4 FINE-SCALE ARCHITECTURE OF THE CIS REGULATORY LANDSCAPE

In this chapter I will:

- Discuss that LD is a useful property of the genome for association studies at the large scale, but that it can impede the fine-mapping of functional variants.
- Outline a number of approaches employed to enable localization and identification of functional variants.
- Present a strategy used to scan all cis eQTLs detected for a given gene and to identify those that tag independent effects on gene expression.
- Describe the genetic architecture of the cis regulatory landscape and show that multiple regulatory elements can interact to regulate expression in cis.

4.1 FROM GENOME-WIDE ASSOCIATION HITS TO FUNCTIONAL VARIANTS

The power of a SNP to show association with a phenotype is related to its correlation coefficient with the causal variant (Ioannidis, Thomas et al. 2009). This correlated structure of variants in the genome has made it possible to carry out GWAS and identify a plethora of associations between genetic variants and complex traits. However, the variants discovered are not necessarily the ones that give rise to phenotypes, but are more likely tags of functional drivers. Furthermore, when a locus is identified by SNP association, the causal mutation itself need not be a SNP (Altshuler, Daly et al. 2008). For example variants in the *IRGM* gene were found to be associated with CD, but subsequent analysis indicated that the causal mutation is most likely a

deletion upstream of the promoter affecting tissue-specific expression (McCarroll, Huett et al. 2008).

Using GWAS-detected regions as a starting point, the field is currently focusing on strategies for the localization and identification of true functional variants. It is only when these variants are discovered that it will be possible to piece together the biological pathways and processes sculpting complex traits and disease risk. Fine-mapping and identification of functional variants is not an easy task as the correlated structure between variants can impede fine-mapping, with patterns of LD determining the number of markers required to detect and fine-map an association (Mackay, Stone et al. 2009). If a group of markers is in high LD, it is only necessary to genotype one of them as a proxy for all others in the LD block. In pure breeds of dogs for example, where LD blocks are large, only a few markers are required to detect candidate regions. However it is not possible to localize functional variants precisely using this approach (Sutter and Ostrander 2004). In species such as *Drosophila*, LD declines rapidly over short physical distances and knowledge of all sequence variants is necessary for association mapping (Carbone, Jordan et al. 2006), but localization of variants with an impact on the phenotype is precise. Given the extent of LD in humans, genetic variants are likely to have a number of close proxies (Slatkin 2008). A detailed survey of 5 Mb of the human genome (Encyclopedia of DNA Elements or ENCODE regions) genotyped and sequenced in HapMap individuals, revealed that over half of all common SNPs have at least 10 other SNPs in their proximity with an $r^2 > 0.8$ (International HapMap Consortium 2005).

Fine-mapping established associations involves selecting a set of non-redundant SNPs that are in perfect, or near perfect correlation (Ioannidis, Thomas et al. 2009). The rationale behind this approach is that one of the variants selected is the functional driver of the phenotype. Consequently, fine-mapping requires detailed knowledge of variation. Currently the most complete catalogue of human genetic variation is the

HapMap Phase 2, (four million SNPs genotyped for four geographically distinct populations), which covers roughly 30% of common variants. A much more detailed assay of variation will be provided by the ongoing 1000 Genomes Project which involves sequencing the genomes of 1,000 individuals (http://www.1000genomes.org). Deeper sequencing will subsequently reveal rarer variants (International HapMap Consortium 2005).

GWAS interrogating regulatory variation are also faced with the same issues when it comes to localization of functional variants. Association studies of SNP genotypes with transcript levels reveal that for most genes multiple cis eQTLs exist (Stranger, Nica et al. 2007; Dimas, Stranger et al. 2008; Dimas, Deutsch et al. 2009). In such cases, it is likely that most variants mapping to the same genetic locus and are in high LD do not tag independent regulatory effects. On the contrary, SNPs with promising association signals are those that are not in LD and are expected to contribute independent effects to the phenotype of interest (Ioannidis, Thomas et al. 2009). Single loci however may harbour multiple independent functional variants, as is the case of chromosome 8q24 which contains seven independent risk alleles for prostate cancer (Haiman, Patterson et al. 2007).

