3 RTC – empirical method for integrating regulatory variants with complex trait associations

The biological interpretation of the plenitude of GWAS signals (WTCCC 2007; Eeles, Kote-Jarai et al. 2008; Zeggini, Scott et al. 2008) is very challenging since most candidate loci fall either in gene deserts or in regions with many equally plausible causative genes. Following the concurrent progress in understanding the genetic basis of regulatory variation (Cheung, Spielman et al. 2005; Dixon, Liang et al. 2007; Goring, Curran et al. 2007; Stranger, Forrest et al. 2007), differential gene expression has been proposed as a promising intermediate layer of information to aid this interpretation (Emilsson, Thorleifsson et al. 2008). Most commonly, interrogating the GWAS SNPs themselves for significant associations with gene expression has been employed to explain some of the GWAS results (Moffatt, Kabesch et al. 2007; Barrett, Hansoul et al. 2008). However, the ubiquity of regulatory variation throughout the human genome (Dixon, Liang et al. 2007; Stranger, Nica et al. 2007) makes coincidental overlaps of eQTLs and complex trait loci very likely. This likelihood is a direct consequence of the correlation structure in the genome (linkage disequilibrium - LD), which makes functionally unrelated variants statistically correlated.

As sample sizes increase, allowing the discovery of larger numbers of eQTLs of smaller effect size and as the expression experiments will be performed in a larger variety of tissues, we can envisage that almost every gene will have an associated eQTL under a certain condition. Consequently, the probability that any of these will map to a genomic region where a GWAS SNP also resides is very high. Therefore, it is important to emphasize that while it is very tempting to infer potential causal mechanisms based on such overlaps, this would be a naïve inference in the absence of additional supporting evidence for causality. In the long run, this will not only be an issue for gene expression, but also for any other cellular phenotype. Association studies for intermediate phenotypes with possible relevance to complex traits are underway and their results will overlap some of the GWAS signals. The biological meaning of these overlaps will again need to be evaluated in the context of the genome's correlation structure.

It is not evident though, how to model each genomic region with overlapping association signals in the absence of information about the history of the region. Accounting for the historical parameters of a region under the coalescent, while desirable, is computationally and practically not feasible since the human population history is too complex to properly model and small deviations or slightly incorrect assumptions could create false signals or reduce power.

In order to distinguish such accidental co-localizations (Chen, Zhu et al. 2008; Plagnol, Smyth et al. 2009) from true sharing of causal variants, I propose here an empirical methodology instead. This directly combines eQTL and GWAS data while accounting for the LD of the region harbouring the GWAS SNP. In this chapter, I demonstrate the value of the approach by predicting the regulatory impact of several GWAS variants in *cis* and *trans* and I also show that the correlation strength (r^2, D') between the GWAS SNP and the eQTL is not a sufficient predictor of regulatory mediated disease effects. This work has been described in (Nica, Montgomery et al. 2010).

3.1 Current GWAS signals are enriched for regulatory variants

To identify likely causal effects (not variants since full sequencing data is not available at this point), I took advantage of published association data catalogued in the NHGRI database (Hindorff, Sethupathy et al. 2009) and gene expression data generated in LCLs derived from HapMap 3 individuals (see Methods). In this study, I limited the expression analysis to the 109 CEU individuals (European origin), as they are the closest in ancestry to the majority of individuals in published GWAS studies. I used the NHGRI database (accessed 02.03.09) to extract 976 GWAS SNPs with minor allele frequency (MAF) > 5% that were also genotyped in the HapMap 3 CEU, thus allowing to test the exact GWAS SNPs for associations with differential gene expression in LCLs. In total 17673 genes were examined. To discover eQTLs, I used Spearman Rank Correlation (SRC). This method captures the vast majority of associations discovered with standard linear regression (LR) models, with the additional advantage that it's not affected by outliers and hence has more power and allows direct comparison of nominal P-values (Stranger, Nica et al. 2007). I looked for both proximal (*cis*) and distal (*trans*) effects as follows: variants within 1Mb on either side of the transcription start site (TSS) of a gene are considered to be acting in *cis,* while those at least 5 Mb downstream or upstream of the TSS or on a different chromosome are considered to be acting in *trans*.

