger, Nica et al. 2007). This approach accounts for population differentiation and prevents detection of spurious associations. For each of the 14,456 probes in each multiple population panel, expression values were permuted among individuals of a single population followed by regression analysis of the grouped multi-population expression data against the grouped multi-population permuted nsSNP genotypes. Four significance thresholds were selected (0.05, 0.01, 0.001, 0.0001) and an association to gene expression was considered significant if the nominal p-value from the linear regression test was lower than the 0.01 tail of the distribution of minimal permuted p-values. Correction for false positives was carried out as described in section 2.6.2.

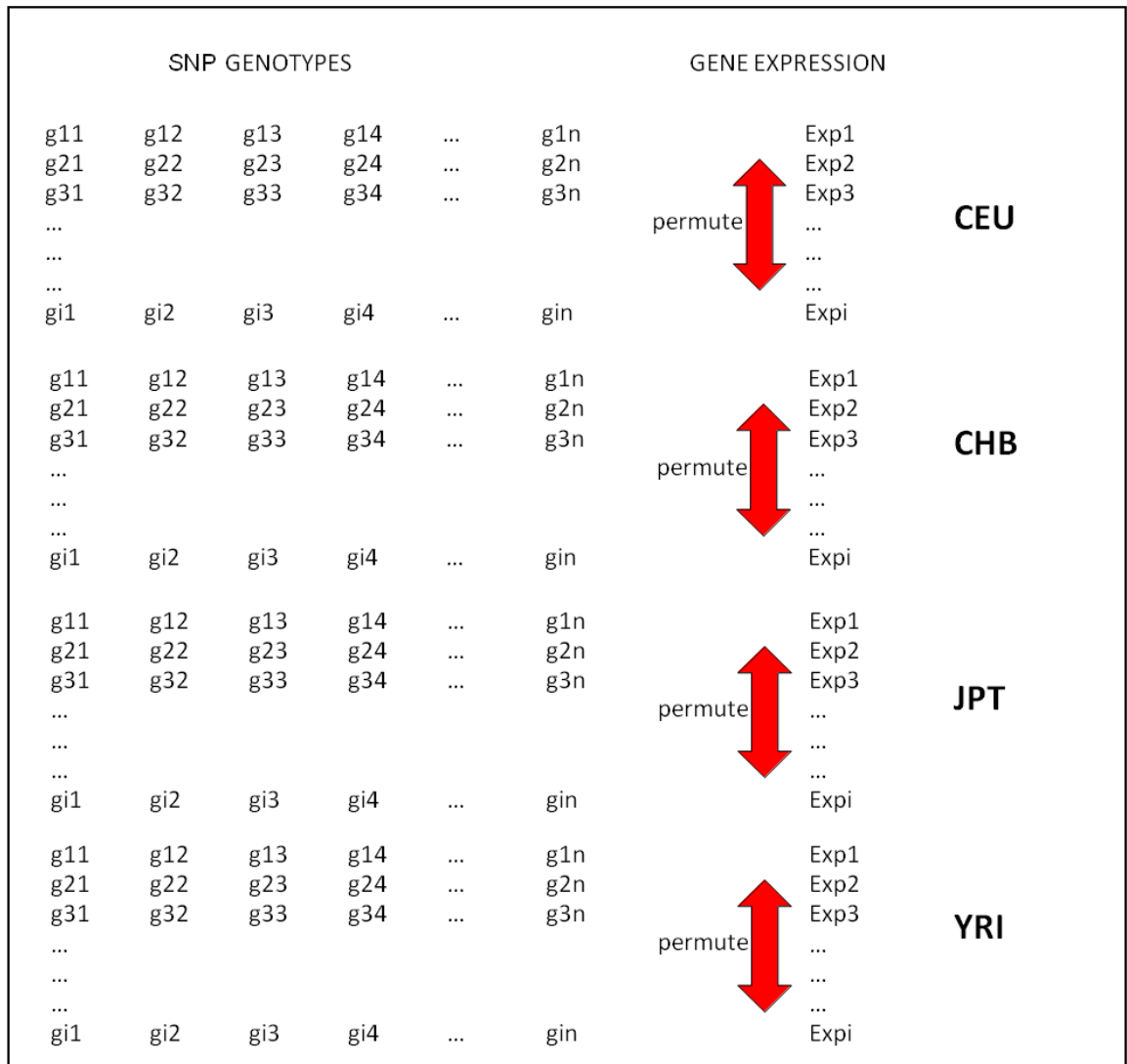


Figure 10. Conditional permutations were used to determine significance levels for each gene in the multiple population analysis. Ten thousand permutations of expression levels (Exp1, Exp2, Exp3...Exp_i) relative to genotypes (g₁₁, g₂₁, g₃₁...g_{i1}) were performed for each gene, in an approach where the correlated structure of expression values within each population was retained. This was achieved by randomizing expression data within each population. This approach accounts for population differentiation and prevents detection of spurious associations.

2.6.4 eQTL-nsSNP linkage disequilibrium analysis

Differential expression of nsSNPs is most likely driven by regulatory variants tagged by eQTLs that are in LD with the nsSNP. To address this, I explored the distribution of r^2 (a

