2 Materials and methods

2.1 Resources

The data presented and analysed in this thesis has been derived from samples belonging to three major resources, briefly outlined below: the HapMap, MuTHER and GenCord projects. Table 2.1 summarizes the number of available samples, SNPs and transcripts per resource and the respective thesis chapters where they have been analyzed.

Chapter	Resource	Samples by Tissue	SNPs	Mapped Probes	Mapped Genes
3	HapMap 3 (CEU)	109 LCL	1,186,075	21,800	18,226
4	MuTHER	156 LCL, 160 SKIN, 166 FAT	865,544	27,499	18,170
5	GenCord	75 LCL, 75 fibroblasts, 75 T-cells	1,428,314	26,651	17,945

Table 2.1. Summary of resources (samples, SNPs and transcripts) used throughout the thesis.

2.1.1 HapMap

The International HapMap project is a large-scale collaboration launched in 2002 to identify and catalogue common human genetic variation (Consortium 2003). DNA from LCLs derived from individuals of different population ancestry has been genotyped in an attempt to discover the vast majority of common human SNPs (MAF ≥ 5%). HapMap 3, the current and largest phase of the project (http://hapmap.ncbi.nlm.nih.gov/) is comprised of over 4 million SNPs genotyped from individuals of the Phase 1 and 2 populations (180 CEU, 90 CHB, 91 JPT, 180 YRI) and approximately 1.5 million SNPs genotyped in 760 individuals of seven new populations (90 ASW, 100 CHD, 100 GIH, 100 LWK, 90 MEX, 180 MKK, 100 TSI).

In this thesis, I analysed data from the subset of unrelated HapMap 3 CEU individuals (N=109) in the study described in Chapter 3.

2.1.2 MuTHER

The MuTHER (Multiple Tissue Human Expression Resource) project was funded by the Wellcome Trust in 2007 as a coordinated program of analysis aiming to enhance our knowledge about common trait susceptibility. By generating detailed genetic (genotyping and resequencing) and genomic (mRNA expression, methylation status) information from a range of tissues collected from ~1000 twins, the MuTHER project will constitute a major resource for understanding the relationships between sequence variation and disease phenotypes (http://www.muther.ac.uk/).

LCLs, fresh lymphocytes, fat, muscle and skin biopsies have been obtained from a maximum of 855 twins (318 monozygotic, 537 dizygotic) from the well-characterised Twins UK Resource (Spector and Williams 2006). This sample of volunteers was recruited by media campaigns without selecting for particular diseases or traits. All twins received a series of detailed disease and environmental questionnaires and the majority of individuals have been clinically assessed at several time points for hundreds of phenotypes related to common diseases or intermediate traits. All individuals recruited in this study were Caucasian female twins aged between 39 and 70 years old.

At the time of writing, whole-genome genotyping and expression profiling of the full set of 855 twins was underway. A sample subset representing the pilot phase of the MuTHER project had been profiled in advance in three tissues: LCL, skin and fat. Skin punch biopsies (N=196) were taken from a relatively photo-protected area adjacent and inferior to the umbilicus. The fat sample was then carefully dissected from the same skin biopsy incision. A peripheral blood sample to generate lymphoblastoid cell lines (LCL) was taken contemporaneously. The biopsies were performed by Daniel Glass at KCL following the technique steps described in Appendix 1.

Chapter 4 describes the analysis I performed on the MuTHER pilot project data.

2.1.3 GenCord

The GenCord project was initiated at the University of Geneva Hospital and consists of a collection of cell lines derived from the umbilical cords of 85 individuals of Western European origin. The primary goal of the project was to serve as a resource facilitating discovery and comparison of eQTLs across multiple tissues while controlling for confounding factors such as different population samples or differences in technological and statistical methods employed. Umbilical cord was chosen due to its accessibility and the potential of harvesting multiple tissues from the same sample. Following appropriate consent and ethical approval (Dimas, Deutsch et al. 2009), cord blood and cord tissue was obtained per each sample in order to derive three cell-types: primary fibroblasts, EBV-immortalized lymphoblastoid cell lines (LCL) and primary T-cells. All pregnancies were full or near full term (38-41 week) ensuring age homogeneity of the samples.

GenCord LCL data was used in the control experiment I describe in Chapter 3. GenCord data from LCLs, fibroblasts and T-cells was used in the analysis I present in Chapter 5.

2.2 SNP genotyping

Genetic variation data (SNP genotypes) from HapMap 3, MuTHER and GenCord has been analysed throughout the course of my PhD, primarily to identify associations with gene expression variation (eQTL discovery, section 2.4).

SNP detection has been performed mostly on Illumina’s whole-genome genotyping platforms using the Infinium HD technology. This enables dense, uniform genome coverage by typing a representative set of tag SNPs. The Infinium II assay workflow is described in Figure 2.1.