4.2 NARROWING DOWN THE REGION OF INTEREST

Mapping eQTLs has two components: detection and localization (Mackay, Stone et al. 2009). eQTL detection depends on effect sizes and allele frequencies and delimits a broad genomic region harbouring regulatory elements. Localization or fine-mapping of eQTLs depends on the recombination frequency between regulatory elements and markers. Many approaches have been employed to fine-map eQTLs mostly by narrowing down the region likely to harbour the regulatory variant. In general, the smaller the space outlined by significant associations, the narrower the region that has

to be surveyed for variation, although this can be complicated by local patterns of LD, population history and non-genetic factors. Despite all this, one of the ways forward is to make use the properties of the genome (e.g. information about recombination hotspot intervals) and integrate data from various fields to limit the size of the genomic space to be scanned. Some approaches employed thus far are discussed below.

One approach, employed by Veyrieras et al (2008) who studied HapMap Phase 2 LCLs, involved taking the position of the most significant SNP as an estimate of the location of the functional site. The authors point out that this is only a rough proxy and that these SNPs are unlikely to be true functional variants since: a) HapMap Phase 2 contains only about a third of common SNPs, b) some significant SNP associations may arise if the SNP is in LD with CNVs and c) non-functional SNPs in strong LD with the causal SNP may have lower p-values just by chance. A Bayesian hierarchical model incorporating information about the physical location of SNPs, as well as SNP functional annotation was used to create a high-resolution map of cis regulatory variation. Thirty three percent of most significant eQTLs were found to map within 10 kb of the TSS, and immediately upstream of transcription end site (TES). The former are likely to be polymorphisms that affect the strength of TF binding sites and influence the rate of transcription. The latter may have an impact on microRNA binding and subsequent transcript degradation. eQTLs were also found to be more frequent in exons compared to introns, suggesting that these polymorphisms may affect transcript stability or rate of degradation.

Another study interrogating cis regulatory variation employed allelic expression to measure the relative expression of alleles within a sample, assaying both primary (unspliced) transcripts and mRNA (Pastinen and Hudson 2004). This approach yields direct (vs. statistically inferred) relationships between SNPs and cis regulatory differences (Verlaan, Ge et al. 2009), but does not detect differences in transcript levels driven by variants unlinked to the primary transcript. Allelic expression screening in

LCLs and primary osteoblasts revealed that even for genes that were expressed in both tissue types, identical haplotypes exerted different effects in ~ 50% of the cases. Therefore the same haplotype can display different regulatory effects depending on the tissue it is acting in. (Note that in this study each tissue type originated from one of two populations. Both populations however were of Northern European origin).

A third study investigated the relationship between expression levels of 4,200 genes and proportion of European ancestry in LCLs from African American individuals (Price, Patterson et al. 2008) who inherit variable proportions of African and European ancestries. It was shown that expression differences in individuals of different ancestry proportions reflect expression differences between African and European populations. Using information on an individual's ancestry at the location of a gene whose expression was being analysed, ancestry effects were employed to quantify the relative contributions of cis and trans regulation of human gene expression. The authors estimated that 12 ± 3% of all heritable variation in human gene expression is due to cis variants. However, as they point out, distinction between cis and trans was somewhat imprecise due to the extended length (> 10Mb) of segments of continental ancestry in African Americans.

The examples above illustrate that association analyses testing marker panels cannot differentiate causal SNPs from proxies. Identifying causal variants will be aided by obtaining a more complete catalogue of genetic variation (e.g. 1000 Genomes), but also by cataloguing variants with a functional role on a genome-wide scale. This is the aim of the ENCODE Project (Birney, Stamatoyannopoulos et al. 2007), whose ultimate goal is to find all functional elements in the genome across different cell types. In its pilot phase, a number of techniques were employed to analyse 1% (30Mb) of the human genome.

With the same aim in view, two recent studies focused on identifying regulatory elements across the genome. In the first study Heintzman et al (2009) used a chromatin

immunoprecipitation (ChIP)-based microarray method to identify promoters, enhancers and insulators in multiple cell types and investigate their role in cell type-specific gene expression. Over 55,000 potential transcriptional enhancers were identified, marked with highly cell type-specific histone modification patterns. The patterns detected correlated strongly to cell type-specific gene expression programmes on a global scale and were functionally active in a cell type-specific manner. In contrast, the chromatin state at promoters, as well as binding of CTCF (a major protein involved in insulator activity), were largely invariant across diverse cell types. The second study used *in vivo* mapping of p300 binding to identify regulatory sequences that control the spatial and temporal expression of genes (Visel, Blow et al. 2009). p300 is a near-ubiquitously expressed transcriptional co-activator and a component of enhancer-associated protein assemblies. ChIP of p300, followed by massively parallel sequencing led to mapping of several thousand p300 binding sites in mouse embryonic forebrain, midbrain and limb tissue. Eighty six of the identified sequences were tested in a transgenic mouse assay and enhancer activity was detected in nearly all cases.

