5 CELL TYPE SPECIFICITY OF CIS REGULATORY VARIATION

In this chapter I will:

- Underline that most studies investigating regulatory variation to date explore expression in a single cell type.
- Stress the value of documenting cell type-specific regulatory variation.
- Describe a resource and experimental strategy that enable detection of eQTLs across cell types.
- Outline that the majority of eQTLs identified using this resource are cell type-specific.
- Emphasize the value of large collections of LCLs.

5.1 THE VALUE OF STUDYING DIFFERENT CELL TYPES

Variation influencing gene expression can manifest itself as gene expression differences among populations, among individuals in a population, among tissues, and in response to environmental factors. As discussed in the previous chapters, the genetic basis of the first two types of gene expression variation has been investigated in a number of studies with the quantification of mRNA in one tissue and the identification of eQTLs in a single or multiple populations (Adams, Kerlavage et al. 1995; Reymond, Marigo et al. 2002; Su, Cooke et al. 2002). The complex developmental program in higher eukaryotes however results in a vast set of highly specialized cell types, whose fate is determined to a large extent by the combination of expressed genes and their level of expression. During development, but also in differentiated cells, some genes exhibit ubiquitous patterns of expression while others display tissue-specific activity (Myers, Gibbs et al. 2007; Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008). The extent to which

genetic variation manifests itself as tissue-specific gene expression patterns remains unknown and eQTL cell type specificity remains underexplored. A handful of studies have identified eQTLs in certain human (Myers, Gibbs et al. 2007; Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008) and mammalian (Cotsapas, Williams et al. 2006; Campbell, Kirby et al. 2008) tissues but a systematic study comparing eQTLs across a wide range of cell types, while controlling for confounding associations, such as population samples and differences in technology or statistical methodology, is lacking in humans. Studies in model organisms however are highlighting the value of interrogating regulatory variation systematically and in a tissue-specific context (Petretto, Mangion et al. 2006; Huang, Shifman et al. 2009).

The importance of documenting cell type-specific regulatory variation is high given the role of gene expression patterns in determining cell type during development, in shaping higher level phenotypes and in determining disease risk. In cases such as asthma (Moffatt, Kabesch et al. 2007) and colorectal cancer (Valle, Serena-Acedo et al. 2008) documenting genetic control of gene expression variation is likely to shed light on mechanisms of disease pathogenesis. Furthermore there is growing evidence that causative variants identified in GWAS are likely to behave in a cell type-specific manner (Wellcome Trust Case Control Consortium 2007). Cataloguing cell type-specific regulatory variation can therefore serve to connect biological pathways controlling cellular activities in health and disease (Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008; Wu, Delano et al. 2008).

The case of CD, an autoimmune inflammatory disease of the gastrointestinal tract, illustrates the critical role of eQTLs in elucidating disease pathogenesis. GWAS revealed a strong signal in a 1.25 Mb gene desert of chromosome 5p13.1 (Libioulle, Louis et al. 2007; Wellcome Trust Case Control Consortium 2007). Expression association studies quantifying transcript levels in LCLs (Libioulle, Louis et al. 2007), revealed that the same region showed a strong association with transcript levels of

PTGER4. Knockout mice for *PTGER4* have increased susceptibility to colitis, rendering this gene a strong susceptibility candidate for CD (Servitja, Pignatelli et al. 2009).

In cases such as the above, where disease and gene expression signals map to the same chromosomal location (Figure 24 a), integrating information from both sources may provide important clues about the genes and functional pathways involved in disease pathogenesis. CD is an immune system disease and studying expression in immune system-derived LCLs has proven informative in terms of pointing to candidate genes. In this respect LCLs are a relevant cell type to study for CD. In the case of other phenotypes however (e.g. diabetes) expression association signals in LCLs may not yield signals that track disease association (Figure 24 b). Interrogating expression in pancreatic-islet β-cells might provide more clues for the pathogenesis of diabetes (Nica and Dermitzakis 2008).

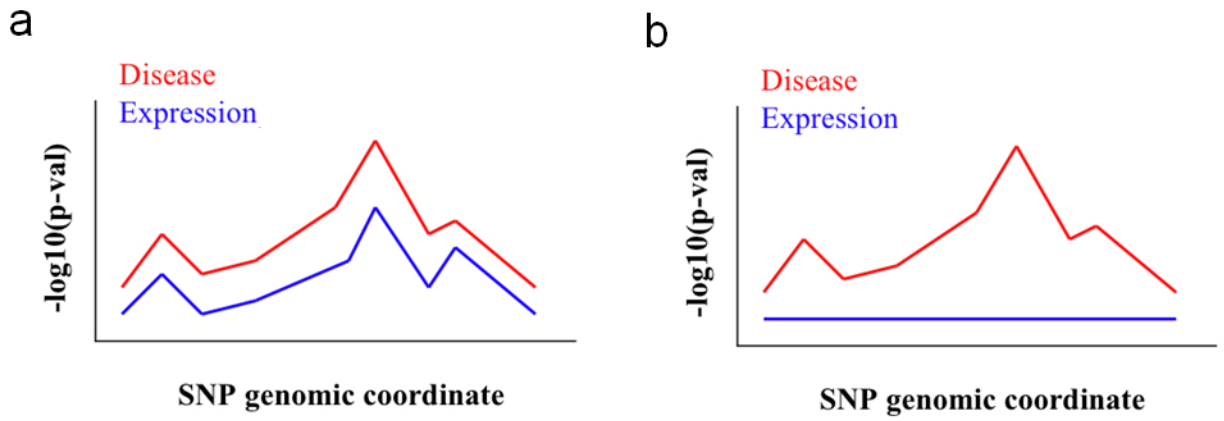

Figure 24. Disease and expression signals from genome-wide association studies (GWAS). The x axis represents chromosomal location, the y axis shows the significance of association for SNPs along the chromosome. In **a)** expression and disease association signals track one another, implying that expression of the particular gene in the cell type studied may be involved in disease pathogenesis. This is the case for Crohn disease (CD) where a SNP on chromosome 5 was associated with expression levels of *PTGER4* in LCLs and also showed a significant association to disease. In **b)** expression and disease association signals do not track one another, implying that expression of the particular gene in the cell type studied is probably not relevant for the disease. Given the important role of gene expression in disease pathogenesis, it is necessary to investigate multiple cell types to determine whether there are cases in which

expression signals mirror those of disease. In this way it will be possible to identify loci with a functional role in disease pathogenesis. Figure adapted from (Nica and Dermitzakis 2008).

It is not clear how straightforward it will be to determine which cell type or tissue is relevant for a particular disease or complex trait. As with candidate gene studies it may turn out that in some cases the relevant cell type is not the one that was identified as a candidate based on existing biological knowledge. Interrogating expression in blood-derived LCLs for example has proven useful for identifying genes implicated in the pathophysiology of autism (Nishimura, Martin et al. 2007). Gene expression profiles from males with autism and non-autistic controls clearly distinguished cases from controls. It is yet not clear how many cell types and tissues will be adequate to provide a catalogue of regulatory variation, but this approach contributes to efforts using functional genomic information to interpret the biological effects of disease or complex trait variants. To date, such efforts are hindered by the limited availability of the relevant cell type to perform the functional assays. Understanding the degree of tissue-specificity of regulatory variation will enable us to assess how much we are missing by interrogating only a limited number of tissues and will provide clues as to how many tissues will be required to capture the spectrum of functional consequences of disease-causing variants (McCarthy and Hirschhorn 2008; Nica and Dermitzakis 2008).

In this study, I assessed cell type specificity of variants impacting gene expression by quantifying mRNA levels in three cell types from each of 85 individuals, and by identifying shared and cell type-specific eQTLs. I also explored the fine-scale architecture of cis regulatory landscapes conditioning on cell type, to determine the extent to which genes are regulated by common or cell type-specific regulatory elements. This work has been described in (Dimas, Deutsch et al. 2009).

5.2 DETECTING CIS EQTLS IN THREE CELL TYPES

Eighty five individuals from the GenCord resource were studied to explore the cell type-specific distribution of cis regulatory variation. GenCord is a collection of cell lines derived from umbilical cords of individuals of Western European origin (see section 2.3.3). Sample collection was performed systematically on full term or near full term pregnancies, to ensure homogeneity for sample age. mRNA levels were quantified in primary fibroblasts, LCLs, and primary T-cells for 48,804 probes using the illumina WG-6 v3 Expression BeadChip array. Data from 22,651 probes, mapping to 17,945 autosomal RefSeq genes (15,596 Ensembl genes) were analysed. The same samples were genotyped on the illumina 550K SNP array. Following quality control (SNPs with missing data were removed) and minor allele frequency filtering (MAF ≥ 5%), 394,651 SNPs were used for association testing. PCA detected ten potential outlier individuals from the genotype data (Figure 25) who were subsequently removed from the analysis. eQTL discovery and all other properties of the results for 75 vs. 85 individuals were almost identical (Figure 26).