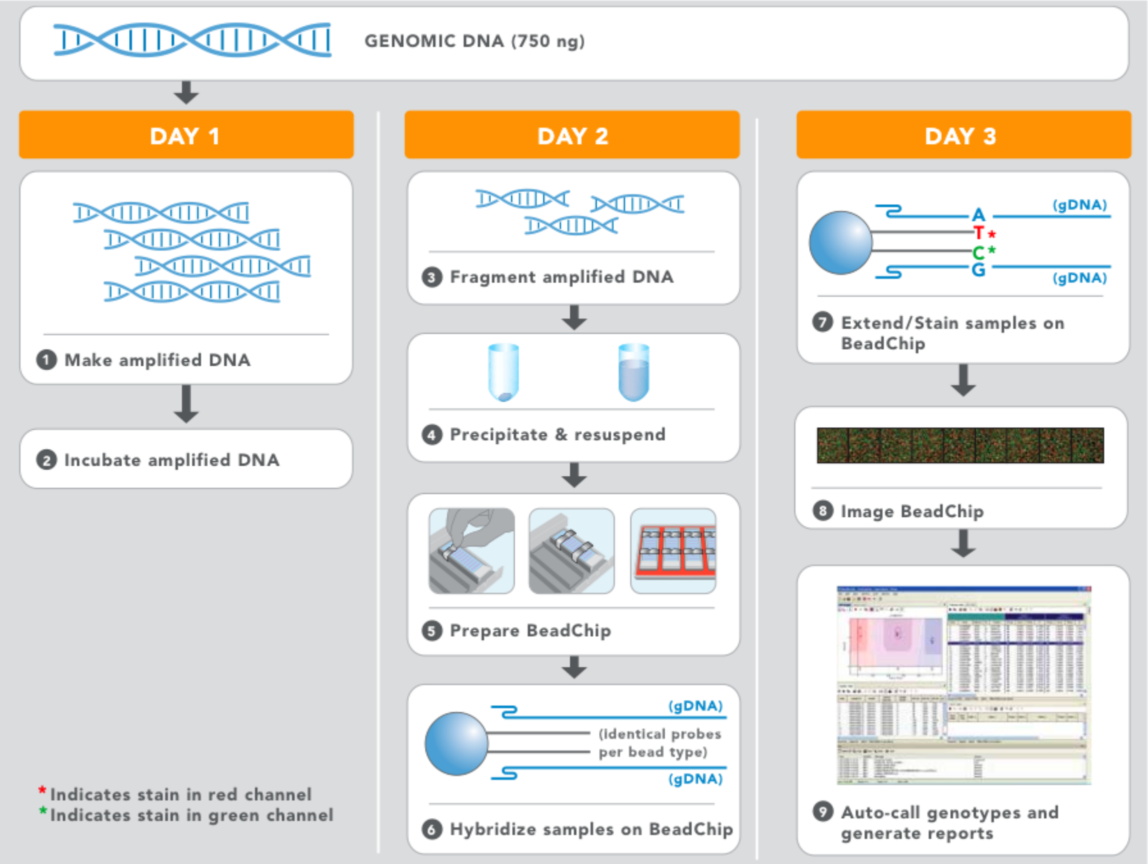

GENOMIC DNA (750 ng)
DAY 1
DAY 2
DAY 3
1 Make amplified DNA
2 Incubate amplified DNA
3 Fragment amplified DNA
4 Precipitate & resuspend
5 Prepare BeadChip
(gDNA)
(identical probes per bead type)
(gDNA)
6 Hybridize samples on BeadChip
(gDNA)
A
T*
C*
G
(gDNA)
7 Extend/Stain samples on BeadChip
8 Image BeadChip
9 Auto-call genotypes and generate reports
*Indicates stain in red channel
*Indicates stain in green channel

Figure 2.1. Illumina II assay protocol. The Infinium II whole-genome genotyping assay uses a single bead type and dual colour channel approach. During Step 1 and Step 2, a DNA sample of relatively low required quantity (750 ng suffice for assaying 500,000 SNPs) is amplified and incubated overnight. The amplification has no appreciable allelic partiality. Following the amplification, the product is fragmented in an enzymatic process (Step 3). After precipitating and resuspending the DNA (Step 4), the BeadChip is prepared for hybridization (Step 5). The DNA samples are applied onto the BeadChips and incubated overnight, thus allowing the fragmented DNA to hybridize to locus-specific 50-mers on the chips which are covalently linked to one of the > 500,000 chip bead types (Step 6). One bead corresponds to each allele per SNP locus. After hybridization, an enzymatic base extension process ensures allelic specificity and the products are subsequently fluorescently stained (Step 7). Finally, the BeadArray Reader (Step 8) detects the fluorescence bead intensities, which are in turn analyzed by calling algorithms and translated into genotypic information (Step 9). Figure and assay protocol description from www.illumina.com

HapMap

HapMap genotypes have been generated by the International HapMap Consortium and are publicly available on the HapMap website (http://hapmap.ncbi.nlm.nih.gov/). The release used in this thesis (HapMap version 27, NCBI Build 36) contains SNP genotype data generated from 1,301 HapMap 3 samples collected using two platforms: the Illumina Human1M (by the WTSI) and the Affymetrix SNP 6.0 (Broad Institute). Data from the two platforms have been merged and the subset of SNPs passing the following QC criteria kept: 1) Hardy-Weinberg p-value > 10^{-6} per population; 2) genotype missingness < 0.05 per population; 3) <3 Mendel errors per population; 4) SNP must have an rsID and map to a unique genomic location. For the analysis presented in Chapter 3 I have used all common (MAF ≥ 5%) autosomal SNPs from the unrelated CEU HapMap 3 individuals (N=109). This dataset amounts to 1,186,075 SNPs.