In order to assess the overall impact of the currently known GWAS SNPs on expression, I contrasted their *cis* and *trans* effects to those of a random set of SNPs, representing the null. In a QQ plot (Figure 3.1), I compared the distributions of the best *cis* and *trans* association p-values per SNP for the 976 GWAS SNPs (observed) to 1000 sets of most significant p-values of 976 random SNPs each (expected). The 1000 random sets of 976 SNPs were sampled to have identical MAF distribution to the GWAS SNPs.

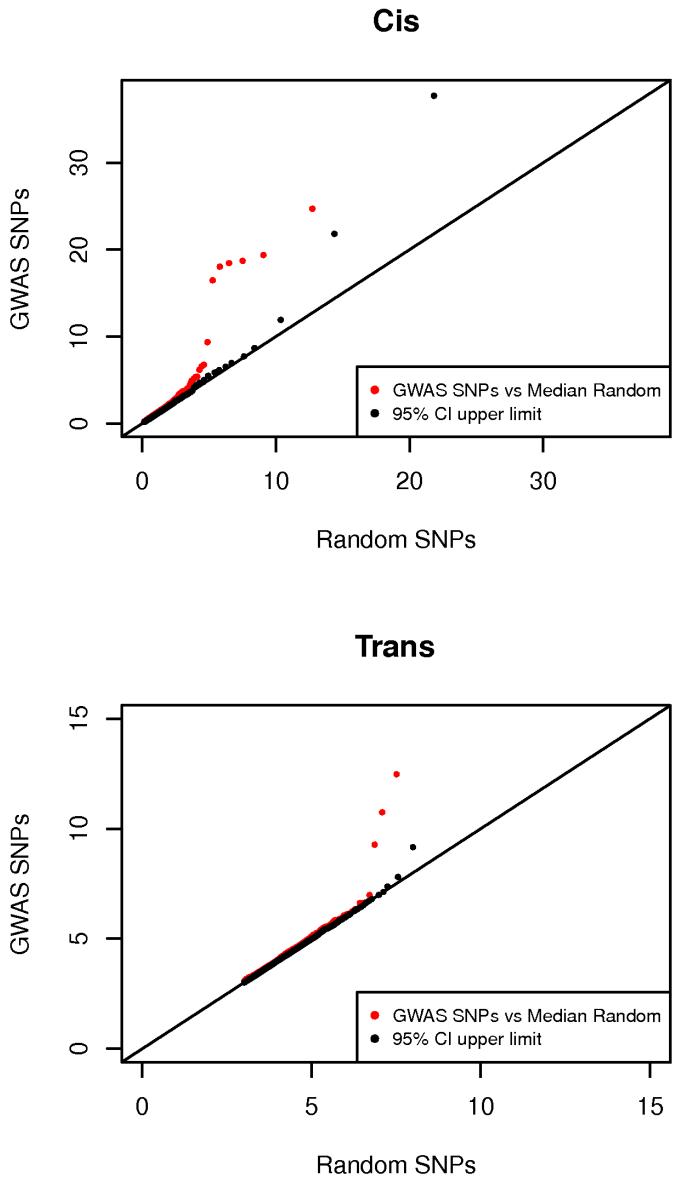

Figure 3.1. Excess of regulatory variants among GWAS signals. QQ plot depicting the excess of significant regulatory signal in GWAS data (976 NHGRI SNPs). For both the *cis* and *trans* analyses, the $-\log_{10}$(P-value) of the best associations per SNP are plotted. In red, the distribution of these values for GWAS SNPs is compared to that of the median of 1,000 sets of 976 random SNPs with same MAF distribution. In black, the estimated upper limit of the 95% confidence interval is plotted.

In *cis*, I observe a much stronger regulatory signal in the GWAS data compared to random (Figure 3.1). The significant difference between the two becomes apparent above a $-\log_{10}$(P-value) = 4. In *trans,* I also detect a more significant regulatory signal for GWAS SNPs compared to random, however not as strong as in *cis*. This is to be expected given that the much greater statistical space explored in *trans* limits the power to detect such effects.

Nevertheless, despite their confinement to one tissue type - LCLs, these comparisons support the overall explanatory potential of regulatory variation for the biological effects of GWAS variants. As expected given the nature of the tissue, the phenotypes responsible for this enrichment are immunity related (Figure 3.2).