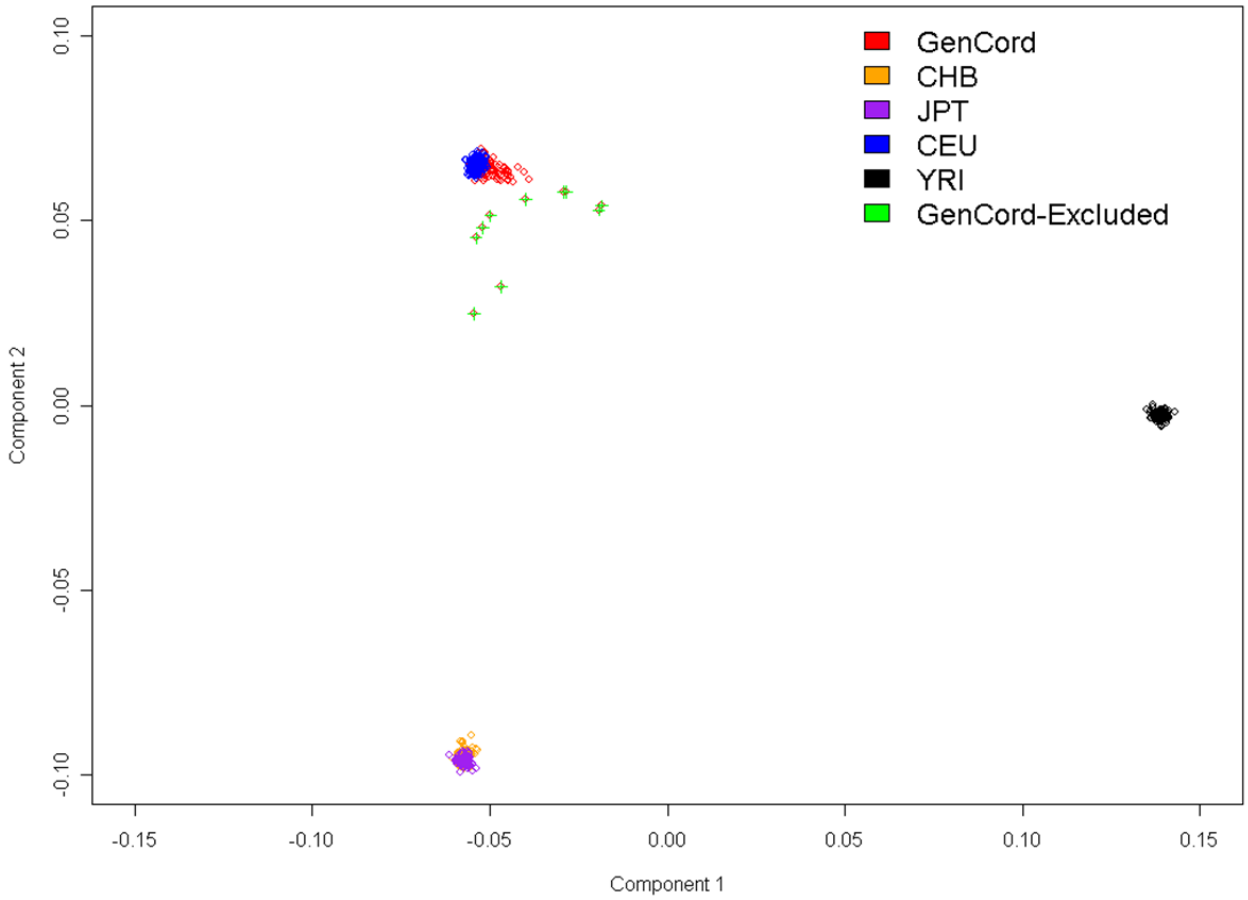

Figure 25. Principal components analysis (PCA) of the GenCord and HapMap Phase 2 populations. GenCord individuals were clustered with the HapMap populations (CEU, CHB, JPT and YRI) to assess relative population stratification in the samples. Given the observed clustering along the first two principal components, ten outliers were removed from the analysis (GenCord-Excluded).

I explored associations in cis, by testing SNPs mapping within a 2 Mb window centred on the TSS of genes. SRC was used to test for association between SNP genotype and mRNA levels, after intensity normalization and $\log_2$ transformation, performed separately for each cell type. A total of 6,083,130 tests were performed and significance thresholds for each gene were assigned through permutations. For 75 individuals at the 0.01 permutation threshold I discovered 2,146, 2,155 and 2,046 genes with significant cis eQTLs in fibroblasts, LCLs and T-cells respectively, with an

estimated FDR of 7%. At the stricter 0.001 permutation threshold, I discovered 427, 442 and 430 genes with significant cis associations in fibroblasts, LCLs and T-cells respectively, with an estimated FDR of 4% (Figure 26). The genomic distribution of detected associations at the 0.001 threshold in each cell type is shown in Figure 27.

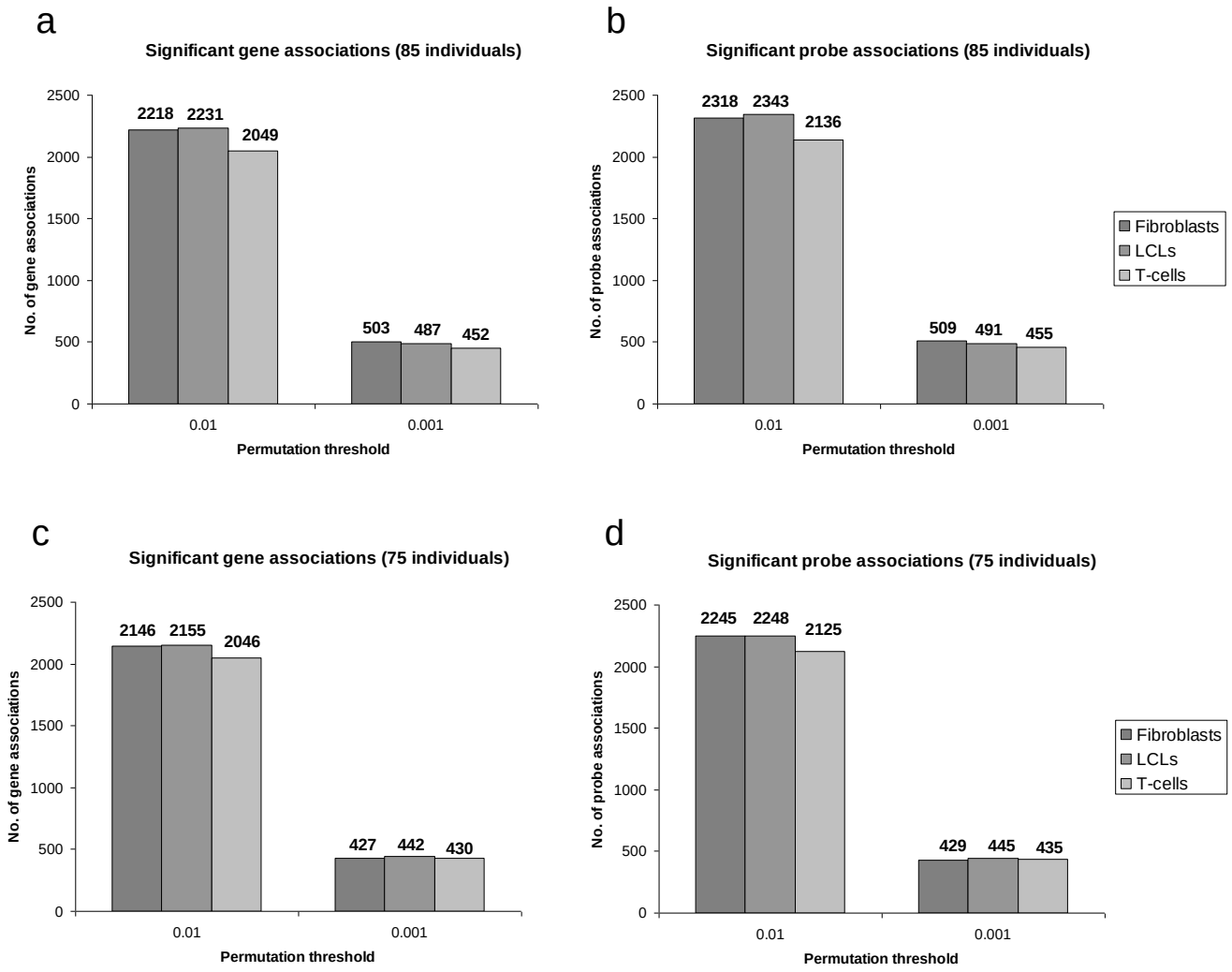

Figure 26. Significant gene and probe associations in GenCord cell types at the 0.01 and 0.001 permutation thresholds. Numbers on top of the histogram bars represent counts of associations. **a)** and **b)** show gene and probe associations detected using 85 individuals and **c)** and **d)** show gene and probe associations detected using 75 individuals, after removal of 10 outliers. Association detection was highly similar in both analyses, with comparable estimated false discovery rates (FDR = 7% for genes and 10% for probes for the 0.01 permutation threshold and 3% for genes and 3-4% for probes at the 0.001 permutation threshold for both the 85 and 75 individuals analyses).

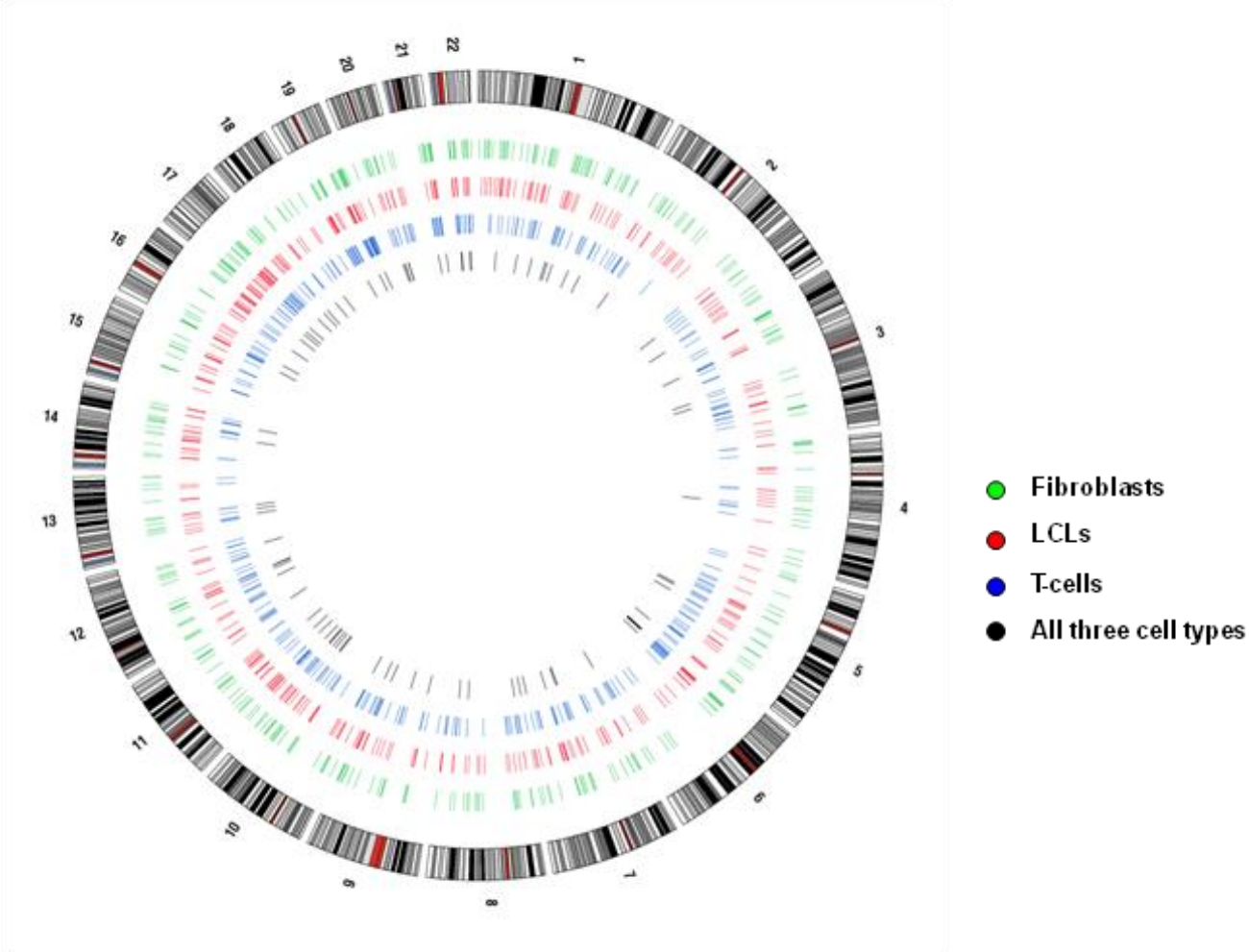

Figure 27. Genome-wide map of cis eQTLs in GenCord three cell types. cis eQTLs at the 0.001 permutation threshold are shown as colour-coded lines on their corresponding chromosomal location. Internal black lines represent genes with eQTLs in all cell types.