measure of statistical correlation between alleles at two loci) (Hartl and Clark 2007) for eQTL-nsSNP pairs in which the nsSNP: a) showed a significant association with its own gene's expression levels and b) showed no such association. LD values were calculated by a pairwise estimation, for eQTLs and nsSNPs genotyped in the same individuals, and that mapped within a 100 kb window of each other (Ensembl 46). LD values were calculated by Daniel Rios at the EBI. The distributions of r^2 estimates for eQTLs-nsSNP pairs with and without an associated nsSNP were compared using a Mann-Whitney (M-W) test. Significant results were those for which M-W p-value ≤ 0.05 .

2.6.5 Allele-specific expression assay

2.6.5.1 DNA and RNA preparation for allele-specific expression assays

The association tests employed make predictions about DE nsSNPs. ASE assays were used for the biological verification of a subset of these predictions (Figure 11). Genomic DNA (gDNA) and total RNA were extracted from LCLs of the unrelated CEU and YRI HapMap individuals (Coriell) using Qiagen's AllPrep kit. RNA was treated with Turbo DNA-free (Ambion) to minimize gDNA contamination. The RNA was concentrated and further cleaned with RNeasy MinElute columns (Qiagen). Total RNA and gDNA were quantified using a Nanodrop Spectrophotometer (Thermo Scientific) and either QuantiT RNA or DNA reagents (Invitrogen). Double stranded (ds) cDNA was synthesised from 250 ng of cleaned RNA. The first strand was synthesised with Superscript III (Invitrogen) and random hexamers. The second strand was synthesised with DNA polymerase I (Invitrogen), ribonuclease H (Invitrogen) and dNTPs. The 96-well plate containing the ds cDNA samples was cleaned using Multiscreen PCR plate (Millipore). This work was carried out by Matthew Forrest at the WTSI.