MuTHER

The pilot MuTHER samples have been genotyped at WTSI using in parallel Illumina's 1M-Duo and 1.2M-Duo custom chips on different subsets of individuals. Before further filtering, there were 106 samples with call rate (CR) ≥ 0.90 on the 1.2M and 88 samples with CR ≥ 0.90 on the 1M chip. Combined intensity files were created for Illuminus (Teo, Inouye et al. 2007) by retaining on a per-chromosome basis only SNPs common to both chips. Additionally, any SNPs that moved position between the two chips were removed. Following further quality checks (Hardy- Weinberg p > 10^{-4}, MAF > 1%), 865,544 SNPs were kept for analysis. The QC analysis was performed by Simon Potter at WTSI.

The set of successfully genotyped samples was overlapped with individuals having corresponding expression data available. This amounted to the following sample set per tissue: 156 LCL, 160 skin and 166 fat individuals (Chapter 4).

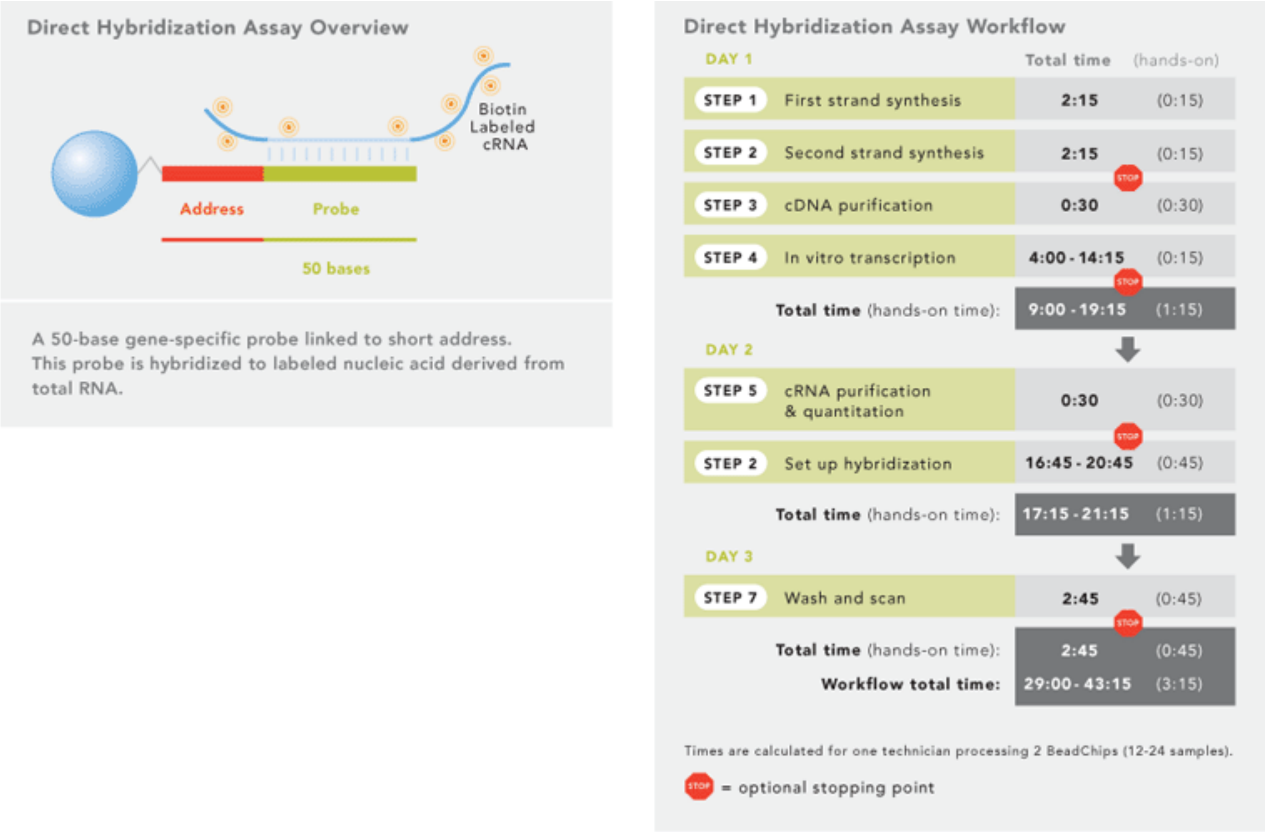
GenCord

The 85 GenCord individuals were genotyped for approximately half a million SNPs each using Illumina's 550K SNP array. DNA samples were extracted from cord tissue LCLs with the Puregene cell kit (Gentra-Qiagen, Venlo, The Netherlands). This work was carried out by Samuel Deutsch and colleagues in Stylianos Antonarakis' lab at UGMS. Principal component analysis (PCA) was performed on the genotype data to detect potential outliers. Following this analysis performed by Stephen Montgomery at the WTSI, ten individuals were removed. After further QC analysis (removing SNPs with missing data), 394,651 SNPs with MAF ≥ 5% were kept for analysis (Chapter 3).
To increase the power to detect associations with expression, GenCord genotypes were imputed onto the reference HapMap 2 data using the BEAGLE software (Browning and Browning 2007). Following imputation, QC was performed whereby SNPs with imputation quality scores < 0.9 (24,7078 SNPs) and those failing MAF (<5%) or Hardy-Weinberg equilibrium checks (total of 67,718 SNPs) were removed. This work was performed at UGMS by Eugenia Migliavacca (imputation) and Tuuli Lappalainen (QC). A final set of 1,428,314 SNPs in 75 individuals was used for the analysis In Chapter 5.