In this study I dissected the fine-scale architecture of the cis regulatory landscape using eight of the eleven HapMap Phase 3 populations. I designed and applied a strategy to filter all cis eQTLs detected for a given gene and identify those that tag independent regulatory elements. I also explored the extent to which pairs of interacting variants shape expression levels in cis to highlight the complexity and multidimensionality of gene regulation. At the time of writing this work was in preparation for publication.

4.3 HAPMAP PHASE 3 CIS EQTLS

With the availability of additional populations, as well as additional individuals per population, HapMap Phase 3 provides greater power for eQTL detection within and

across populations. SRC was used to test for association in cis between SNP genotypes (of approximately 1.2 million SNPs per population) and transcript levels of 18,226 Ensembl genes, independently in each population and considering only unrelated individuals (Table 8). All SNPs mapping in a 2 Mb window, centred on the TSS of genes were tested and correction for significance was through permutations. Gene expression for GIH, LWK, MEX and MKK was PCA-corrected and analysed against non-PCA-corrected genotypes (see section 2.3.2.1). This work was carried out in collaboration with Barbara Stranger and Stephen Montgomery at the WTSI.

	SNPs	Probes	Total tests in cis
CEU	1,223,718	21,800	26,690,513,298
CHB	1,115,926	21,800	24,339,461,986
GIH	1,174,223	21,800	25,598,061,400
JPT	1,096,051	21,800	23,905,968,361
LWK	1,249,643	21,800	27,242,217,400
MEX	1,163,286	21,800	25,359,634,800
MKK	1,284,097	21,800	27,993,314,600
YRI	1,306,038	21,800	28,485,994,818

Table 8. HapMap Phase 3 SNPs, probes and total association tests performed in cis.

At the 0.01 permutation threshold of significance, roughly 180 genes are expected to have one significant association by chance. We detected 657, 774, 698, 795, 773, 472, 947 and 799 genes in CEU, CHB, GIH, JPT, LWK, MEX, MKK and YRI populations respectively (estimated FDR of 20-40%) (Table 9). From a non-redundant union of 3,130 gene associations (18% of all genes tested) 1,074 (34%) were shared in at least two populations and 63 (2%) had a significant association in all eight populations.

	0.01 permutation threshold	
	genes	FDR
CEU	657	0.28
CHB	774	0.24
GIH	698	0.26
JPT	795	0.23
LWK	773	0.24
MEX	472	0.39
MKK	947	0.19
YRI	799	0.23
Non-redundant	3,130	
> 2 populations	1,074	
8 populations	63	

Table 9. HapMap Phase 3 cis significant gene associations.

To explore the location and strength of cis eQTLs for each of the eight populations, the distance of the most significant cis eQTL per gene was mapped relative to the TSS. In agreement with previous studies (Stranger, Nica et al. 2007; Veyrieras, Kudaravalli et al. 2008) a strong signal was found close to the TSS, with no discernable trend in a 5′ or 3′ direction (Figure 20). This symmetrical trend has also been documented in the analysis of the ENCODE Consortium (Birney, Stamatoyannopoulos et al. 2007) and is likely to reflect variation in core regulatory sequences such as promoter elements.