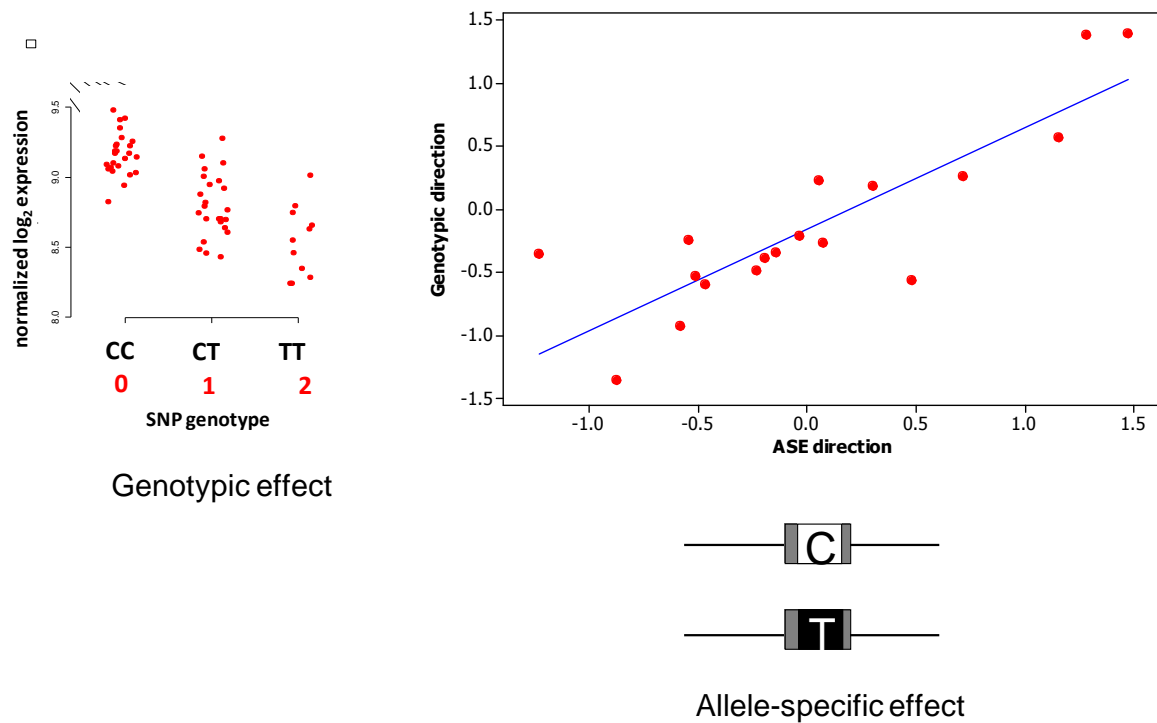


Figure 11. Statistically predicted vs. biologically verified differentially expressed (DE) nsSNPs. Allele-specific expression (ASE) assays were used to verify differential expression of nsSNPs (cis eQTLs) predicted from statistical analyses. In this scatterplot the experimentally determined allelic effect (x axis) of a C/T nsSNP is compared to the genotypic effect predicted from the association test (y axis).

2.6.5.2 *illumina allele-specific expression array*

A custom made Oligo Pool All (OPA) array (illumina) based on the Golden Gate assay was used to assay ASE. Only exonic SNPs ≥ 45 bp from both exon edges were chosen for submission to illumina for assay design, to ensure that the assay would work equally well for genomic and cDNA. SNPs that failed according to illumina's design scores were discarded. Paired ds cDNA and gDNA were dried down in 96-well plates and re-suspended in 5 μ l of HPLC purified water. Golden Gate assays were then run for all samples using the manufacturer's standard protocol for gDNA (i.e. ds cDNA was treated exactly the same way as gDNA). Reactions were hybridised to 8 \times 12 Sentrix

Array Matrix (SAM) Universal Probe Sets so that 96 arrays could be run in parallel. Each bead type (probe) is present on a single array on average 30 times. All reactions were run in duplicate, so that each LCL had two ds cDNA replicate and two gDNA replicate hybridizations. SAMs were scanned with a Bead Station (illumina). A total of 1,536 assays were interrogated on the array, but only 141 were nsSNPs from this study and only 28 were selected based on data quality for further analysis. This work was carried out by Matthew Forrest at the WTSI.

2.6.5.3 *Allele-specific expression assay data pre-processing*

Data from each array were summarised by calculating the per bead type average of 4 quantities after outlier removal: the $\log_2(\text{Cy3})$ and $\log_2(\text{Cy5})$ intensities, average log-intensities ($1/2\log_2(\text{Cy5}.\text{Cy3})$) and log-ratios ($\log_2(\text{Cy5}/\text{Cy3})$). Outliers were beads with values more than three absolute deviations from the median. Arrays with low dynamic range (determined using an inter-quartile range cut-off of < 1 for either the $\log_2(\text{Cy3})$ or $\log_2(\text{Cy5})$ summary intensities) were discarded. The summarised data were normalized by median centring of log-ratios. Normalisation was carried out in R using the *Beadarray* package (Dunning, Smith et al. 2007) by Matthew Ritchie at the CRI. Direction of expression (high/low) was assigned to alleles for nsSNPs fulfilling the threshold criteria from the association study (adjusted $r^2 \geq 0.27$; i.e. the nsSNP explained at least 27% of the variance in gene expression so the effect is expected to be large) and the ASE assay (average cDNA log-intensity ≥ 12 within a population).