2.3 Gene expression quantification

Transcript levels in HapMap (LCL), GenCord (LCLs, fibroblasts, T cells) and MuTHER (LCL, skin, fat) samples were quantified at WTSI using Illumina's whole-genome gene expression arrays. HapMap and GenCord data are also publicly available at http://www.sanger.ac.uk/resources/software/genevar/.

Whole-genome expression profiling is based on the direct hybridization technology developed by Illumina (Figure 2.2).



A 50-base gene-specific probe linked to short address. This probe is hybridized to labeled nucleic acid derived from total RNA.

Direct Hybridization Assay Workflow

		Total time	(hands-on)
DAY 1			
STEP 1	First strand synthesis	2:15	(0:15)
STEP 2	Second strand synthesis	2:15	(0:15)
STEP 3	cDNA purification	0:30	(0:30)
STEP 4	In vitro transcription	4:00 - 14:15	(0:15)
	Total time (hands-on time):	9:00 - 19:15	(1:15)
DAY 2			
STEP 5	cRNA purification & quantitation	0:30	(0:30)
STEP 2	Set up hybridization	16:45 - 20:45	(0:45)
	Total time (hands-on time):	17:15 - 21:15	(1:15)
DAY 3			
STEP 7	Wash and scan	2:45	(0:45)
	Total time (hands-on time):	2:45	(0:45)
	Workflow total time:	29:00 - 43:15	(3:15)

Times are calculated for one technician processing 2 BeadChips (12-24 samples).

STOP = optional stopping point

Figure 2.2. Direct hybridization assay overview and workflow. Figure from www.illumina.com

The protocol features first the amplification of the starting RNA material via first- and second-strand reverse transcription, followed by a single in vitro transcription (IVT) amplification that incorporates biotin-labelled nucleotides. The resulting cRNA is purified, hybridized to the array and labelled with Cy3-streptavidin (Amersham Biosciences, Little Chalfont, UK). The fluorescence emission by Cy3 is scanned and quantified with Bead Station (Illumina).

More than 48,000 unique bead types (one for each of the 47,294 transcripts plus controls) are represented on the array. Each bead contains several hundred thousand copies of gene-specific 50mer probes covalently attached. The probes are derived from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) and UniGene databases. The beads are assembled into 3 µm diameter wells, generating an average 30-fold redundant information for each probe. These background-corrected values for a single bead type are summarized by Bead Studio (Illumina software) and outputted to the user as a set of 47,294 intensity values for each individual hybridization.

HapMap

Total RNA was extracted from LCLs derived from the HapMap 3 individuals (Coriell). Gene expression was quantified using Illumina's commercial array Sentrix Human-6 Expression BeadChip version 2. For each RNA extraction, two one-quarter scale Message Amp II reactions (IVTs) (Ambion, Austin, Texas, USA) were performed using 200 ng of total RNA, to produce cRNA. To assay transcript levels, 1.5 µg of the cRNA were hybridized to the whole-genome expression array. Six arrays were run in parallel on each individual BeadChip. The experimental work was carried out by Catherine Ingle, James Nisbet and Magdalena Sekowska at the WTSI.

To combine information from the two replicate hybridizations, raw data was normalized on a $\log_2$ scale by quantile normalization (Bolstad, Irizarry et al. 2003) across replicates of a single individual followed by median normalization across all individuals from a single population. Normalization was performed by Stephen Montgomery at WTSI.

Of the >48,000 probes represented on the array, only a trustable subset was chosen for further analysis. The Sentrix Human-6 Expression BeadChip version 2 array covers over 24,000 unique, curated RefSeq genes, as well as genes with less well-established annotation. Only probes corresponding to well-annotated RefSeq genes were kept at this point. Additionally, probes were matched to corresponding Ensembl genes (Ensembl 49 NCBI Build 36) using SSAHA (Sequence Search and Alignment by Hashing Algorithm) (Ning, Cox et al. 2001). Following the SSAHA run 22,512 probes were mapped to 19,862 Ensembl genes. Probes mapping to multiple Ensembl genes were removed, as well as ones mapping to sex chromosomes. After filtering, a non-redundant set of 21,800 probes (corresponding to 18,226 Ensembl genes) was used for association analysis. Mapping and selection of probes for final analysis was carried out by Antigone Dimas at WTSI.

MuTHER

RNA was extracted from LCLs, skin and fat samples derived from the pilot MuTHER individuals. Gene expression was measured using Illumina's HumanHT-12 version 3 whole-genome array, as explained previously (in this case, each sample had three technical replicates). The experimental work was carried out by James Nisbet and Magdalena Sekowska at WTSI and by Amy Barrett and Mary Travers at WTCHG.

Log_2-transformed expression signals were normalized separately per tissue as follows: quantile normalization was performed across the 3 replicates of each individual followed by quantile normalization across all individuals.