5.3 Replication of cis eQTLs detected in LCLs

There has been long debate about the stability of eQTLs detected in LCLs from different samples, experiments and technologies, as well as the use of large collections of these cell lines. In the present study I assessed how well previously described eQTLs from the CEU HapMap Phase 2 (International HapMap Consortium 2007; Stranger, Nica et al. 2007) are replicated in GenCord LCLs. The expectation is that a large proportion of

eQTLs will be shared, as both populations are of European descent and share similar allele frequency spectra. Due to differences in probe sequence content between the illumina v1 array (used for HapMap Phase 2 CEU) and the illumina v3 array (used for GenCord) it was possible to compare a small subset of SNP-probe associations. Comparisons were made for cases where the SNP was present in both HapMap and GenCord and the probe had identical sequence between illumina v1 and v3 expression arrays. Strict filtering was performed to avoid confounding effects arising from: a) differences in probe efficiency, b) the possibility that probes covered alternative splicing products from the same gene and c) the occurrence of probes in the v1 array containing SNPs. Of the 5,898 SNP-probe pairs that survived the 0.001 permutation threshold in HapMap Phase 2 CEU, 137 SNP-probe pairs (44 probes, some associated with multiple SNPs) were also tested in GenCord LCLs. The distribution of nominal (uncorrected) p-values from the association test for these SNP-probe pairs is greatly enriched for very low p-values, with 114 nominal p-values < 0.001 (83%) (Figure 28). Therefore, previously detected eQTLs were well-replicated, despite the long separation time between tests, demonstrating the stability of LCLs. These data highlight the value of large collections of LCLs from different cohorts for studies of gene expression and disease interpretation.

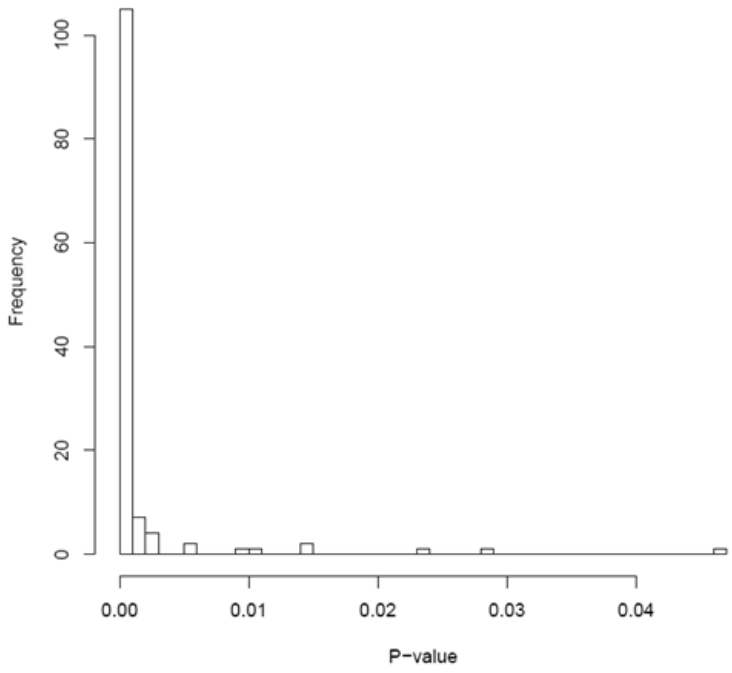

Figure 28. Replication of nominal (uncorrected) p-values in GenCord of SNP-probe associations initially identified as significant in HapMap Phase 2. I tested 137 identical SNP-probe pairs (44 probes) in GenCord LCLs. Of these, 114 SNP-probe pairs (83%) have a nominal p-value < 0.001 in GenCord LCLs, suggesting good replication of eQTLs between experiments.

5.4 SHARING AND CELL TYPE SPECIFICITY OF CIS EQTLS

Having established the robustness of eQTLs through replication, I interrogated the cell type specificity of regulatory effects by exploring genes with cis eQTLs that were: a) shared in all three cell types, b) shared in two cell types and c) cell type-specific. At the 0.001 permutation threshold, I identified a non-redundant set of 1,007 genes with cis eQTLs of which 86 (8.5%) were shared in all three cell types, 120 (12%) were shared in exactly two of the cell types and 801 (79.5%) were cell type- specific (Table 13 for genes, Table 14 for probes; results for the 0.01 permutation threshold are also shown). The proportion of cell type-specific eQTLs was similar to previous estimates of eQTL tissue specificity and alternative splicing reported in a study interrogating two tissue types, sampled however from different groups of individuals (Heinzen, Ge et al. 2008).

		0.01 permutation threshold	% shared	0.001 permutation threshold	% shared
All 3 cell types	Fibroblasts - LCLs - T cells	227	4.5	86	8.5
Exactly 2 cell types	Fibroblasts - LCLs	296	5.9	38	3.8
	Fibroblasts - T cells	270	5.4	35	3.5
	LCLs - T cells	288	5.7	47	4.7
Cell type specific	Fibroblasts	1,353	26.9	268	26.6
	LCLs	1,344	26.7	271	26.9
	T cells	1,261	25.0	262	26.0
Total significant in each cell type	Fibroblasts	2,146		427	
	LCLs	2,155		442	
	T cells	2,046		430	
3 cell type union		5,039	100.0	1,007	100.0
Total tested		15,670		15,670	

Table 13. Cell type-shared and specific gene associations. This table shows gene associations that were: i) shared in all three cell types, ii) shared in two cell types and iii) cell type-specific.

		0.01 permutation threshold	% shared	0.001 permutation threshold	% shared
All 3 cell types	Fibroblasts - LCLs - T cells	170	3.0	79	7.7
Exactly 2 cell types	Fibroblasts - LCLs	231	4.1	35	3.4
	Fibroblasts - T cells	212	3.8	35	3.4
	LCLs - T cells	225	4.0	49	4.7
Cell type-specific	Fibroblasts	1,632	29.1	280	27.1
	LCLs	1,622	28.9	282	27.3
	T cells	1,518	27.1	272	26.4
Total significant in each cell type	Fibroblasts	2,245		429	
	LCLs	2,248		445	
	T cells	2,125		435	
3 cell type union		5,610	100.0	1,032	100.0
Total tested		22,651		22,651	

Table 14. Cell type-shared and specific probe associations. This table shows probe associations that were: i) shared in all three cell types, ii) shared in two cell types and iii) cell type-specific.

The relative sharing of gene associations across cell types is shown in Figure 29 a and Figure 30 (probe associations shown in Figure 29 b). The degree of gene (and probe) sharing is an overestimate of overlapping genetic effects as expression of genes for which eQTLs were identified in all three or at least two cell types is not necessarily driven by the identical regulatory elements.

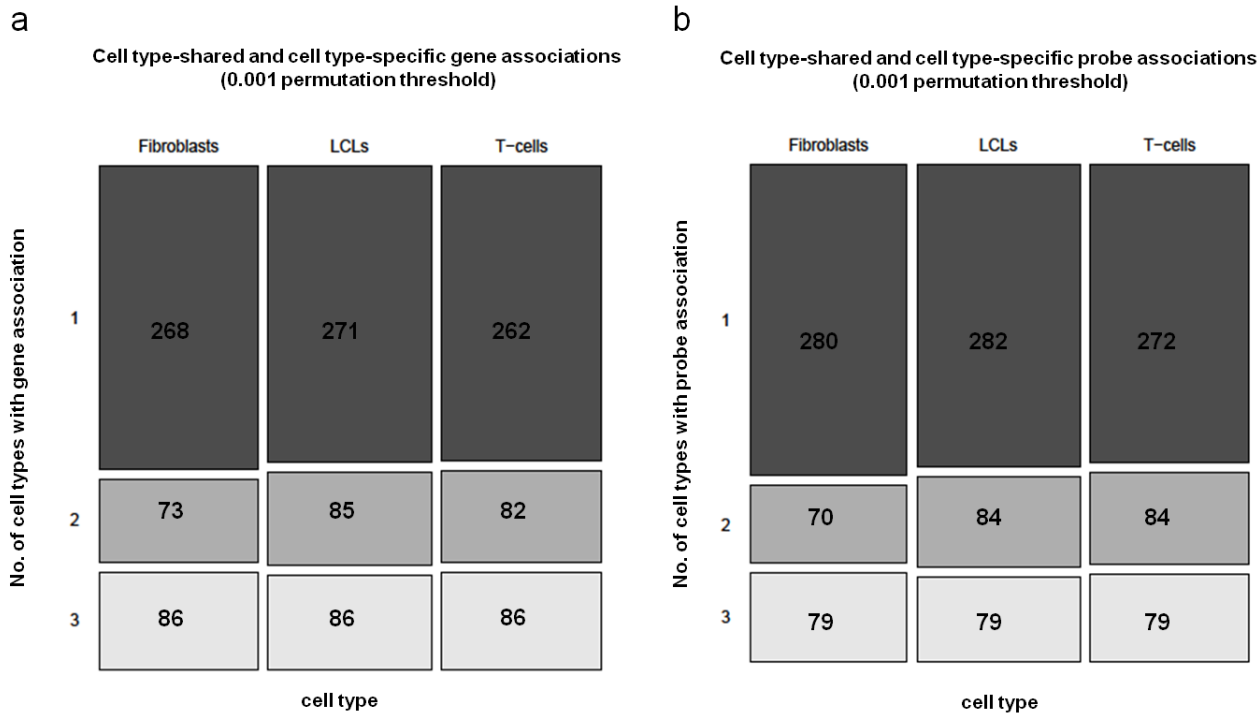

Figure 29. Relative sharing of significant genes and probes in three cell types. Cell type-shared and cell type-specific associations for **a)** genes and **b)** probes (0.001 permutation threshold). Each bar indicates the full fraction of genes or probes for which eQTLs were detected in each cell type. Light grey indicates the fraction of genes/probes with eQTLs overlapping in all three cell types, dark grey indicates the fraction of genes/probes with an overlap in at least one other cell type, and black indicates the fraction of genes/probes with cell type-specific eQTLs.