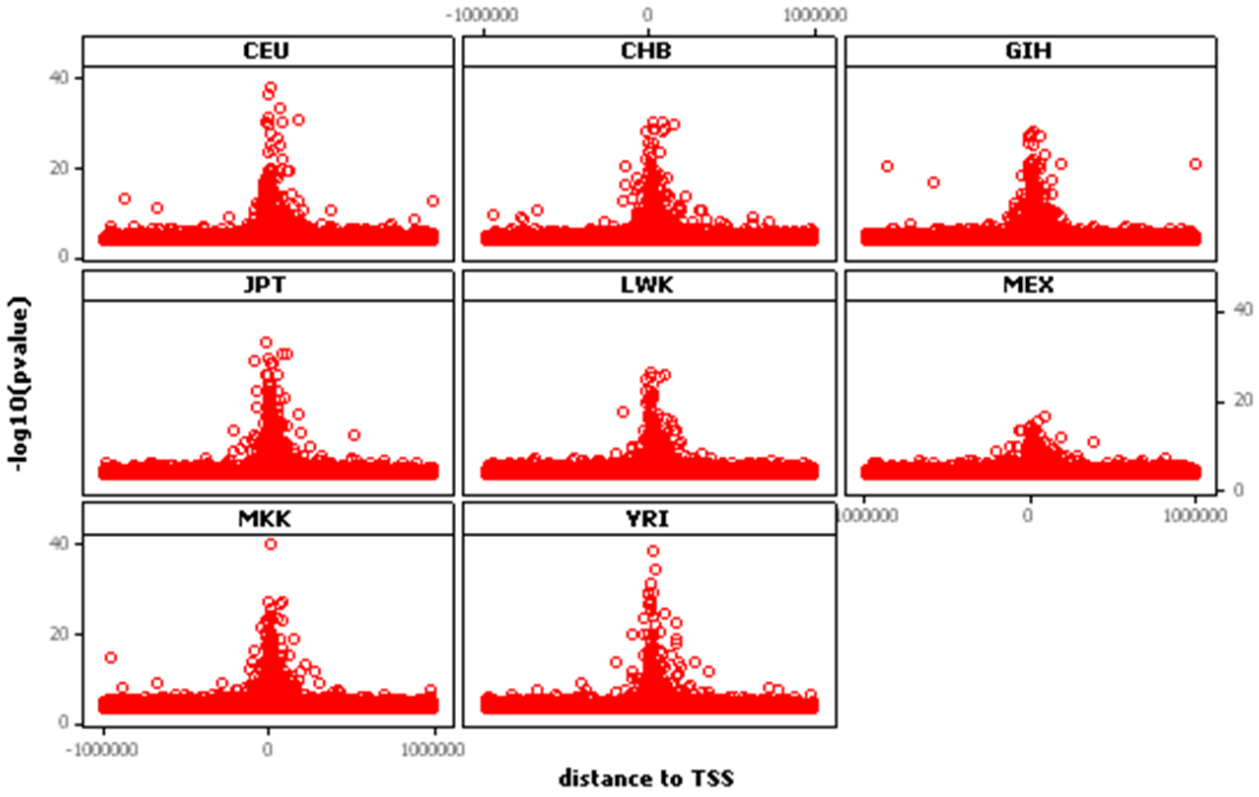

Figure 20. Distance (in bases) of the most significant cis eQTL per gene to the transcription start site (TSS). For HapMap Phase 3 populations (0.01 permutation threshold) the strength and abundance of cis eQTLs decrease with increasing distance from the TSS.

4.4 INDEPENDENT REGULATORY INTERVALS

Over half of the genes with a significant association at the 0.01 permutation threshold possess more than one SNP with a significant association in each of the eight populations (Figure 21 a and Figure 21 b). Multiple eQTLs identified for a given gene most probably tag the effects of the same regulatory element. Gene regulation however is dependent on the joint action of multiple regulatory elements (Figure 22) and the aim of this study was to identify cis eQTLs that tag independent regulatory effects (independent eQTLs or regulatory intervals).