The >48,000 probes targeting more than 25,000 genes are derived from RefSeq (Build 36.2, Rel 22) and UniGene (Build 199). To select probes corresponding to well-annotated genes, Illumina's v3 probes were mapped to unique Ensembl gene IDs by combining and cross-checking two methods. The first approach used probe annotations to RefSeq IDs provided by Illumina, which were further queried with BioMart (Ensembl 54) for corresponding Ensembl genes IDs. RefSeq IDs mapping to multiple Ensembl Genes were excluded, and only autosomal genes retained. This step was performed with the help of Tsun-Po Yang at WTSI. The second approach used BLAT (Kent 2002) to map the 50-mer probe sequences to Ensembl transcripts and to extract genomic locations matching all 50 bases of the probe sequence. Probes with unique perfect match to the genome and corresponding transcripts matching to the same genes were kept. This approach was performed by Josine Min at WTCHG. The union of the two mapping approaches after excluding 196 conflictingly matching probes resulted in 27,499 probes corresponding to 18,170 autosomal genes available for association analysis.

GenCord

Total RNA was extracted from LCLs, fibroblasts and T-cells of the 85 GenCord individuals. Two one-quarter scale Message Amp II reactions (Ambion) were performed for each RNA extraction with 200 ng of total RNA. 1.5 μg of cRNA was hybridized to Illumina's WG-6 v3 Expression BeadChip array to quantify transcript abundance as described previously. Each RNA sample had two technical replicates. This work was carried out by Catherine Ingle, James Nisbet, and Magdalena Sekowska at the WTSI.

The expression raw data was normalized independently for each cell type as follows: the intensity values were $\log_2$ transformed, quantile normalized per sample replicates and median normalized 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.

The WG-6 v3 Expression BeadChip array covers over 27,000 unique coding transcripts. For some of them, well-established annotation exists (7,000 transcripts have provisional annotation). In addition, the array covers non-coding transcripts, as well as experimentally confirmed mRNA sequences aligning to EST clusters. Again, only probes with good or provisional annotation (mapping to RefSeq genes) were selected of the total 48,000 probe set 36,156 probes with Refseq IDs were queried for their corresponding Ensembl gene IDs in Biomart (Ensembl 50, NCBI Build 36). Of these, 22,651 probes had a uniquely assigned Ensembl gene ID and did not map to either chromosomes X or Y. These probes corresponding to 17,945 RefSeq genes and 15,596 Ensembl genes respectively were used for subsequent analysis. Selection of the final probe list was done by Antigone Dimas at WTSI.

2.4 eQTL discovery

Associations between SNP genotypes and normalized expression values were run using Spearman Rank Correlation (SRC) and additive linear regression (LR). SRC was exclusively used to detect eQTLs (Chapter 4) while LR was used to quantify the proportion of expression variance unexplained by the SNP genotypic classes (Chapter 3, Chapter 5). I considered SNPs within a 1Mb window on either side of a gene’s transcription start site (TSS) as *cis*-acting while SNPs located further than 5 Mb away either side of a gene’s TSS or SNP-gene pairs on different chromosomes as *trans*-acting.

2.4.1 Association analysis

Before association, the SNP genotypes were numerically encoded (0, 1 or 2) to represent the counts of alphabetically sorted alleles at each locus (e.g. counting the number of G alleles for an A/G SNP: AA = 0, AG = 1, GG = 2) (Figure 2.3).

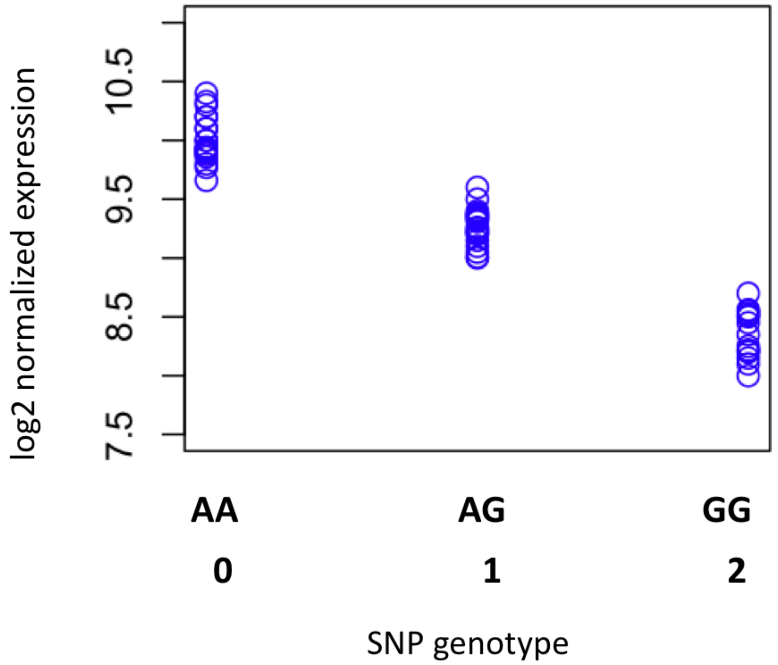


Figure 2.3. SNP-gene association example. The A/G SNP in this schematic example is plotted against a gene's corresponding normalized $\log_2$ expression values. In this case, the A allele at the SNP locus predisposes individuals to have higher expression values of the respective gene.