As expected, a proportion of variation controls expression levels in a similar way across cell types and this most probably reflects regulation of processes common to all cells. At the 0.001 permutation threshold, of the genes with cis eQTLs common to two or more cell types, 124 (12.3%) were shared between fibroblasts and LCLs, 121 (12.0%)

were shared between fibroblasts and T-cells, and 133 (13.2%) were shared between LCLs and T- cells (Table 14). Increased eQTL sharing between LCLs and T-cells is most likely due to the related function and common developmental origin of these cells.

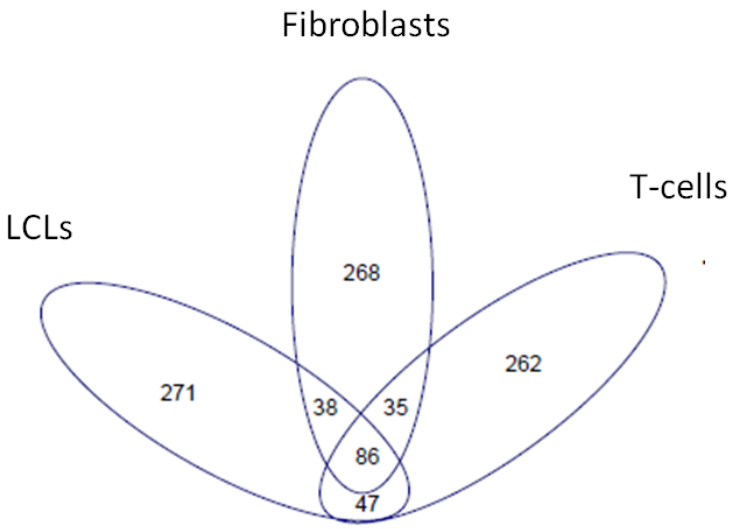

Figure 30. Venn diagram of genes with cis eQTLs in three cell types. Cell type-specific counts, two-way and three-way sharing is shown. Figure by Manolis Dermitzakis.

The most striking result from this analysis is the prominence of cell type specificity. 268 (26.6% of total), 271 (26.9%) and 262 (26.0%) of gene associations were found only in fibroblasts, LCLs and T-cells respectively (Table 13). It is plausible that cell type-specific eQTLs can arise if a gene is expressed in one cell type, but not in another. To test this I explored the medians and variances of gene expression in each cell type, and found that genes with cell type-specific signals had significantly higher expression variance in the cell type where the eQTL was detected (M-W p-value < 0.0001 for all comparisons). Medians of gene expression values for the same genes were either marginally significantly or not significantly different, meaning that all genes included in this analysis were largely expressed in all cell types. Furthermore, it is estimated that all genes with cell type-specific cis eQTLs are expressed to some level in

all three cell types. This suggests that the majority of cell type specificity is not a result of the presence or absence of gene expression between cell types, but is due to differential expression resulting from cell type-specific use of regulatory elements.

5.5 Dissecting eQTL cell type specificity

To dissect the nature of the overlap of cis eQTLs across cell types, I compared the direction of the allelic effect (i.e. assignment of high/low expression to eQTL alleles) between pairs of cell types in cases where SNP-gene associations were significant for both cell types. The direction (sign of Spearman rho) was in complete agreement for all pairwise cell type comparisons at the 0.001 permutation threshold (Figure 31) (99% agreement for 0.01 permutation threshold). This observation implies that regulatory variants are active across cell types in the same manner.

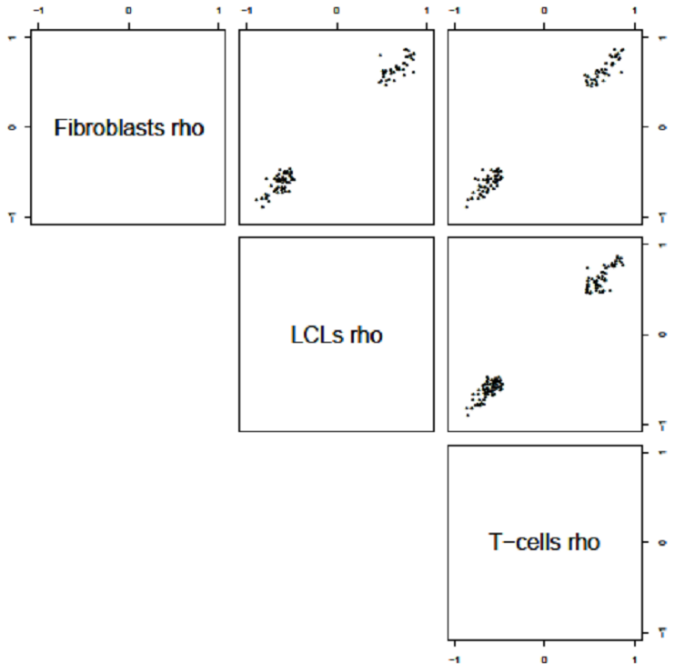

Figure 31. Comparison of the direction of the allelic effect of overlapping SNP-probe associations between pairs of cell types. The plots indicate the value of Spearman rho (effect size) for the same SNP-probe associations between cell types at the 0.001 permutation threshold. In all cases the direction of the allelic effect (indicated by the sign of rho) is the same.

To assess the strength of the cell type specificity observed, I performed RMA on cell types. Cell type specificity is expected to be reflected in the SNP x cell type interaction term, where any cell type-specific association is expected to have a significant interaction term. For cell type-specific eQTLs I found 61 % enrichment of low p-values in RMA (quantified by estimation of FDR (Storey and Tibshirani 2003) (Figure 32). No such enrichment was observed for cell type-shared eQTLs. RMA however is relatively limited in this type of analysis, as the power to detect an interaction term is never maximized. This is because reversal of allelic effect between cell types is not observed.

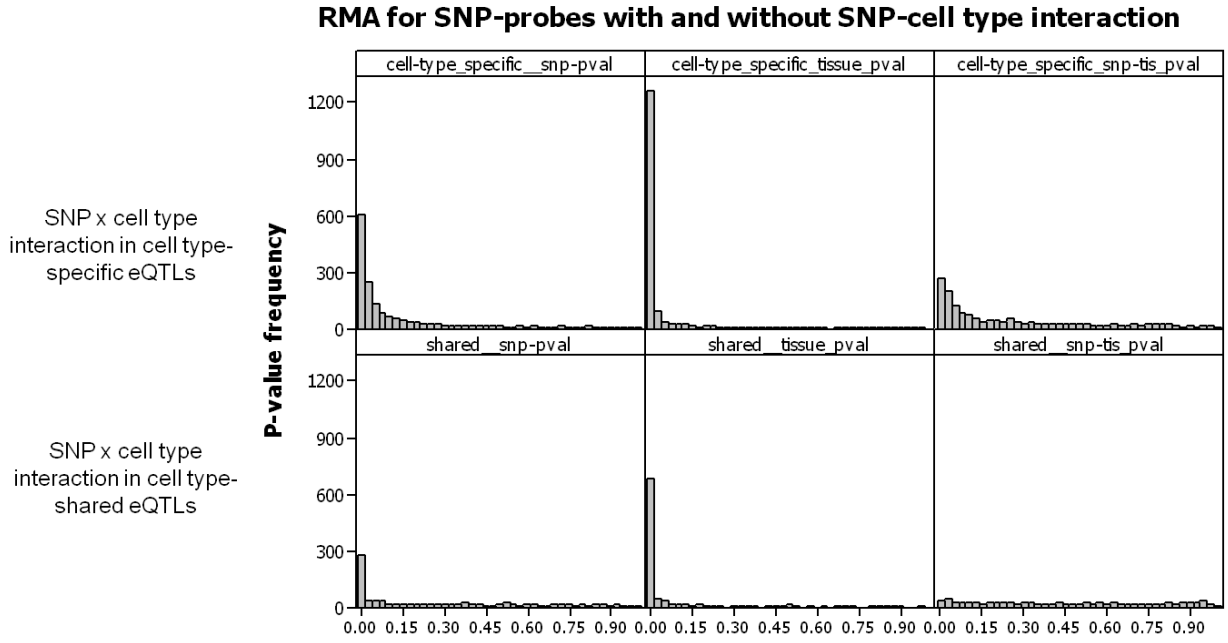

Figure 32. Repeated-measures ANOVA (RMA) to confirm eQTL cell type specificity. RMA association testing (using cell type as the repeated measure) of SNP-probe pairs significant in all three, exactly two and in only one cell type confirmed cell type specificity. Enriched low p-values were observed for SNP-cell type interactions corresponding to those associations that were defined as cell type-specific from the association overlap analysis.

ASE assays were used to validate a subset of cell type-specific eQTLs discovered for genes that also possess transcript SNPs. The ratio of the two transcript alleles was measured in individuals who were double heterozygotes for both the eQTL and the transcript SNP. For 35 transcript SNPs (seven in genes with fibroblast eQTLs, 14 in LCL eQTL genes and 14 in T-cell eQTL genes) extensive allelic imbalance was observed for the cell type in which the eQTL was detected (Figure 33). This imbalance was not observed for ratios of the same eQTL-transcript SNP pairs in the two cell types where the eQTL was not detected (paired t-test p-value = 5.6×10^{-7}). Taken together, these results confirm the signal of cell type specificity statistically and experimentally.

Limited sharing of associations between cell types may arise as a consequence of winner's curse (Goring, Terwilliger et al. 2001; Lohmueller, Pearce et al. 2003; Ioannidis 2008). A cross-threshold assay of sharing revealed that overlapping associations among cell type pairs increased slightly at relaxed significance thresholds for one cell type (Figure 34). Even with relaxed thresholds however over half of associations detected remain cell type-specific.

To further quantify the extent of winner's curse I selected significant SNP-probe pairs from one cell type, and explored their nominal (uncorrected) p-value distribution in the other two cell types. As expected, these distributions were enriched for low p-values, reflecting associations that are shared between cell types (Figure 35). When SNP-probe associations with significant associations in the secondary cell type were removed (i.e. shared associations at the same and at the lower significance threshold), the resulting nominal p-value distributions demonstrated only small enrichment for low p-values.