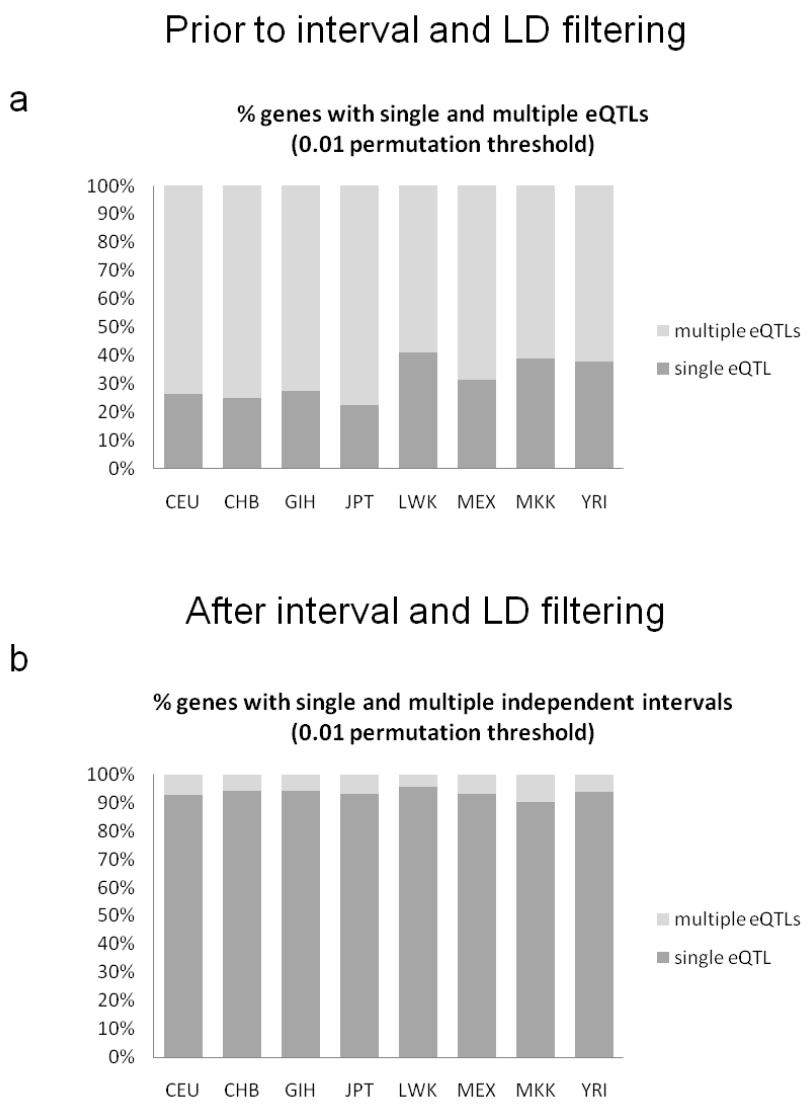

Figure 21. Percent of genes with multiple cis eQTLs and independent intervals. a) shows the % of genes possessing multiple cis eQTLs prior to recombination hotspot interval mapping and LD filtering and **b)** shows the % of genes possessing multiple independent cis eQTLs (intervals) (0.01 permutation threshold).

A detailed description of the strategy employed to do this has been given in section 2.7.1. Briefly, for a given gene eQTLs were mapped in recombination hotspot intervals, the most significant eQTL per interval was retained and remaining eQTLs were filtered further to exclude the least significant variant from variant pairs with a D' > 0.5. This rigorous filtering strategy ensures that surviving eQTLs tag the effects of

independent regulatory elements. Furthermore, since filtering is strict the count of independent cis eQTLs most likely represents the lower bound of the true number of regulatory elements controlling the expression of genes. As expected, the number of cis eQTLs detected for each gene after filtering is much lower (Figure 21 c and Figure 21 d).