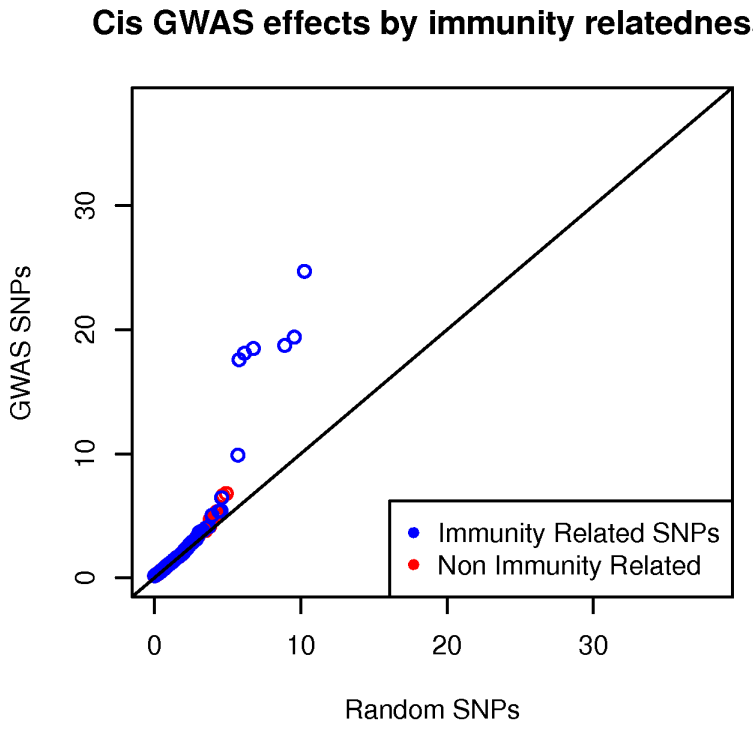

Figure 3.2. *Cis* regulatory enrichment stratified by immunity relatedness. The $-\log_{10}$(P-value) of the best associations per GWAS SNPs and a set of random SNPs are plotted. As expected given the tissue (LCLs), immunity related phenotypes are mainly responsible for the enrichment.

3.2 RTC score to distinguish between causal effects and coincidental overlaps

To identify the subset of causal effects from the regulatory enrichment observed, I focused only on the genomic regions harbouring either *cis* or *trans* eQTLs. I split the genome into recombination hotspot intervals based on genome-wide estimates of hotspot coordinates from McVean et al. (McVean, Myers et al. 2004). Limiting the search space for causal effects to these intervals is a reasonable conventional approach, as the lack of recombination events between the reported associated SNPs and the functional variants they are tagging enabled the discoveries through GWAS in the first place.

Given the abundance of *cis* eQTLs in the human genome, mere interval overlap is not sufficient to claim that a co-localized *cis* eQTL and a GWAS SNP are tagging the same functional variant. However, if the GWAS SNP and the eQTL do tag the same causal SNP, it is expected that removing the genetic effect of the GWAS SNP will have a marked consequence on the eQTL association. Starting from this hypothesis, I developed an empirical method to uncover regulatory mediated associations with complex traits. For all genes with a significant *cis* eQTL (0.05 permutation threshold as defined in Methods) (Stranger, Nica et al. 2007) in a given interval, I created corrected phenotypes from the residuals of the standard LR of the GWAS SNP against normalized expression values of the gene for which an eQTL exists. The residuals capture the remaining unexplained expression variance after the removal of the GWAS SNP effect. The SRC analysis was redone, this time with the pseudo phenotype, and the adjusted association P-value retained. Depending on the internal LD structure of the hotspot interval, the correlation between the GWAS SNP and the eQTL will vary, hence so will the P-values after and before correction. One way to assess the relevance of the GWAS SNP to the eQTL is to compare its correction impact to that of all other SNPs in the interval. For this purpose, I defined a Regulatory Trait Concordance (RTC) Score for each gene-GWAS SNP combination as a ratio taking into account the ranking of the correction with respect to all SNPs in the interval ($Rank_{GWAS\ SNP}$) and the total number of tested SNPs (N_{SNPs}) (see Methods).

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

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. As such, the RTC score will always be in the range (0,1], with values close to 1 indicating that the GWAS effect is the same as the eQTL effect.