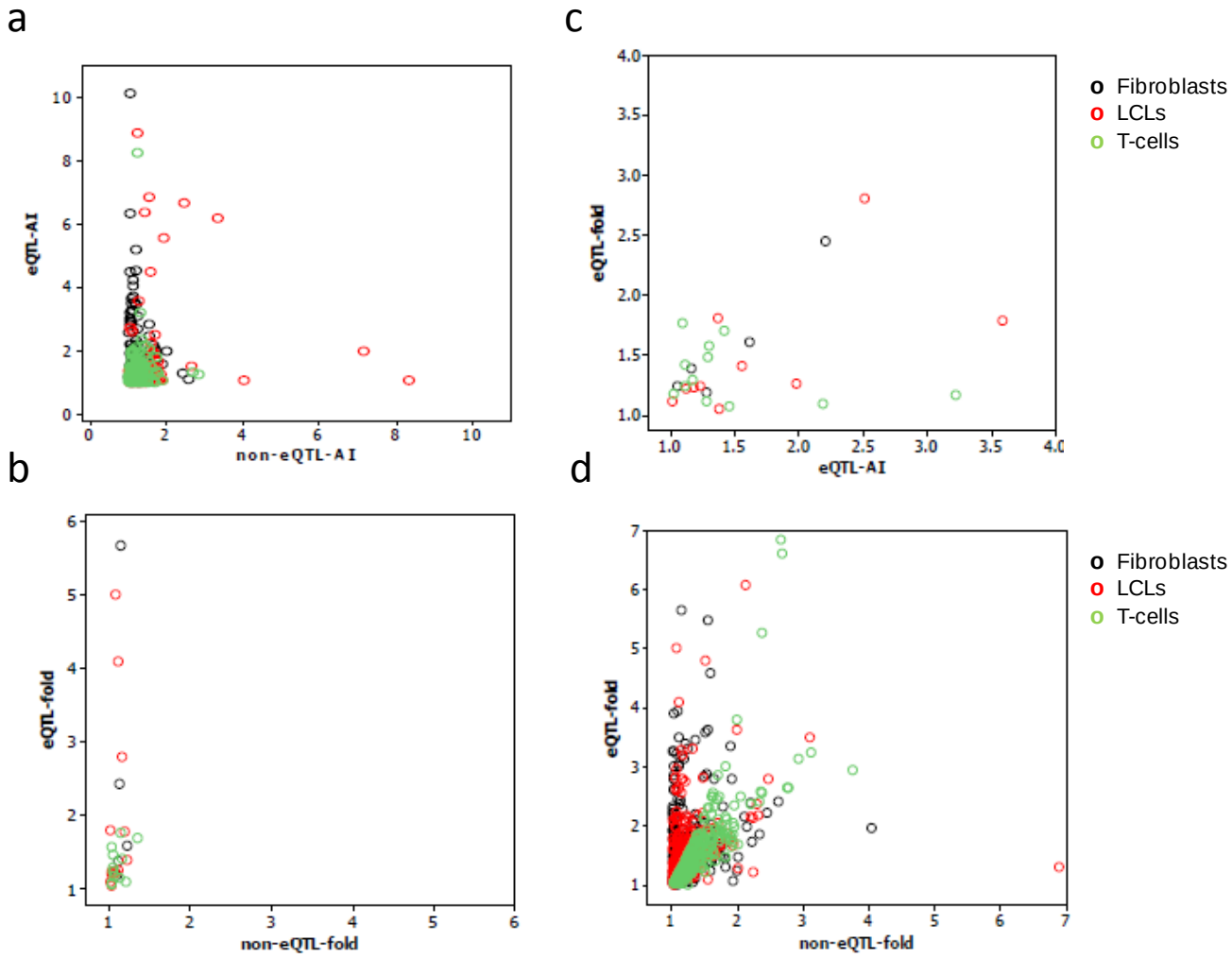

Figure 33. Allele-specific verification of eQTLs. a) Degree of allelic imbalance in double heterozygote individuals (for eQTL and transcript SNP) for 35 assayed transcript SNPs. The y axis shows the ratio of the two alleles in the cell type where the eQTL was initially discovered for each individual, and the x axis shows the mean of the ratio for the other two cell types for each individual. Data points are colour-coded to indicate cell type. The degree of allelic imbalance is more pronounced in the eQTL cell type vs. the non-eQTL cell types. **b)** Fold change difference in expression between the medians of the two homozygote classes of the population for the subset of 35 eQTLs that were confirmed by allele-specific expression (ASE). The plot shows fold change in the eQTLs cell type (y axis) and the non-eQTL cell types (x axis). As expected, the pattern is very similar to the one observed in a). **c)** The fold change estimated from the ratio of homozygotes (y axis) and allelic imbalance (x axis). The correlation is very strong and highly significant (Pearson's correlation coefficient r = 0.685, p-value < 0.0001). **d)** Fold change between the medians of the two homozygote classes of the population for the eQTL cell type (y axis) and the non-eQTL cell types (x axis). As expected the fold change is substantially higher for the eQTL cell type with a mean fold change of 1.55 and a range of 1.07 to 2.65.

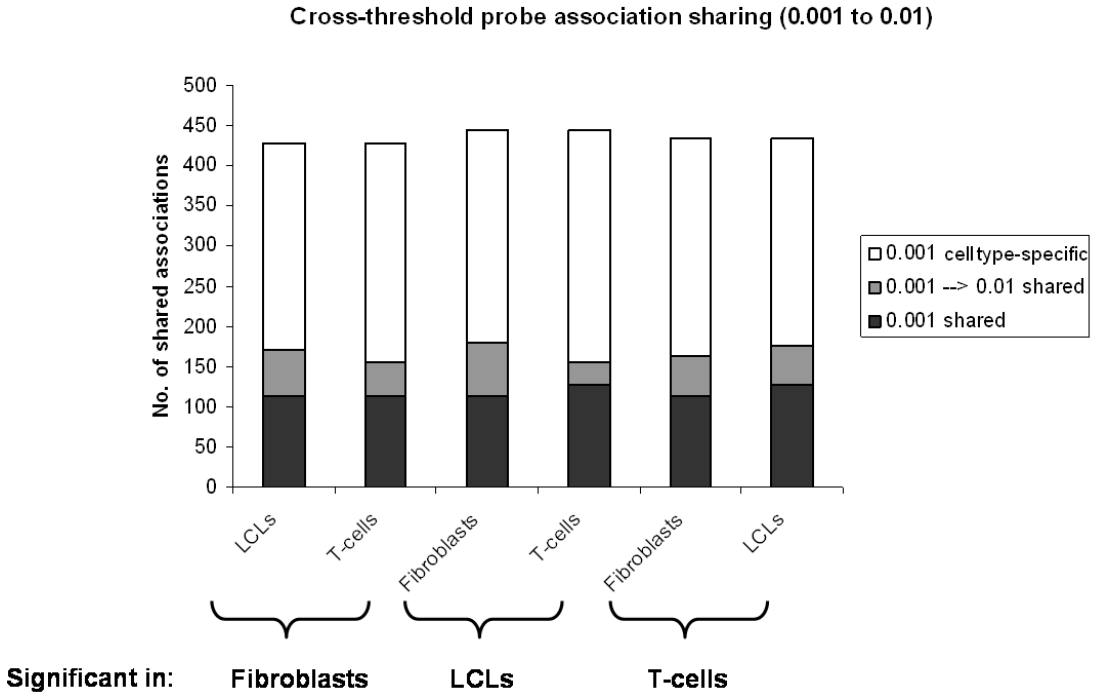

Figure 34. Cross-threshold probe association sharing (exploring the extent of winner's curse). I explored whether association sharing in cell type pairs increases when the significance threshold is relaxed for one cell type. Probe association sharing was found to increase from 28-35% to 40-50% when considering significant associations at the 0.001 permutation threshold in one cell type and the 0.01 permutation threshold in the replication cell type.

I thus quantified the fraction of significant cis eQTLs from one cell type that is not nominally significant (p-value prior to correction > 0.05) in either of the other two cell types. Using this principle of replication, it is estimated that 54%, 50% and 54% of cis eQTLs in fibroblasts, LCLs and T-cells respectively are cell type-specific, amounting to 69% of all cis eQTLs at the 0.001 permutation threshold. Consequently the limited overlap of cis eQTLs between cell types is unlikely to result from winner's curse and a substantial fraction of eQTLs is truly unique to each cell type.

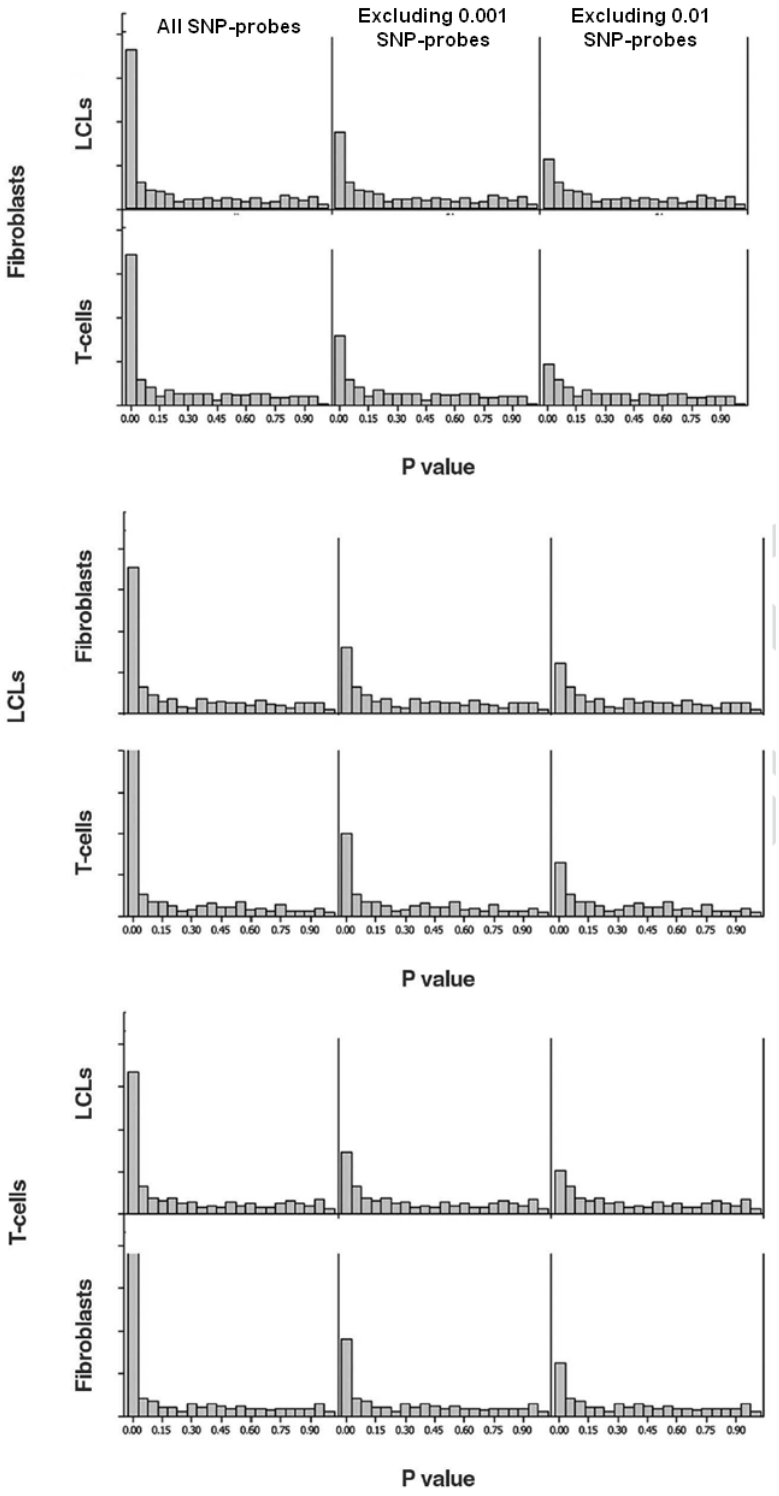

Figure 35. SNP-probe pair nominal (uncorrected) p-value distributions for the two secondary cell types conditional on the reference cell type eQTL (0.001 permutation threshold). The panels on the horizontal axis correspond to secondary cell type p-values for: i) all SNP-probes, ii) excluding SNP-probes significant at the 0.001 permutation threshold and iii) excluding SNP-probes significant at the 0.01 permutation threshold.