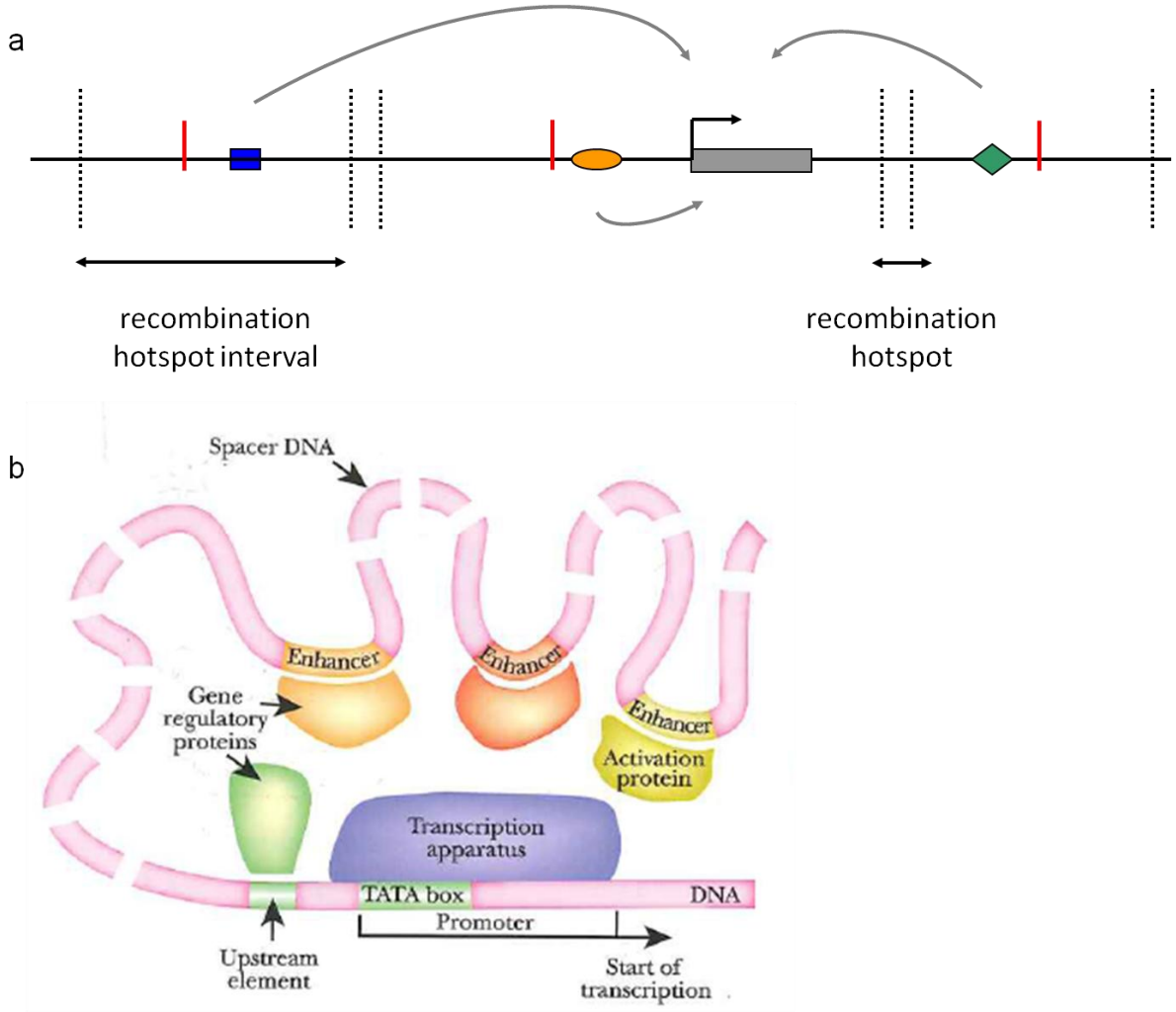

Figure 22. Multiple independent regulatory elements control gene expression. a) Regulatory elements interact with each other to control levels of transcription. In this example independent regulatory elements (with variation in the population) are shown in blue, orange and green and map in different recombination hotspot intervals. The red bars represent SNPs tagging the effects of these elements. **b)** The action of multiple elements controls transcription initiation. Folding of DNA allows numerous activators bound to enhancer sequences to make contact with the basal transcription complex. From (Clark 2005).

At the 0.01 permutation threshold the number of genes possessing multiple independent intervals ranged from 5-10% across the eight populations (Table 10). Specifically, 50 genes with multiple eQTLs (8% of all genes tested), 46 (6%), 44 (6%), 55 (7%), 36 (5%), 34 (7%), 97 (10%) and 52 (7%) were detected for the CEU, CHB, GIH, JPT, LWK, MEX, MKK and YRI populations respectively. Taken together, multiple independent regulatory intervals were detected for approximately seven percent of genes. This observation is in agreement with a mechanism for gene regulation involving the coordinated action of multiple elements (Figure 22).

	0.01 permutation threshold							
No. of intervals	CEU	CHB	GIH	JPT	LWK	MEX	MKK	YRI
1	607	728	654	740	737	438	850	747
2	40	35	43	48	34	30	77	46
3	6	9	1	4	2	4	16	3
4	1			2			4	1
5	1	2						2
6	1							
7								
8								
9								
10								
11	1							
12								
13				1				
Total	657	774	698	795	773	472	947	799
genes with ≥ 2 intervals	50	46	44	55	36	34	97	52
% genes with ≥ 2 intervals	7.61	5.94	6.30	6.92	4.66	7.20	10.24	6.51

Table 10. HapMap Phase 3 independent eQTLs (intervals) at the 0.01 permutation threshold.