2.4.1.1 Spearman Rank Correlation (SRC)

SRC is a non-parametric test assessing the degree of statistical dependence between two variables (X and Y). A monotonic function is fitted to describe the correlation between X and Y (e.g. X = genotype, Y = expression). No other assumption is made about the relationship between the two variables, which are rank-ordered. In our case for example, expression values are ordered low to high and ranked accordingly (1..n), irrespective of their actual numerical value. This makes sure that outliers do not have a high impact on estimating the correlation between X and Y. The degree and direction of this correlation is reflected in the ρ (rho) coefficient, calculated as below, where *n* is the number of observations and d_i is the difference between the ranks of each observation on the two variables ($d_i = x_i - y_i$):

$$\rho = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)}$$

When two observations for the same variable are equal (tied), they are each assigned the average corresponding rank. A perfect Spearman correlation ($\rho = 1$ or $\rho = -1$) occurs when each of the variables is a perfect monotone function of the other. The sign marks the direction of the correlation: $\rho > 0$ (positive correlation) if Y tends to increase when X increases and $\rho < 0$ (negative correlation) if Y tends to decrease when X increased. A nominal p-value for the association test is also reported.

2.4.1.2 Additive linear regression (LR)

In a LR model, the relationship between two variables is explored by fitting a linear equation to the observed values. For the work presented in this thesis, the following main effects additive model was used to test for SNP-gene expression associations:

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

Here, the dependent variable Y_i is a probe's normalized $\log_2$ expression value quantified in individual *i* (*i* = 1..n) and the explanatory variable X*i* is the corresponding numerically encoded genotype. ε_i are independent normally distributed random variables with mean 0 and constant variance (Stranger, Forrest et al. 2005). b_i is the slope of the fitted regression line ($b_i = 0$ if there is no association between the genotype and the expression values). How well the regression model fits the data can be estimated from the inspection of the residuals i.e. the vertical distances of each point from the regression line. The residuals quantify the proportion of the variance in the dependent variable (Y - expression) that cannot be accounted for by the explanatory variable (X - genotype). As such, the most common regression technique employs minimizing the sum of squared residuals.

2.4.2 Multiple testing correction

The statistical significance of associations between SNP genotypes and gene expression levels was assessed using permutations (Churchill and Doerge 1994; Doerge and Churchill 1996). The $\log_2$ normalized expression values of each probe were permuted 10,000 times relative to the genotypes of the SNPs in the tested window (2MB in *cis*). The minimal p-value association of each run was retained generating thus a distribution of 10,000 values corresponding to the best random SNP-probe associations. Significance was assessed for different threshold levels (0.5, 0.01, 0.001 and 0.0001) by comparing the tail of the distribution of the 10,000 minimal p-values for each gene to the observed association p-value (e.g. an association was considered significant at the 0.0001 threshold if the nominal observed p-value was lower than the 0.0001 tail of the distribution of minimal permuted p-values) (Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007).

2.5 Recombination hotspot mapping and LD filtering

To restrict the search space for causal regulatory effects and refine eQTL signals, I have made use of the genome's correlation structure (LD). Specifically, I used recombination hotspot coordinates derived from the statistical analysis of the variation data generated by the HapMap 2 project (Release 22, Build 36) (McVean, Myers et al. 2004) (Myers, Bottolo et al. 2005). The recombination hotspots inferred are typically 1-2 kb long and are surrounded by much larger regions (defined here as recombination hotspot intervals) essentially devoid of recombination (Paigen and Petkov 2010). All autosomal SNPs in HapMap 3 CEU, MuTHER and GenCord have been mapped to recombination hotspots and hotspot intervals. The mapping serves both to restrict the search for functional regulatory variants explaining GWAS signals (Chapter 3, Chapter 5) and also for refining eQTL signals by identifying independent regulatory effects and comparing them across multiple tissues (Chapter 4).

In Chapters 3 and 5, GWAS results are tested for explanatory regulatory effects. For this purpose, given any GWAS SNP, I focus on the recombination hotspot interval where it resides and where also at least one eQTL co-localizes. Limiting the search space for causal effects to these intervals with independent recombination history is a reasonable approach, as few or no recombination events are expected between the reported associated SNPs and the functional variants they are tagging.

In Chapter 4, I aim to characterize in detail the landscape of regulatory variation across LCLs, skin and fat. For this reason, I refine the discovered eQTL signals to likely independent effects per gene. The strategy employed is the following: after mapping significant eQTLs to recombination hotspot intervals, the most significant SNP per gene per interval is kept. Furthermore, to avoid long-range correlations which can extend over recombination hotspots, an additional LD filtering step is performed so that for each pair of significant eQTLs with D' > 0.5, the least significant SNP is ignored. The choice of D' over r^2 as LD filtering metric is based on their distinctive properties. Both metrics relate to D, the basic unit of LD measuring the deviation of haplotype frequencies from equilibrium state (Lewontin and Dunn 1960). For two SNPs with alleles (A,a) and (B,b) respectively:

$$D = f(AB) - f(A)f(B)$$

where f(X) is the frequency of the X allele. If D is significantly different from 0, LD occurs. D', calculated as below, ranges from 0 to 1, with D' =1 denoting complete LD while values towards 0 indicating linkage equilibrium, i.e. historical genetic independence.

$$if D \geq 0, \quad D' = \frac{D}{D_{\max}}$$
$$if D < 0, \quad D' = \frac{D}{D_{\min}}$$

r^2 is the statistical coefficient of determination, or the measure of correlation between a pair of variables (SNP genotypic classes in this case). Also in the range of 0 to 1, $r^2 = 1$ indicates that one SNP is directly predictive of the other (perfect correlation) and lower values denote the decay of their correlation (r^2 approaches 0) (Wang, Barratt et al. 2005).