5.6 INDEPENDENT EQTLS

Experimental data are accumulating in an effort to annotate the regulatory landscape around genes (Birney, Stamatoyannopoulos et al. 2007). In agreement with previous studies (Stranger, Nica et al. 2007; Veyrieras, Kudaravalli et al. 2008) I found that, on average, the strength and density of cis associations detected for a given gene decay symmetrically with increasing distance from the gene's TSS (Figure 36). As discussed in Chapter 4, the correlated structure of variants within a genomic region due to LD enables association studies as it reduces the number of markers required for testing association with a phenotype, but can impede fine-mapping. The strategy described in 2.7.1 was used to identify independent eQTLs. eQTLs were mapped in recombination hotspot intervals, the most significant eQTL per interval was retained and the least significant eQTL from eQTL pairs with D' < 0.5 between intervals was excluded to derive independently-acting cis eQTLs. At the 0.001 permutation threshold and averaged across three cell types, 5.1% of genes with identifiable eQTLs possess more than one independent interval carrying a significant eQTL (Table 15). In LCLs this number is 4.5% which is comparable to 7.6% of genes with multiple independent eQTLs detected for the HapMap Phase 3 CEU population (Table 10).

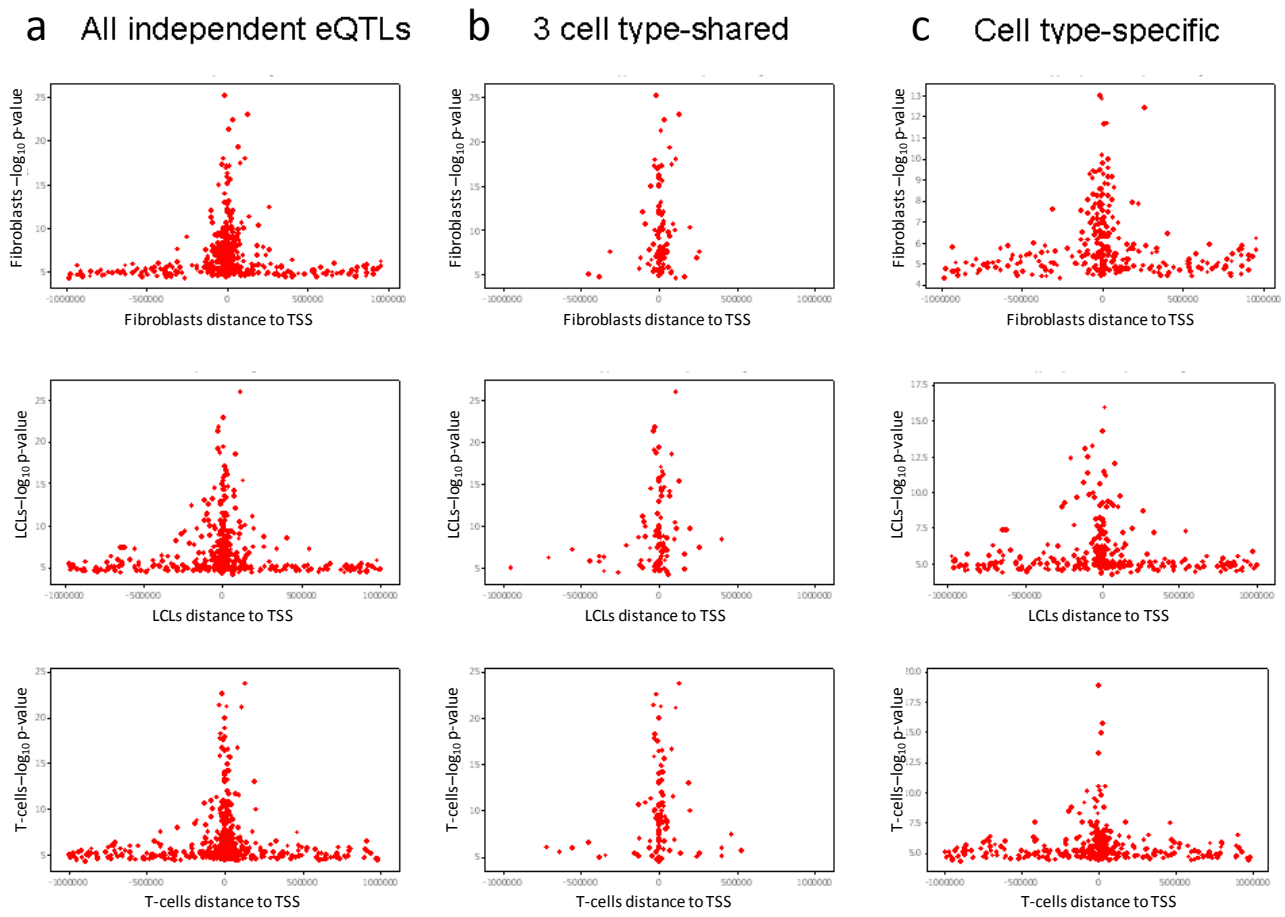

Figure 36. Localization of cis eQTLs (0.001 permutation threshold). a) Distance (in bases) to transcription start site (TSS) of all independent cis eQTLs in each cell type. **b)** Shared cis eQTLs in all three cell types. **c)** Cell type-specific cis eQTLs. Shared cis eQTLs cluster around the TSS whereas cell type-specific cis eQTLs span a wider range of distances from the TSS.

	0.01 permutation threshold			0.001 permutation threshold		
No. of intervals	Fibroblasts	LCLs	T-cells	Fibroblasts	LCLs	T-cells
1	1,875	1,902	1,788	408	422	403
2	237	217	230	18	16	27
3	28	32	27	1	2	0
4	6	3	1	0	1	0
5	0	1	0	0	1	0
Total genes	2,146	2,155	2,046	427	442	430
% genes with ≥ 2 intervals	12.6	11.7	12.6	4.4	4.5	6.3

Table 15. Number of independent cis eQTLs (regulatory intervals) per gene.

This implies that for a fraction of genes and in all cell types considered, multiple cis regulatory variants act to determine expression levels. To further dissect the fine structure of regulatory variant sharing between genes, I repeated the overlap analysis but compared overlap of independent eQTLs (intervals rather than genes) across cell types. When the union of significant genes at the 0.001 permutation threshold was considered, only 6.9% of intervals were found to be shared across all three cell types. 9.7% were shared in exactly two cell types and 83.4% were cell type- specific (Figure 37 a and Table 16). The degree of interval sharing between cell types increases as genes that have shared expression associations in at least two (Figure 37 b and Table 17) and in all three cell types (Figure 37 c and Table 16) are considered.

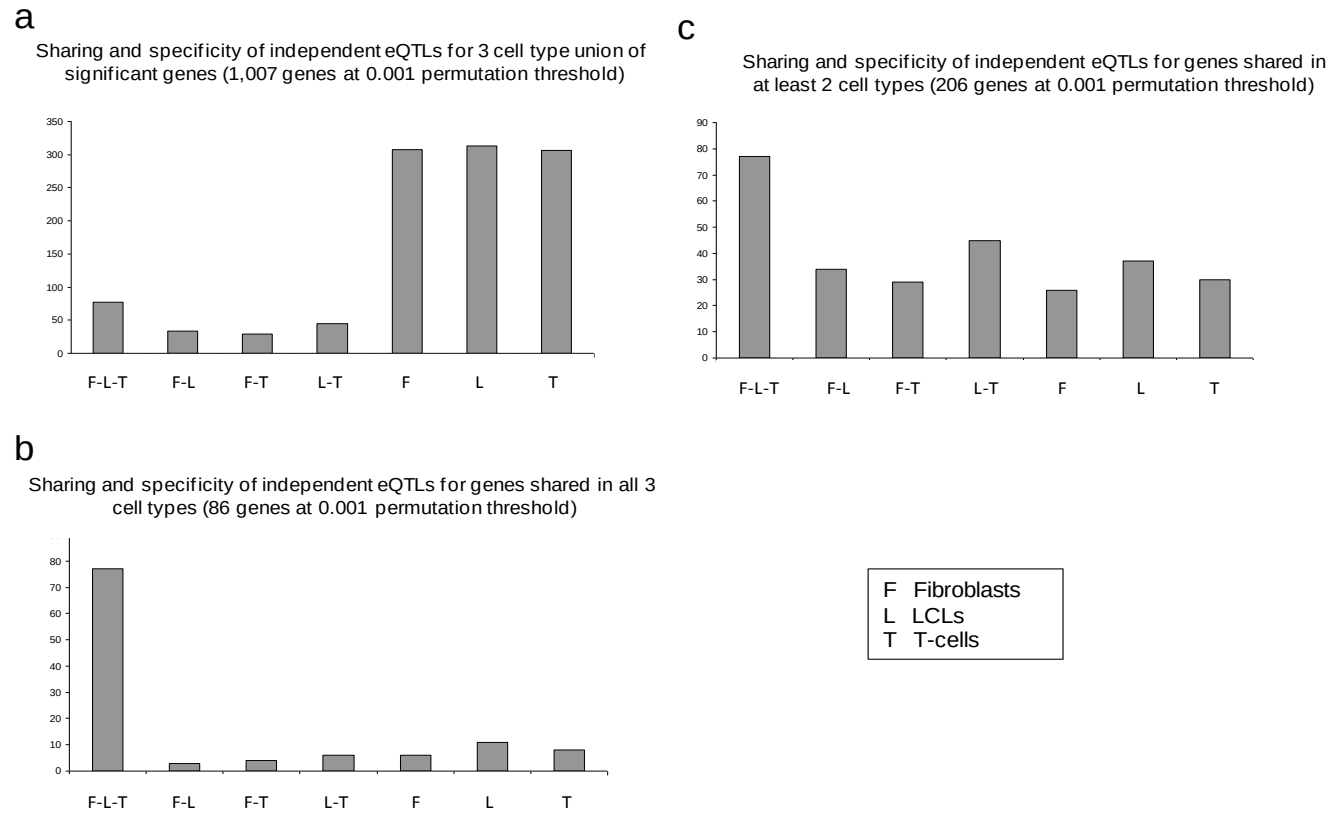

Figure 37. Fine-scale overlap of regulatory signals in three cell types (0.001 permutation threshold). Cell type-shared and specific independent intervals for **a)** the union of genes with a significant association, **b)** genes shared in at least two cell types and **c)** genes shared in all cell types.