2.6.6 *Amino acid substitution effect*

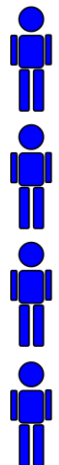
Given that nsSNPs are likely to be functional I explored three aspects of the resulting amino acid substitution: a) relative position of substitution on the peptide, as a percent of peptide total length. b) hydrophobicity change in peptide resulting from the amino acid substitution. For each pair of variant sequences the hydrophobicity at the position

of the variant amino acid was calculated using the Kyte-Doolittle algorithm (Kyte and Doolittle 1982) and a window size of seven amino acids (centred on the variant amino acid). The difference between hydrophobicity scores was then taken for each of the variant pairs in the dataset. c) Pfam score change in peptide sequence resulting from the amino acid substitution (Finn, Tate et al. 2008). All sequences were searched against the profile-HMM library provided by the Pfam database (release 22.0) using hmmpfam from the HMMer software package (version 2.3.2, <http://hmmer.janelia.org/>) and a default cut off E-value of 10. Only the HMM_ls library was used so that domain assignments to a pair of variant sequences were comparable. The set of Pfam domain assignments were then filtered such that only the domains that overlapped with the SNP position and that at least one of the domain assignments from a pair of variant sequences scored above the Pfam defined gathering threshold, were considered in the subsequent analysis. The difference between the two E-values was taken for each of the variant pairs in the dataset. Pfam scores were provided by Robert Finn from the Pfam team at the WTSL.

2.6.7 Impact of eQTL-nsSNP interaction in trans

The impact of interactions between eQTLs and nsSNPs on gene expression in trans was tested for the CEU population. In a previous study (Stranger, Forrest et al. 2005), trans effects were found to be weak in the YRI population and the number of individuals in the CHB and JPT Phase 2 populations limit the power for detection of trans effects. To test the trans effect of eQTL-nsSNP interactions I pooled the minor allele homozygote and the heterozygote into a single genotypic category and then coded genotypes as 0 (major allele homozygote) or 1 (heterozygote and minor allele homozygote) for both eQTL and nsSNP. As a result, four possible eQTL-nsSNP genotypic combinations are possible: 0-0, 1-0, 0-1, 1-1 (Figure 12). Analysis of variance (ANOVA) was performed

using the R software package (R Development Core Team 2008) to test the effects of: the eQTL, the nsSNP, and the eQTL × nsSNP interaction term against gene expression phenotypes in trans. In each case the gene from which the eQTL-nsSNP pair originated was excluded from the association test. To ensure all genotypic combinations were present and to avoid outlier effects, tests were carried out for 22 SNP pairs with low LD ($D' \leq 0.5$) between eQTL and nsSNP and a MAF ≥ 0.1 for both variants.



	eQTL		nsSNP		genotype
	CC	0	GG	0	0-0
	CT	1	GG	0	1-0
	CC	0	AG	1	0-1
	TT	1	AA	1	1-1

Figure 12. Genotypes for eQTL and nsSNP were combined and used to test for impact of the eQTL-nsSNP interaction on gene expression in trans. The minor allele homozygote and the heterozygote were collapsed into a single genotypic category (red circles represent genotypes that were pooled for each variant). This resulted in two genotypic categories coded as 0 (major allele homozygote) or 1 (heterozygote and minor allele homozygote) for both eQTL and nsSNP. As a result, when combining eQTL-nsSNP genotypes four combinations (shown in the right column) are possible: 0-0, 1-0, 0-1, 1-1.

To assess significance of interaction p-values a single permuted dataset of expression values relative to combined genotypes was generated and the p-value distributions of interaction terms for observed and permuted data were compared. To further evaluate the robustness of observed interactions, I permuted eQTL genotypes

relative to nsSNP genotypes and gene expression phenotypes, and re-ran the ANOVA association test for the top ten most significant interactions.

2.7 eQTL FINE-SCALE ARCHITECTURE STUDY (CHAPTER 4)

2.7.1 Recombination hotspot interval mapping and LD filtering

LD is a useful property of the genome as it enables genome-wide mapping of variation associated with a phenotype. At a smaller scale however LD impedes fine-mapping as multiple correlated variants can show a significant association with a trait. The aim of this study was to identify those cis eQTLs that tag the effects of independent regulatory elements and in this way detect independent cis regulatory signals for a gene. To do this I mapped eQTLs in recombination hotspots and recombination hotspot intervals using data on the recombination patterns in the genome (McVean, Myers et al. 2004; Myers, Bottolo et al. 2005; Winckler, Myers et al. 2005).