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

While r^2 quantifies the statistical correlation between two variants, D' is a measure of their historical relationship which is biologically more meaningful. For example, two correlated SNPs in between which no recombination event occurred (D' =1) but which have different MAFs (low r^2) can be tagging the same functional effect (e.g. a single independent regulatory variant residing in the respective hotspot interval). The stringent D' threshold (which corresponds to an even lower r^2) provides thus a more suitable method to filter for historically independent effects. When comparing across tissues, this filtering ensures that true shared effects (interval-gene combinations) are contrasted and not just genes, which would be inaccurate in cases when the same gene is regulated by different functional variants in different tissues.

2.6 RTC scoring scheme (Chapter 3, Chapter 5)

The Regulatory Trait Concordance (RTC) method was developed in order to detect the subset of GWAS signals which could be explained by significant regulatory effects and identify the genes whose expression levels they mediate. For this purpose, I used expression data from two resources: HapMap 3 and GenCord. The whole-genome expression quantification experiments on the MuTHER pilot samples were performed towards the end of my PhD and were not available for analysis at that time. eQTLs discovered in LCLs derived from HapMap 3 CEU and GenCord individuals were tested in Chapter 3, while eQTLs detected in the three GenCord tissues (LCLs, fibroblasts, T-cells) were overlaid with GWAS results in Chapter 5. I next describe the RTC method and the main experiments it has been used for.

2.6.1 Method overview

I assess the likelihood of a shared functional effect between a GWAS SNP and an eQTL by quantifying the change in the statistical significance of the eQTL after correcting for the genetic effect of the GWAS SNP. The correction is performed using a LR model. The GWAS SNP is first regressed against normalized expression values of the gene for which an eQTL exists. The residuals capture the remaining unexplained expression variance after the removal (correction) of the GWAS SNP effect. This resulting pseudo phenotype is used to redo the SRC association with the eQTL genotype. It is expected that if the GWAS SNP mediates the disease effect through a change in gene expression due to a regulatory variant (eQTL) then correcting out the GWAS SNP effect will have a marked consequence on the eQTL i.e. the eQTL SNP – gene association p-value after correction will be much less significant than the association p-value before correction. The p-value estimates however, are affected also by the LD structure of the investigated region: the correlation between the eQTL and the GWAS SNP but also between each of the two and the actual functional variants (most often unknown) influence the correction outcome. Given that part of the change in the p-values will be attributed to LD, it is necessary to account for this correlation in each interval of interest.

I account for the LD structure in each hotspot interval separately by ranking (Rank $_{\text{GWAS SNP}}$) the impact on the eQTL (quantified by the adjusted association P-value after correction) of the GWAS SNP correction to that of correcting for all other SNPs in the same interval. The rank denotes the number of SNPs which when used to correct the expression data, have a higher impact on the eQTL (less significant adjusted P-value) than the GWAS SNP (i.e. Rank$_{GWAS\ SNP}$ = 0 if the GWAS SNP is the same as the eQTL SNP, Rank$_{GWAS\ SNP}$ = 1 if of all the SNPs in the interval, the GWAS SNP has the largest impact on the eQTL etc). By taking into account the total number of SNPs in the interval (N$_{\text{SNPs}}$), this ranking can be compared across different genes and intervals. For this purpose, the RTC score is defined as follows:

$$RTC = \frac{N_{SNPs} - Rank_{GWAS\ SNP}}{N_{SNPs}}$$

The RTC score ranges from 0 to 1, with values closer to 1 indicating causal regulatory effects. The highest RTC statistic (RTC = 1) is obtained for the lowest correction ranking (Rank$_{GWAS\ SNP}$ = 0) corresponding to cases when the GWAS SNP is identical to the eQTL. As expected in these instances, correcting the eQTL SNP with itself removes the largest possible amount of variance, more so than with any other SNP in the region. Cases when the eQTL and GWAS SNP are identical are impossible to resolve with the RTC or any other method. They are however still informative, indicating that the pattern of association between the SNPs in that region and the disease phenotype and gene expression respectively are identical.

2.6.2 RTC properties under simulations

Before applying it to large-scale expression datasets, I investigated the properties and robustness of the RTC score with respect to D' and r^2, the two most common LD metrics. Both possible scenarios were tested: the null hypothesis (H_0) when a GWAS disease SNP (dSNP) and a co-localizing eQTL would tag two different causal variants and the alternative hypothesis (H_1) when the eQTL and dSNP tag the same functional variant. For this purpose, I have simulated causal SNPs (cSNP), eQTLs and dSNPs under different scenarios varying the LD levels between them as well as the LD pattern of the hotspot interval where they reside. The dSNP emulates the most significant trait-associated SNP typically reported by GWAS studies, while the cSNP represents the actual functional variant, unknown most of the times. For each simulated case, the cSNP was first masked, then the RTC was calculated and its performance evaluated. I used the HapMap 3 CEU *cis* eQTLs (315 genes at 10^{-3} permutation threshold) to create the list of cSNPs.