In all cases however, there still remains a substantial fraction of cell type-specific independent eQTLs even for genes that had at least one cis eQTL in common in all three cell types.

		Independent eQTL sharing (0.001 permutation threshold)					
		3 cell type union significant genes	%	At least 2 cell type shared genes	%	3 cell type shared genes	%
Genes		1007		206		86	
3 cell type shared independent eQTLs	Fibroblasts - LCLs - T cells	77	6.9	77	27.7	77	67.0
Exactly 2 cell type shared independent eQTLs	Fibroblasts - LCLs	34	3.1	34	12.2	3	2.6
	Fibroblasts - T cells	29	2.6	29	10.4	4	3.5
	LCLs - T cells	45	4.1	45	16.2	6	5.2
cell type specific independent eQTLs	Fibroblasts	307	27.6	26	9.4	6	5.2
	LCLs	313	28.2	37	13.3	11	9.6
	T cells	306	27.5	30	10.8	8	7.0
Total		1111	100.0	278	100.0	115	100.0

Table 16. Independent eQTL (interval) sharing for significant genes (0.001 permutation threshold).

I further evaluated the distribution of independent eQTLs with respect to the TSS and their effect size, conditioning on sharing and specificity across cell types. Cell type-shared eQTLs tend to be of higher significance and larger effect size (Spearman rho) and cluster tightly around the TSS (Figure 36 for significance and Figure 38 for effect size). On the contrary, cell type-specific eQTLs tend to be of lower effect size and are more widely distributed around the TSS (Figure 36). This is in agreement with recent studies (Heintzman, Hon et al. 2009; Visel, Blow et al. 2009) showing that enhancer elements, which are found at greater distances from the gene, are more tissue-specific than basic regulatory elements such as promoters which map close to the TSS. Furthermore, the count of independent eQTLs per gene was significantly correlated with the number of transcripts per gene, for genes with significant cis eQTLs (Pearson's correlation coefficient = 0.049, p-value = 0.117 for the 0.001 permutation threshold, and Pearson's correlation coefficient = 0.105, p-value < 0.0001 for the 0.01 permutation

threshold). This suggests that regulatory complexity is correlated with transcript complexity raising the possibility that some of the regulatory variant signals may mediate genotype-specific choices for alternative TSSs or alternative splicing.

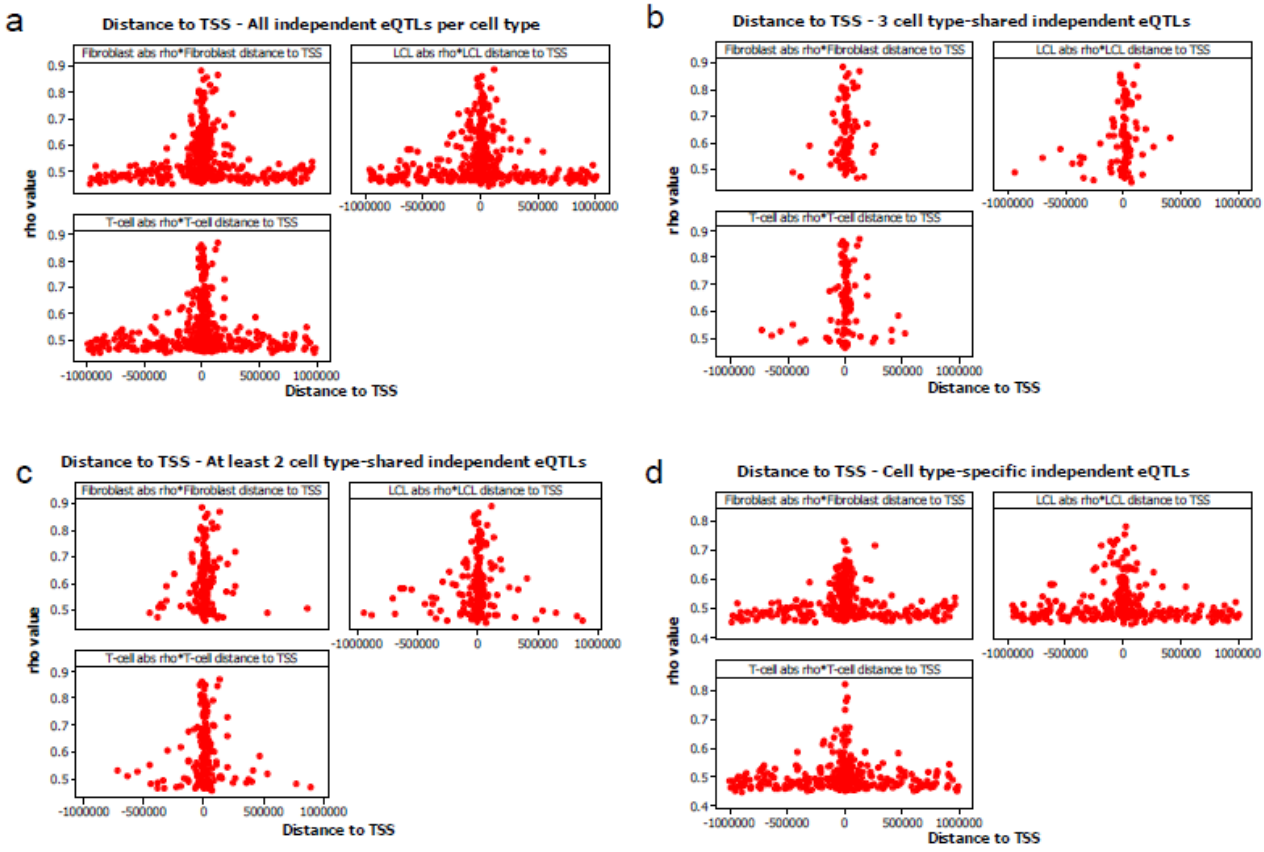

Figure 38. Effect size (Spearman rho) of independent cis eQTLs (0.001 permutation threshold) as a function of the distance (in bases) to a gene's transcription start site (TSS). a) shows all independent cis eQTLs discovered in each cell type, **b)** and **c)** show three cell type-shared and two cell type-shared independent cis eQTLs respectively and **d)** shows independent cis eQTLs specific to one cell type only.

The complexity of the regulatory landscape is illustrated in the case of TSPO, an outer mitochondrial membrane protein of peripheral tissues (Papadopoulos, Baraldi et al. 2006) with a role in cholesterol transport, immunomodulation and apoptosis (Casellas, Galiegue et al. 2002) (Figure 39). At the 0.01 permutation threshold six

independent intervals were identified for this gene: one shared in all three cell types, three fibroblast-specific and two LCL-specific intervals. Additionally, four alternate transcriptional splice variants, encoding different isoforms, have been characterized for this gene and TSPO receptors are found in many tissues of the human body.

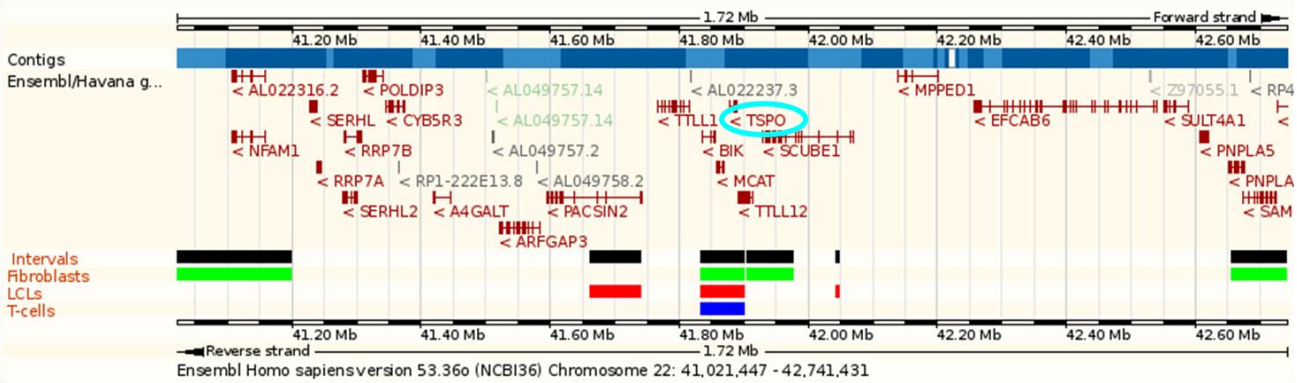

Figure 39. Complex genetic architecture around the *TSPO* gene. *TSPO* (blue oval) encodes for an outer mitochondrial membrane protein with a role in cholesterol transport, immunomodulation and apoptosis. Six independent intervals have been identified for this gene in three cell types: one shared in all three cell types, and 5 cell type-specific. Figure created using the Ensembl genome browser (http://www.ensembl.org).

Regulatory complexity also takes the form of a single independent interval regulating the expression of multiple genes (interval pleiotropy). I explored the number of associated genes per independent interval and found that at the 0.001 permutation threshold over 6% of intervals are associated with the expression of more than one gene (this number increases to almost 19% at the 0.01 permutation threshold). An example of a single eQTL influencing eight genes is shown in Figure 40. In such cases it may be interesting to explore whether the genes influenced by a common regulatory interval are components of the same pathway or network. The multidimensionality caused by cell type specificity, regulatory region promiscuity and genetic variation highlight the challenges to be faced when a wider range of conditions and context-dependent effects (cell types, tissues, developmental stages) are considered.