The RTC score captures the LD structure of each tested region by taking into account the correction at all SNPs for every recombination hotspot interval. In addition, this ensures that RTC estimates are not up weighted in intervals with low number of SNPs (e.g. an extreme hypothetical case would be an interval with two SNPs only, the eQTL and the GWAS SNP; in this case the ranked correction at the eQTL would be high - Rank $_{\text{GWAS SNP}}$ = 1, as there is no other SNP in the interval to test; nevertheless, given just the 2 SNPs in the interval, the RTC score would only be 0.5 = (2 – 1) / 2). While this is not a problem for overestimating confident RTC scores, a caveat of the method is that intermediate values are equally discarded when in fact estimations derived from intervals with more SNP information should be up scaled (i.e. an RTC = 0.7 in an interval with 150 SNPs is more considerable than an RTC = 0.7 in an interval with 10 SNPs). Adjusting the value of the RTC score based on the SNP content of each region is a pending further development of the method. Meanwhile, one way to maximize the information content in each interval would be to include imputed SNP data. Given that the p-value associations prior to and after GWAS SNP correction are calculated with a non-parametric ranked test (SRC), it would be possible to use the estimates of allele dosage instead of the direct genotypes. This strategy has been shown to have comparable results to methods that take genotype uncertainty into account (Guan and Stephens 2008) and along with the SRC test as well as the permutations-based eQTL assignment, it should not be sensitive to outliers. A thorough evaluation of the use of imputed data to estimate RTC scores remains to be performed as a further improvement of the test.

3.3 RTC properties

The properties and robustness of the RTC score were investigated under the null hypothesis (H_0: eQTL and GWAS are tagging two different causal SNPs) and the alternative hypothesis (H_1: same causal SNP). For this purpose, I have simulated causal SNPs (cSNP), eQTLs and dSNPs (see Methods) varying the LD levels between them as well as the LD pattern of the hotspot interval where they reside. The cSNPs were then

masked and subsequently, the RTC score was calculated under these different LD scenarios for both hypotheses.

The RTC score is uniformly distributed under the null, when the simulated causal eQTL SNP (c-eQTL) and the causal disease SNP (c-dSNP) are different (Figure 3.3, left panel). Under the H_1 on the other hand, the RTC score is right skewed, with a clear enrichment for values close to 1 recovering the single causal SNP effect (Figure 3.3, middle panel).

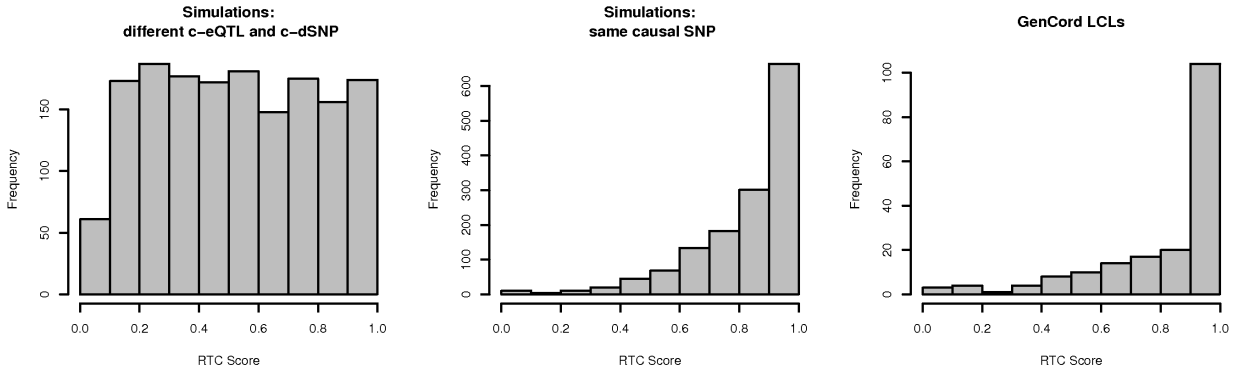

Figure 3.3. RTC score distribution following simulations. The RTC score is uniformly distributed for simulated eQTLs and dSNPs tagging two different causal variants in the same interval (left panel). The RTC Score is right-skewed for simulated eQTLs and dSNPs tagging the same functional variant (middle panel). The RTC score is sensitive to associations tagging a common functional variant in non-simulated data, when the GWAS trait is gene expression (GenCord LCL samples – right panel).

The simulations show that the complexity and variability of the LD structure in the genome impede the simple use of correlation metrics to infer shared causal effects. The statistical correlation (r^2) between the eQTL and the dSNP is not on its own sufficient to predict whether they tag the same cSNP. The RTC outperforms r^2 since it is able to recover causal effects even for low correlated pairs (Figure 3.4).