A recombination hotspot interval was defined as the space between two recombination hotspots and represents a segment of DNA with an independent recombination history (McVean, Myers et al. 2004). Recombination hotspot intervals were constructed using hotspot coordinates estimated from HapMap Phase 2 data and coordinates were lifted over to Build 36 using the UCSC liftOver tool (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>). eQTLs were mapped in intervals and only the most significant eQTL per interval was considered for further analysis. Correlation between this subset of eQTLs is still possible if LD extends across intervals and to ensure that independent signals were identified, the least significant variant from eQTL pairs with $D' \geq 0.5$ for a given gene was removed. Independent eQTLs (or regulatory intervals) therefore define genomic units likely to carry independent functional regulatory elements. The strategy described is shown in Figure 13.

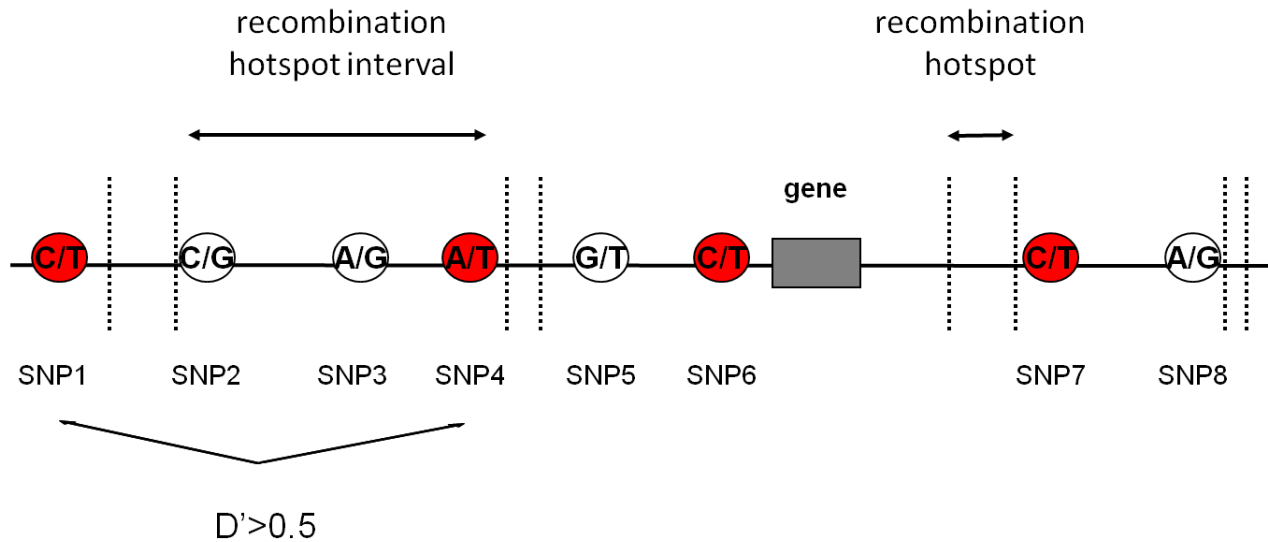


Figure 13. Detecting independent eQTLs (intervals). Recombination hotspot intervals were defined as the space between two consecutive recombination hotspots (McVean, Myers et al. 2004) and represent an approximation of genomic units with an independent history. To identify intervals with an independent effect on gene expression in cis, eQTLs were mapped in recombination hotspot intervals and the most significant eQTL for a given interval (shown in red) was retained. Further control for correlation was performed by excluding the least significant eQTL from eQTL pairs with a $D' > 0.5$ (e.g. SNP1-SNP4, with SNP4 being the most significant of the two). In this example SNP4, SNP6 and SNP7 are independent cis eQTLs defining independent intervals. A modified version of this strategy was used to test the effects of interactions between SNPs that have an impact on gene expression in cis: SNPs with a nominal (uncorrected) p-value < 0.001 were mapped in intervals and SNP pairs with a $D' > 0.5$ were excluded from association testing. In this example an interaction effect would be tested for the following pairs: SNP4-SNP6, SNP6-SNP7, and SNP4-SNP7 (see section 2.7.3).