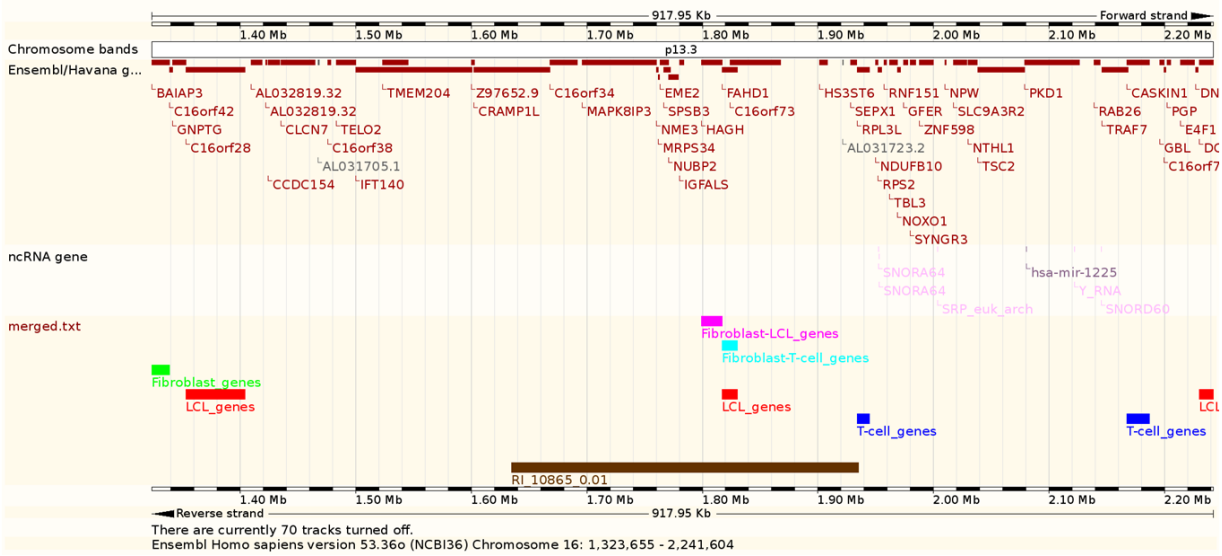

Figure 40. A single independent eQTL in the brown interval (denoted at RI_10865) affects the expression of a total of 8 genes (0.01 permutation threshold). RI_10865 affects the expression of three LCL-specific, two T-cell specific and one fibroblast-specific genes, as well as a gene with a significant association shared in fibroblasts and LCLs and a gene with a significant association shared in fibroblasts and T-cells (find genes). Figure created using the Ensembl genome browser.

5.7 BIOLOGICAL PROPERTIES OF SHARED AND CELL TYPE-SPECIFIC EQTLS

Gene Ontology (GO) (Ashburner, Ball et al. 2000) terms were used to compare biological properties of cell type-specific and shared genes. For cell type-specific associations, I detected an over-representation of properties linked to signal transducer activity, cell communication, development, behaviour, cellular process, enzyme regulator activity, transcription regulator activity and response to stimulus, reflecting properties likely to sculpt cell type-specific profiles. For associations shared in all cell types I found an over-representation of catalytic activity and transport properties (Fisher's exact test p-value < 0.05) (Table 17).

GO_Slim_Description	F_vs_FLT	results.ratio	results.pvalue	L_vs_FLT	results.ratio	results.pvalue	T_vs_FLT	results.ratio	results.pvalue
nucleic acid binding	GO:0003676	0.95	0.74	GO:0003676	1.56	0.00	GO:0003676	1.68	0.00
motor activity	GO:0003774	Inf	1.00	GO:0003774	Inf	0.35	GO:0003774	Inf	1.00
catalytic activity	GO:0003824	0.58	0.00	GO:0003824	0.76	0.00	GO:0003824	0.76	0.00
signal transducer activity	GO:0004871	4.15	0.00	GO:0004871	2.04	0.00	GO:0004871	1.39	0.03
structural molecule activity	GO:0005198	0.74	0.28	GO:0005198	0.41	0.01	GO:0005198	0.99	1.00
transporter activity	GO:0005215	0.77	0.12	GO:0005215	1.14	0.51	GO:0005215	1.14	0.47
binding	GO:0005488	0.93	0.21	GO:0005488	0.78	0.00	GO:0005488	0.98	0.74
electron transport	GO:0006118	0.00	0.01	GO:0006118	1.38	0.77	GO:0006118	0.00	0.01
nucleoside, nucleotide, nucleic acid metabolism	GO:0006139	0.60	0.00	GO:0006139	1.10	0.33	GO:0006139	0.72	0.00
amino acid and derivative metabolism	GO:0006519	1.42	0.42	GO:0006519	2.32	0.02	GO:0006519	1.65	0.22
transport	GO:0006810	0.21	0.00	GO:0006810	0.39	0.00	GO:0006810	0.31	0.00
cell motility	GO:0006928	20.16	0.00	GO:0006928	12.18	0.00	GO:0006928	3.44	0.33
membrane fusion	GO:0006944	0.06	0.00	GO:0006944	0.00	0.00	GO:0006944	0.66	0.51
cell communication	GO:0007154	2.28	0.00	GO:0007154	2.68	0.00	GO:0007154	1.81	0.00
development	GO:0007275	2.38	0.00	GO:0007275	1.53	0.01	GO:0007275	2.13	0.00
physiological process	GO:0007582	0.98	0.73	GO:0007582	0.85	0.01	GO:0007582	1.02	0.73
behavior	GO:0007610	Inf	0.00	GO:0007610	Inf	0.01	GO:0007610	Inf	0.00
pathogenesis	*	*	*	*	*	*	GO:0009405	Inf	0.36
cellular process	GO:0009987	5.80	0.00	GO:0009987	3.24	0.04	GO:0009987	7.65	0.00
antioxidant activity	GO:0016209	Inf	0.00	*	*	*	*	*	*
chaperone regulator activity	*	*	*	GO:0030188	Inf	0.57	*	*	*
enzyme regulator activity	GO:0030234	4.54	0.00	GO:0030234	6.29	0.00	GO:0030234	2.35	0.03
transcription regulator activity	GO:0030528	7.17	0.00	GO:0030528	6.63	0.00	GO:0030528	4.78	0.00
translation regulator activity	*	*	*	*	*	*	GO:0045182	Inf	1.00
regulation of biological process	GO:0050789	Inf	0.00	GO:0050789	Inf	0.02	GO:0050789	Inf	0.14
response to stimulus	GO:0050896	2.79	0.00	GO:0050896	2.42	0.00	GO:0050896	3.33	0.00

over-represented
under-represented

Table 17. GO Slim term comparison for cell type-specific vs. three cell type-shared genes (0.001 permutation threshold). Fisher's exact test significant p-value < 0.05. Biological properties over-represented in cell type-specific genes are shown in red and include signal transducer activity, cell communication, development, behaviour, cellular process, enzyme regulator activity, transcription regulator activity, and response to stimulus. Biological properties under-represented in cell type-specific genes are shown in blue and include catalytic activity and transport.

Entropy of expression for each gene was calculated as an indication of cell type specificity, with lower entropy values reflecting higher specificity. I used data from cell types (tissues) included the GNF/Novartis atlas of gene expression (Su, Wiltshire et al. 2004) (Table 18) and compared entropy between genes with shared vs. cell type-specific cis eQTLs. Genes with fibroblast-specific eQTLs showed consistently and significantly lower entropy values (i.e. were more cell type-specific) compared to shared associations (M-W p-value = 0.0047). This signal was in the same direction, but less prominent, for the other two cell types. This may be due to the fact that fibroblasts are biologically more distant to LCLs or T-cells, or to potential tissue sampling biases in the GNF/Novartis collection.

tissue	description
adipocyte	fat
adrenal cortex	perimeter of the adrenal gland
adrenal gland	endocrine glands on kidneys
amygdala	groups of neurons located within medial temporal lobes of brain
appendix	part of digestive system, blind-ended tube connected to cecum
bone marrow	tissue in the hollow interior of bones, produces new blood cells
bronchial epithelial	lung epithelium
caudate nucleus	nucleus located in basal ganglia of brain, role in learning and memory
cerebellum peduncles	region of brain, role in the integration of sensory perception
ciliary ganglion	parasympathetic ganglion located in the posterior orbit
dorsal root ganglion	nodule on dorsal root (afferent sensory root of spinal nerve)
heart	heart
hypothalamus	small nuclei in brain linking nervous to endocrine system, located above brain stem
kidney	kidney
liver	liver
lung	lung
lymph node	organ consisting of multiple cell types, part of the lymphatic system
ovary	ovary
pancreas	pancreas
pituitary	pituitary
prostate	prostate
salivary gland	salivary gland
skeletal muscle	skeletal muscle
skin	skin
smooth muscle	smooth muscle
spinal cord	spinal cord
superior cervical ganglion	largest of the cervical ganglia, supplies sympathetic innervation to the face
testis	testis
thymus	thymus
thyroid	thyroid
tongue	tongue
tonsil	tonsil
trachea	trachea
trigeminal ganglion	sensory ganglion of the trigeminal nerve (5th cranial nerve)
uterus	uterus
uterus corpus	endometrium
whole blood	whole blood
whole brain	whole brain

Table 18. Tissues used for entropy calculation (GNF/Novartis atlas of gene expression).

5.8 CONCLUSIONS

This study provides a direct comparison of the impact of regulatory variants in a cell type-dependent context. Having controlled for all other confounders such as experimental design, sampling variance and differences in technology, I have demonstrated that variants affecting gene regulation largely act in a cell type-specific manner, and even cell types as closely related as LCLs and T-cells share only a minority of their cis eQTLs. Based on the three cell types tested, it is estimated that 69-80% of regulatory variants are cell type-specific. Regulatory variant complexity correlates with

transcript complexity suggesting genotype-specific effects on alternative transcript choice. In addition, cell type-specific eQTLs are of smaller effect size and tend to localize at greater distances from the TSS recapitulating enhancer element distributions. Importantly, the signal of cell type specificity is primarily due to differential use of regulatory elements of genes that are expressed in almost all cell types. This analysis is also the first to demonstrate robust replication of eQTLs in LCLs between samples collected and transformed decades apart. This is of great importance for the field of human genetics since a large number of cohorts have collections of LCLs whose value has been debated and questioned repeatedly. I argue that LCLs are likely to represent a legitimate biological system that can be used for disease interpretation or other functional studies with all the limitations of cell line specificity. As more tissues are interrogated diminishing returns in discovery of eQTLs are expected, and it is possible that there is a minimum set of tissues that will be informative for the vast majority of regulatory variants. Nevertheless, this study highlights the need for deep and wide interrogation of regulatory variation in multiple cell types and tissues in order to elucidate their differential functional properties. The pattern of cell type specificity is not expected to be limited to regulatory variants, but is likely to apply to protein-coding and other putative functional variants (e.g. epigenetic modifications).