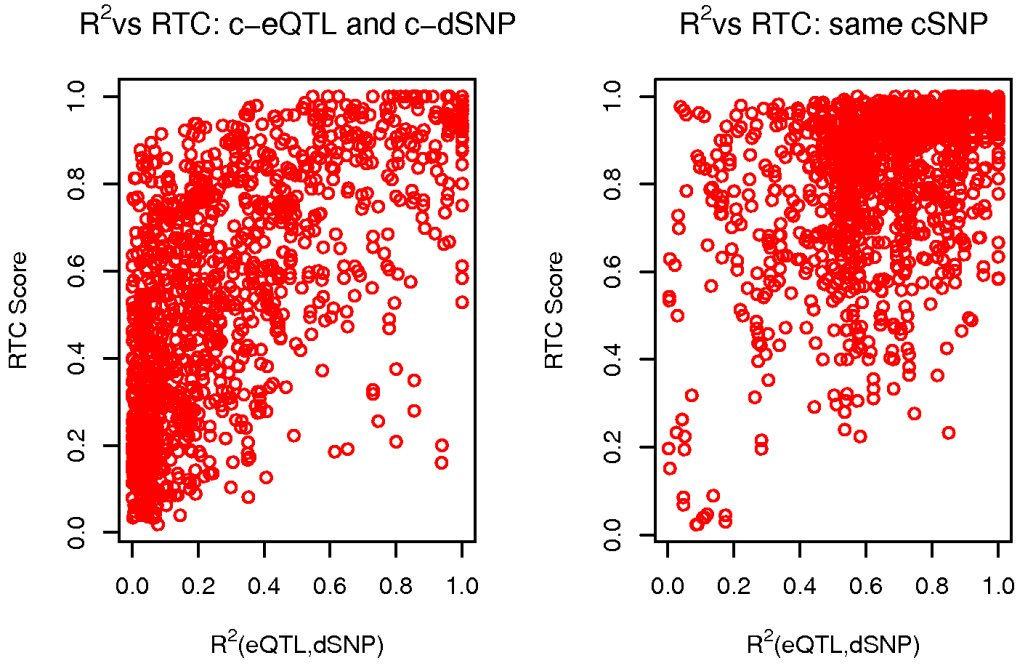

Figure 3.4. Properties of the RTC score when varying r^2. Simulation results depicting the relationship between the RTC score and the r^2 (eQTL, dSNP) when they tag different causal SNPs (H_0: left panel) versus one causal SNP (H_1: right panel). The RTC increases as expected with increased r^2 between the eQTL and the dSNP, but when tagging the same functional variant, various lower pairwise r^2 combinations can determine a high RTC. This makes r^2 on its own insufficient to detect shared causal effects.

The historical correlation metric between eQTLs and dSNPs (D') is also not fully predictive of high RTC scores (Figure 3.5). It can be observed from the H_0 simulation results that D' is not correlated with RTC, meaning that when the eQTL and dSNP tag different functional variants, the RTC score is not high just because D' is high. In addition, while high RTC scoring cases cluster much tighter around high D' values under the H_1 compared to r^2 previously, a high D' is not sufficient to predict causal effects. That is because it would be impossible to distinguish causal from coincidental effects given a perfect historical correlation scenario.