D' was chosen over r^2 to filter for LD as it is a metric of the degree of historical recombination that has occurred between two variants (Hartl and Clark 2007). r^2 on the other hand is an indicator of statistical correlation and not of historical relationships. As a result, if two SNPs have different MAFs, but there has been no historical recombination between them in the samples studied, r^2 can take low values, but $D' = 1$. Under such a scenario, if one SNP displays a strong association with expression levels of a gene, it can be that the other SNP is also associated with expression levels of the same gene, even if r^2 is low. This is possible as the two SNPs may be tagging the same

functional variant since there has been no historical recombination between them. Using a D' threshold (which translates to an even lower r^2) ensures that the two signals are historically independent. Upon recombination hotspot interval mapping and LD filtering, I determined the number of independent intervals detected for each gene at the 0.01 and 0.001 permutation thresholds. This analysis was carried out for HapMap Phase 3 (Chapter 4) and GenCord data (Chapter 5). For GenCord an overlap analysis was carried out to determine the extent to which independent eQTLs (intervals) are shared across the three cell types studied.

2.7.2 Independent eQTL distance to transcription start site

To describe the cis regulatory landscape around genes, p-values and effect sizes (ρ) of the most significant eQTL per gene were plotted relative to the TSS. This was done for both HapMap Phase 3 and GenCord data. HapMap Phase 3 data were analysed by Barbara Stranger at the WTSI.

2.7.3 eQTL-eQTL cis interaction

To further characterise cis regulatory architecture, HapMap Phase 3 SNP pairs were tested for an interaction with an impact on gene expression using the CEU and YRI populations. The interaction model employed in this analysis was identical to that described in section 2.6.7, but instead of testing interactions between regulatory (eQTLs) and protein-coding (nsSNPs) variation, I explored interactions between SNPs likely to tag regulatory variants. Variants tested were not filtered for permutation threshold significance (and are not termed eQTLs), but were SNPs with an observed nominal p-value < 0.001 from the SRC cis association test. These variants were chosen so that SNPs that do not necessarily have large marginal effects are included in the interaction test. (Ideally all SNPs for a given gene should be tested for an interaction to uncover variants

that influence gene expression through interactions. At the time of writing, this was being explored in collaboration with Doug Speed and Simon Tavaré at the CRI). SNPs were mapped in recombination hotspot intervals and the most significant SNP per interval was kept. SNP pairs were constructed for each gene in cis, and pairs with $D' \geq 0.5$ were excluded. ANOVA was used to test the main effects of each SNP, as well as the SNP x SNP interaction term, on gene expression in cis. A single permutation of expression values relative to genotypes was performed to assess significance of the interaction p-values. The p-value distributions of the interaction term for observed and permuted data were compared.

2.8 EQTL CELL TYPE SPECIFICITY STUDY (CHAPTER 5)

2.8.1 Association analysis

GenCord data were used to investigate the cell type specificity of cis eQTLs. A total of 22,651 probes covering 17,945 autosomal RefSeq genes (15,596 Ensembl genes) were tested for cis association with SNP genotypes using SRC. Cis association tests encompassed SNPs mapping in a 2 Mb window centred on the TSS. Following quality control and filtering for $MAF \geq 5\%$, a total of 394,651 SNPs were included in the analysis. Significance thresholds for each gene were assigned after 10,000 permutations of expression values relative to genotypes. To explore sharing and cell type specificity of significant associations, I compared eQTLs and genes across cell types and determined those that passed significance thresholds in all three, in at least two and in only one cell type (overlap analysis).

2.8.2 Repeated-measures ANOVA to investigate eQTL cell type specificity

I used repeated-measures ANOVA (RMA), programmed in R (R Development Core Team 2008), to investigate the robustness of sharing and cell type specificity of associations. I tested shared and cell type-specific SNP-probe pairs identified from the overlap analysis (at the 0.001 permutation threshold), in tests where the repeated measure was the cell type. The analysis was run for pairs of cell types and the significance of the SNP x cell type interaction term was assessed. The expectation is that the interaction term will be significant for those eQTLs that were identified as cell type-specific.