For the H_0 test, the cSNPs were called causal eQTL SNPs (c-eQTLs). For each c-eQTL, I sampled a different causal disease SNP (c-dSNP) from the same recombination hotspot interval, with the requirement that its MAF comes from a distribution identical to that of the GWAS SNPs downloaded from NHGRI (976 GWAS variants) (website accessed 02.03.09). Subsequently, I sampled up to five eQTL-dSNP pairs per interval where the eQTLs and dSNPs are the topmost correlated (r^2) SNPs with the c-eQTL and the c-dSNP respectively. These imitate the typical tagging SNPs reported as having a significant association with gene expression and disease phenotypes respectively. After sampling, I excluded cases where the eQTL and dSNP are identical, as these contradict the H_0. c-eQTL-c-dSNP-eQTL-dSNP quartets mapping to 287 unique hotspot intervals were sampled and tested under H_0. The RTC score was calculated for all simulated eQTL-dSNP pairs in each of the 287 hotspot intervals. The predictive value of the RTC score was compared against standard measures of LD (r^2, D’) between the eQTL and the dSNP.

Under the H_1, the cSNP represents the untyped causal variant mediating the disease association via significant changes in gene expression levels. In this case, both the eQTL and the dSNP tag the same effect. Therefore, up to five eQTL-dSNP pairs were sampled for each hotspot interval harbouring a cSNP under H_1 as follows: the eQTLs were chosen as the top most significant SNPs per eQTL gene - excluding the cSNP; the dSNPs were randomly sampled from the same hotspot interval such that the r^2 between each of them and the cSNP was in the range [0.5,0.9]. Perfectly correlated SNPs ($r^2 = 1$) were excluded, as such cases cannot be resolved. In addition, at any stage of the 5-step iteration process per cSNP, the dSNP was selected to be different from the cSNP and the eQTLs sampled up to that point. cSNP-eQTL-dSNP trios mapping to 290 unique hotspot intervals throughout the genome were sampled and tested under the H_1. For all simulated eQTL-dSNP pairs per each hotspot interval (N = 290), the RTC score was calculated and its predictive value compared against the correlation level (r^2, D') between the eQTL and the dSNP.

Finally, the effect of a region's overall LD pattern on estimating the RTC score was explored. For this purpose, the extent of LD per hotspot interval was calculated as the median r^2 of all pairwise SNP combinations available per interval. Under both H_0 and H_1, the relationship between the median r^2 of a hotspot interval and the RTC was investigated.

The RTC properties as revealed by these analyses are described in Chapter 3.

2.7 MuTHER eQTL analysis (Chapter 4)

2.7.1 Factor analysis

eQTL analysis on the MuTHER pilot data was performed using the discovery framework presented in Section 2.4 of this chapter (Methods). Additionally, eQTL analysis was conducted after accounting for experimental noise and global environmental conditions, which are also known to impact gene expression in a global manner. For this purpose, a Bayesian factor analysis (FA) model (Stegle, Parts et al. 2010) was applied to the expression data in each tissue. This approach uses an unsupervised linear model to account for global variance components in the data, and yields a residual expression dataset that can be used in further analysis.

A wide range of parameter settings was tested for the model, controlling the amount of variance explained by it. This was achieved by setting the parameters of the prior distributions for gene expression precision (inverse variance) and factor weight precision. These random variables are modelled using Gamma distributions, thus their natural exponential family parameters (the prior mean and number of prior observations) were varied. The prior mean was varied from 10^{-6} to 10^{-2} and the number of prior observations from $N*10^{-3}$ to N, where N is the number of observations from data. 120 latent factors were thus learned. For each tissue, the residual dataset that gave the best eQTL overlap between co-twin samples was used in the subsequent eQTL analyses. The prior values used for each dataset are given in Table 2.2. The FA was developed and carried out by Leopold Parts at WTSI.

	Weight prior		Noise prior	
	Mean	**Observations**	**Mean**	**Observations**
LCL	10^{-6}	23	10^{-3}	10
SKIN	10^{-6}	23	10^{-1}	100
FAT	10^{-6}	23	10^{-3}	10

Table 2.2. Factor analysis weight and noise prior values applied to each tissue. Analysis performed on MuTHER pilot samples.

Following FA, the eQTL analysis on the corrected expression data was performed identically to the original detection strategy: SRC followed by multiple-testing correction using permutations.

2.7.2 Estimation of proportion of true positives (π_1)

Overlapping eQTL discoveries at the same threshold is very sensitive to power, as thresholds are driven by statistical significance. Given this, eQTL replication and tissue sharing was quantified also in a continuous way with Storey's qvalue statistic (Storey and Tibshirani 2003). The QVALUE software implemented in the R package qvalue 1.20.0 was used under the default recommended settings. The program takes a list of p-values and computes their estimated π_0 - the proportion of features that are truly null - based on their distribution (the assumption used is that alternative cases tend to be close to zero, while p-values of null features will be uniformly distributed among [0,1]). The quantity $\pi_1 = 1- \pi_0$ estimates the lower bound of the proportion of truly alternative features, i.e. the proportion of true positives (TP). Replication and sharing between two samples is reported as the proportion of TP (π_1) estimated from the p-value distribution in the second sample of independent eQTLs initially discovered in the first sample (exact snp-probe combinations are used).