To address the extent to which gene activity is controlled by common regulatory sequences across populations, I explored sharing of independent eQTLs (intervals). This was done for all regulatory intervals detected in each population and comparison was not restricted to intervals detected for a given gene (the latter analysis was ongoing at the time of writing). At the 0.01 permutation threshold and from a non-redundant

union 3,288 independent intervals, 2,281 (70%) were found in a single population, 404 (12%) were shared in exactly two populations, 201 (6%), 145 (4%), 84 (3%), 65 (2%), 52 (2%) and 56 (2%) were shared in exactly three, four, five, six, seven and all eight populations respectively. Taken together, roughly 31% of intervals were found in at least two populations (Table 11). The high proportion of intervals detected in only one population suggests that even for the same cell type, genes are regulated to some extent by different regulatory elements across populations. Conversely, sharing of intervals implies sharing of regulatory elements. Relative sharing in ≥ five populations increased with higher significance stringency. The lower degree of sharing at the 0.01 permutation threshold may arise as a consequence of winner's curse (Goring, Terwilliger et al. 2001; Lohmueller, Pearce et al. 2003; Ioannidis 2008) which states that the effect sizes discovered when applying specific statistical significance thresholds are inflated compared to true effect size. Consequently the discovery sample usually achieves higher significance than replication samples. In this analysis the degree of sharing across populations may be underestimated if a gene with a significant association in one population barely fails significance correction in a second population. Sharing is likely to be further underestimated due to the fact that eQTL detection is affected by allele frequency differences across populations. Therefore a regulatory element may be active in multiple populations, but detected via an eQTL only in a fraction of these groups.

	0.01 permutation threshold	
No. populations	No. intervals	%
1	2,281	69.37
2	404	12.29
3	201	6.11
4	145	4.41
5	84	2.55
6	65	1.98
7	52	1.58
8	56	1.70
Total intervals	3,288	
>2 populations	1,007	30.63

Table 11. Sharing of intervals for HapMap Phase 3 cis significant genes.

4.5 eQTL-eQTL INTERACTION IN CIS

As outlined above, the genetic architecture of cis regulatory landscapes is complex with multiple regulatory intervals controlling gene expression. To dissect cis regulatory architecture further, I explored the degree to which interactions between variants in cis affect expression levels. DNA sequences containing enhancer elements for example are known to loop over great distances (> 1 Mb) and make physical contact with regulatory elements close to the TSS, in an interaction that affects initiation and rate of transcription (Figure 22 b). To detect such interactions, I applied a similar strategy to that used in Chapter 3 to test for interactions between regulatory and protein-coding variants (also see section 2.6.7). This analysis was carried out for the CEU and YRI populations.

SNPs with a nominal (uncorrected) p-value < 0.001 from the SRC association test were mapped in recombination hotspot intervals and the most significant SNP per interval was retained. SNP pairs with a D' > 0.5 across intervals were excluded from the analysis. Filtering for permutation significance was not performed to include variants that do not necessarily have large marginal effects on the phenotype, but whose impact

on gene expression may be revealed through an interaction. I carried out ANOVA to test the independent effects of each SNP as well as the SNP x SNP interaction term on gene expression in cis. Assuming a uniform distribution of nominal p-values, at the 0.01 nominal p-value threshold approximately 47 and 79 significant associations are expected by chance for the interaction term in the CEU and the YRI populations respectively. I detected 87 and 131 associations corresponding to an estimated FDR of 54% and 60% respectively (Table 12). At the stricter 0.001 nominal p-value threshold, approximately 5 and 8 significant associations are expected by chance for the interaction term in the CEU and the YRI populations respectively. I detected ten and 22 corresponding to an estimated FDR of 47% and 36% respectively (Table 12). Although this is not a very strong signal, given the strict filtering and relatively low power of this analysis, an enrichment of significant interaction terms is observed.

		0.01 nominal p-value threshold			0.001 nominal p-value threshold		
	Total tests	Expected	Observed	FDR	Expected	Observed	FDR
CEU	4,692	46.92	87	0.54	4.69	10	0.47
YRI	7,866	78.66	131	0.60	7.87	22	0.36

Table 12. Expected and observed significant interaction terms for CEU and YRI.

To explore this signal further, I conducted a single permutation of expression levels relative to genotypes. The p-value distributions of observed and permuted interaction terms were compared and an abundance of low p-values was found in the observed data for both populations (Figure 23). This suggests that gene expression in cis is sculpted to a certain extent by interacting regulatory elements.