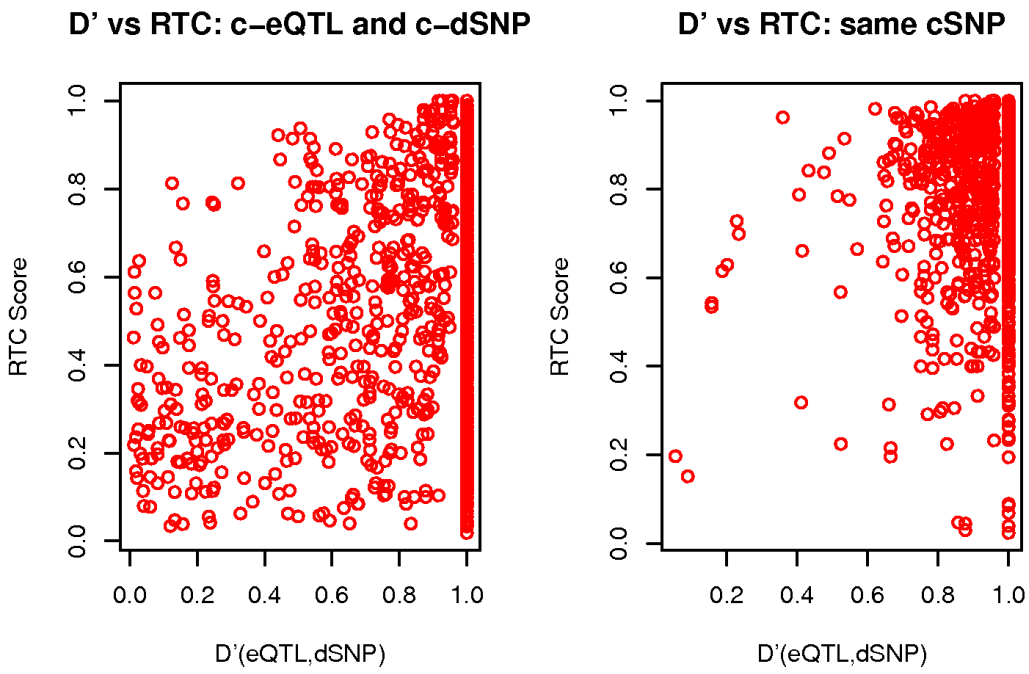

Figure 3.5. Properties of the RTC score when varying D'. Simulation results depicting the relationship between the RTC score and the D' (eQTL, dSNP) when they tag different causal SNPs (H_0: left panel) versus one causal SNP (H_1: right panel). D' is not correlated with RTC, therefore it will not determine high scores on its own in the absence of a common functional variant. Under the H_1, the majority of high RTC scoring pairs have high D', but in the case of a perfect historical correlation scenario, it's impossible to distinguish causal from coincidental effects with D' only.

Finally, the effect of the overall LD pattern in a region of interest on the estimation of the RTC score was investigated. For this purpose, I calculated the median r^2 of each hotspot interval (for all pairwise SNP combinations available per interval) and checked its relationship to the RTC score under the null and alternative hypothesis. It is expected that RTC will perform better in intervals with overall low LD, where the correlation between the eQTL and other non-disease SNPs will decay much faster, making the correction for the dSNP stand out. However, I confirm that the LD of the region does not determine high scores by itself. Intervals of low LD where different c-eQTLs and c-dSNPs reside have a uniform distribution of RTC scores (Figure 3.6, left panel). As expected, the H_1 simulations show that the RTC is most powerful in intervals with low median r^2 (Figure 3.6, right panel).