2.8.3 Allele-specific expression assay

ASE assays were used to validate a subset of cell type-specific eQTLs. Thirty five transcript SNPs (seven in fibroblasts, 14 in LCLs, and 14 in T-cells) in genes with identified cell type-specific eQTLs were tested for ASE in each cell type. The expectation is that allelic imbalance will be observed for the cell type in which the eQTL was detected. 800 ng of total RNA, in a total volume of 20 μ l from fibroblasts, LCLs, and T-cells was converted to cDNA using hexaprimers (Superscript II, Invitrogen). This work was carried out by Christelle Borel at the UGMS. gDNA (~40 ng) from LCLs and cDNA (~30 ng) from each cell type, as well as an RNA control (~30 ng) from 293 T-cells were genotyped using Sequenom's MassArray allele specific assay without competitor (iPLEX Gold assay, Sequenom, San Diego, California, USA). Assays for all SNPs were designed using the eXTEND suite and MassARRAY Assay Design software version 3.1 (Sequenom). Amplification was performed in a total volume of 5 μ L containing the DNA, 100 nM of each PCR primer, 500 nM of each dNTP, 1.25 x PCR buffer (Qiagen), 1.625 mM MgCl₂ and 0.2 U HotStar Taq (Qiagen). Reactions were heated to 95°C for 15 minutes followed by 45 cycles at 94°C for 20 s, 56°C for 30 s, 72°C for 60 s and a final

extension at 72°C for 3 minutes. Unincorporated dNTPs were SAP digested prior to iPLEX Gold allele specific extension with mass modified ddNTPs using an iPLEX Gold reagent kit (Sequenom). SAP digestion and extension were performed according to the manufacturer's instructions with reaction extension primer concentrations adjusted to between 0.731-2.193 μM , dependent upon primer mass. Extension products were desalted and dispensed onto a SpectroCHIP using a MassARRAY Nanodispenser (Sequenom) prior to analysis with a MassARRAY Analyzer Compact mass spectrometer (Sequenom). Allele-specific peak heights from the mass spectra of gDNA and cDNA were analysed to detect transcript SNPs showing allelic imbalance. This work was carried out by Naomi Hammond at the WTSI. The ratio of the two alleles of transcript SNPs was analysed in RNA samples of individuals who were double heterozygotes for both the eQTL and the transcript SNP.

2.8.4 Biological properties of cell type-specific associations

Gene Ontology (GO) terms (Ashburner, Ball et al. 2000) were used to investigate the biological properties of cell type-specific gene associations (at the 0.001 permutation threshold). GO terms were assigned to Ensembl Genes (Ensembl 50) and were then mapped on to their GO Slim ontologies. GO Slim represents a cut-down version of GO and gives a broader overview of the ontology (Ashburner, Ball et al. 2000). Fisher's exact tests were used to compare GO Slim terms corresponding to gene associations that were cell type-specific vs. associations that were shared in all three cell types. Significant associations were those for which Fisher's exact p-value ≤ 0.05 .

2.8.5 Tissue entropy

Gene expression entropy was used as a proxy to gene expression specificity (Jongeneel, Delorenzi et al. 2005; Schug, Schuller et al. 2005; Martinez and Reyes-Valdes 2008). The

expression of genes possessing an eQTL in a single cell type (0.001 permutation threshold) was investigated using the GNF/Novartis expression atlas, which contains expression data for 10,424 Ensembl genes from 38 tissues (Su, Wiltshire et al. 2004). The GNF/Novartis data were used to calculate gene expression entropy as described in Schug et al (2005). Briefly, for expression levels measured in N tissues, the relative expression of a gene g in a tissue t is defined as:

$$p_{t|g} = W_{g,t} / \sum_{1 \leq t \leq N} w_{g,t}$$

where $w_{g,t}$ is the expression level of the gene in that tissue. The entropy of a gene's expression across N tissues is defined as follows:

$$H_g = \sum_{1 \leq t \leq N} -p_{t|g} \log_2(p_{t|g})$$

To assess cell type specificity of associations, I compared the entropy distributions for genes with associations that were cell type-specific vs. associations that were: a) three cell type-shared, b) at least two cell type-shared, c) two cell type union. Significant differences in entropy distributions were those for which M-W p-value ≤ 0.05 .