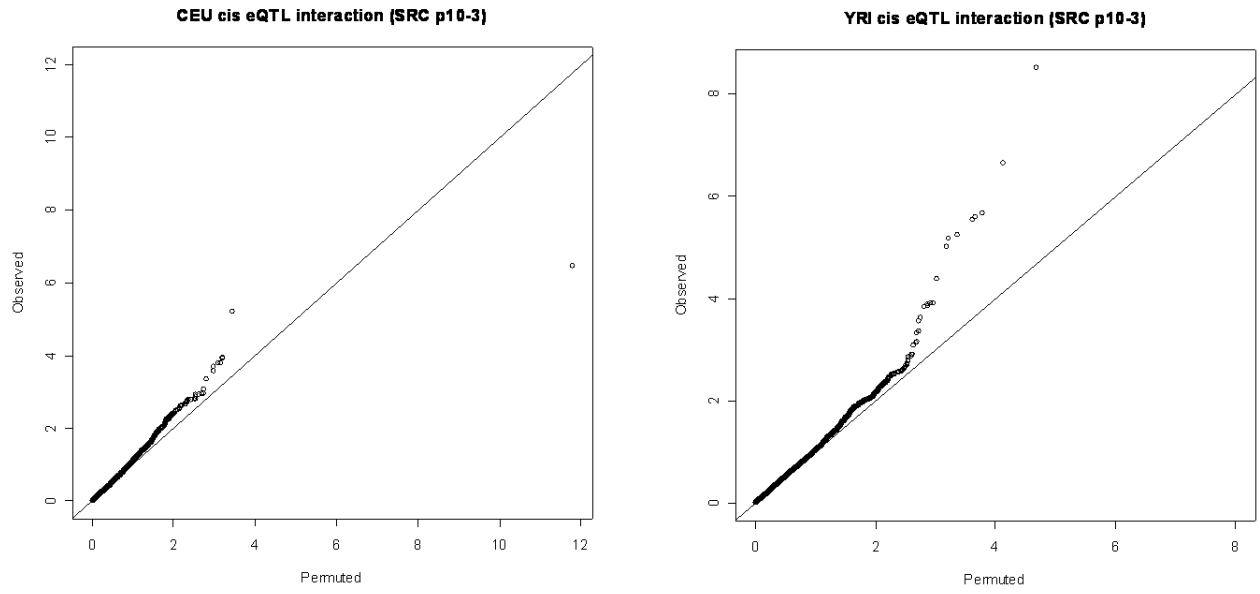

Figure 23. QQ plots of observed vs. permuted cis interaction p-values for the CEU and YRI HapMap Phase 3 populations. The signal at the tail of the observed distributions suggests that interactions between variants in cis influence expression levels of genes.

4.6 CONCLUSIONS

The signals detected in GWAS stem from markers that are not likely to be the causal variants. Furthermore, these markers typically delineate large genomic spaces that harbour causal variants. Replication of signals in independent studies provides corroborating evidence of causality, but the problem of delimiting the space carrying the functional variants remains. In this chapter I have presented a strategy that makes use of the properties of the genome and can be employed to restrict the space likely to contain regulatory elements controlling gene expression in cis. Using eight of the HapMap Phase 3 populations I demonstrated that seven percent of genes (0.01 permutation threshold) across all populations possess multiple independent regulatory intervals. The strategy applied involved strict filtering to remove highly correlated markers likely to tag the same regulatory element. As outlined in section 1.4.1, regulatory element length ranges from a few to a few hundred bp. Recombination

hotspot intervals on the other hand have a median length of 9,000 bp, ranging from a minimum of 998 bp to a maximum 31,495,264 bp. As a result, a single interval may contain multiple regulatory elements. The strategy employed in this study involved selection of the most significant eQTL per interval. Therefore, the number of truly independent eQTLs acting on genes in cis is likely to be higher and the method employed is most probably conservative.

The complexity of the regulatory landscape is further demonstrated through evidence of interactions between genetic variants with a small marginal impact on gene expression in cis. Using the CEU and YRI populations I explored the extent to which SNPs mapping in different intervals jointly affect cis expression levels. Although relatively underpowered, also because the ability to detect an interaction decays substantially when proxies of the functional variants are used, this study presents evidence for a cis interaction between regulatory variants. This approach does not test markers without marginal effects and cannot reveal variants that manifest themselves only in the context of an interaction. Consequently, the extent to which expression is influenced by interactions between variants in cis is likely to be an underestimate. A potentially more informative approach is to test all SNP pairs in the vicinity of a gene. This is currently being explored in collaboration with Doug Speed and Simon Tavaré at the CRI.

This study has highlighted the complex architecture of the cis regulatory landscape. GWAS of phenotypes in which expression levels are likely to play a crucial role should take this observation into consideration. Furthermore, integrating this information with studies on trans gene regulation will help piece together a more complete picture of gene expression control.