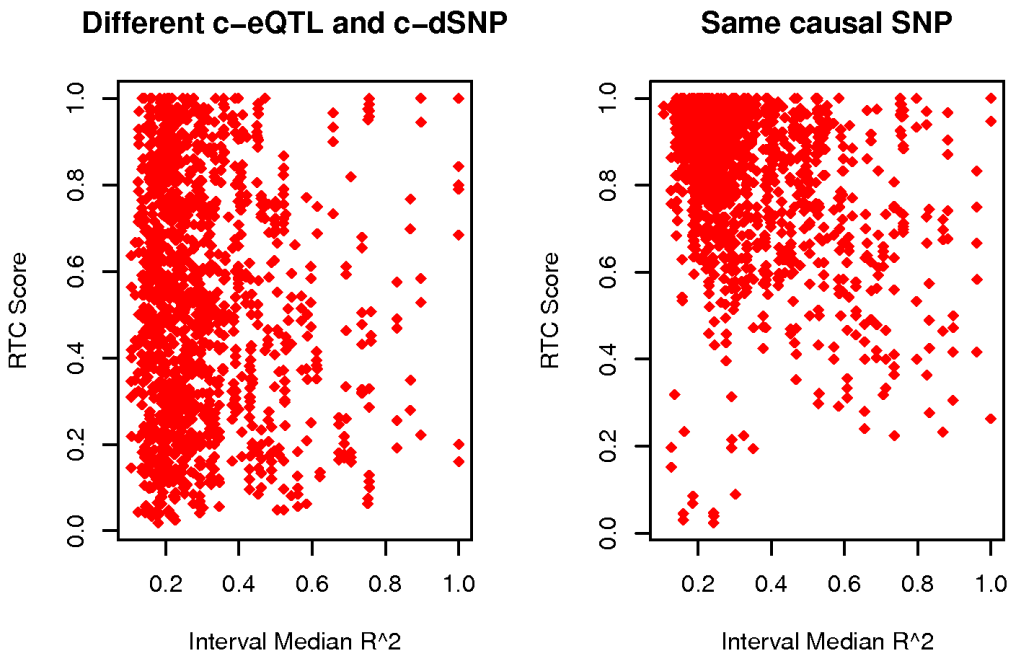

Figure 3.6. Properties of the RTC score when varying the median r^2 of the hotspot interval. Simulation results depicting the relationship between the RTC score and the local LD structure (median r^2) under the null (different causal SNPs - left panel) and alternative hypothesis (same causal SNP - right panel). Under H_0, the RTC score is evenly distributed, therefore intervals with overall low LD will not determine high RTC scores. Under H_1, the RTC performs best in intervals with overall low LD, where the correlation between the eQTL and other non-disease SNPs decays much faster, making the dSNP correction stand out.

3.4 RTC score when both traits are gene expression

In the first instance I tested the RTC method in a positive control experiment where intervals harbouring already identified regulatory associations were analyzed. I used published *cis* eQTLs (10^{-3} permutation threshold) discovered in the same tissue as the HapMap 3 CEU eQTLs (LCLs) but derived from an independent set of samples: 75 individuals of Western European origin from the GenCord resource (Dimas, Deutsch et al. 2009). In this experiment, I considered the GenCord eQTLs as the equivalent of GWAS SNPs and I limited the analysis to intervals with *cis* eQTLs in both datasets. Furthermore, I conditioned the associated genes for the same interval to be identical in the two expression datasets, expecting thus a common functional variant. As a result of this filtering, SNPs in 157 hotspot intervals were tested, associated with differential expression levels of 154 genes. As expected from the H_1 simulations, the RTC score distribution after correcting for the GenCord eQTLs is right-skewed (Figure 3.3, right

panel), suggesting that the scoring method is sensitive to associations tagging the same functional variant. I detect 33 SNP-probe pairs with an RTC score of 1 out of the total 185 tested pairs. Given the marked difference in genotyping density between HapMap and GenCord (~1.2 million SNPs versus ~400,000 SNPs respectively) and the hypothesis that the 157 overlapping intervals share the same functional variant, approximately 3 times more perfect scoring cases (99 pairs with RTC score = 1) are expected than what we observe, had individuals from both datasets been equally densely genotyped. I use the degree of sharing between the eQTLs in the two datasets to derive a reasonable, yet conservative threshold: currently, 105 SNP-probe pairs pass the 0.9 RTC threshold, making it thus a suitable stringent cut-off for calling significant discoveries.

3.5 *Cis* results

Following the positive control analysis, I applied the scoring method in a disease GWAS setting using the NHGRI SNPs described in Section 3.1. The respective 976 common GWAS SNPs map to 784 hotspot intervals. Of these, I focused the *cis* analysis on GWAS intervals (N=130) where at least one significant *cis* eQTL at a 0.05 permutation P-value threshold also resides. For the *trans* analysis, I ordered all 784 GWAS intervals by their most significant *trans* eQTL and kept the topmost 50 intervals for further examination. Table 3.1 summarizes the most confident *cis* results ordered by RTC score. I detect SNP-gene combinations passing the 0.9 threshold for 28 intervals out of the 130, twice as many than expected by chance (13 expected top 10% scoring intervals under the uniform distribution). The RTC method confirms prior results in the literature suggestive of disease effects mediated through expression (*ORMDL3* for asthma risk (Moffatt, Kabesch et al. 2007), *C8orf13* locus for systemic lupus erythematosus risk (Hom, Graham et al. 2008), *SLC22A5* for Crohn's disease (Peltekova, Wintle et al. 2004; Barrett, Hansoul et al. 2008). In addition, other yet unknown candidate genes for a variety of conditions are identified.

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs2064689	Crohn's disease	WDR78	1	1
rs3129934	Multiple sclerosis	HLA-DRB1	1	6
rs2188962	Crohn's disease	SLC22A5	1	5
rs1015362	Burning and freckling	TRPC4AP	1	20
rs2735839	Prostate cancer	C19orf48	1	19
rs6830062	Height	LCORL	1	4
rs2242330	Parkinsons disease	TMPRSS11A	1	4
rs7498665	Body mass index,Weight	EIF3CL	1	16
rs2872507	Crohn's disease	ZPBP2	0.99	17
rs255052	HDL cholesterol	AGRP	0.99	16
rs4549631	Height	TRMT11	0.98	6
rs9469220	Crohn's disease	ILMN_29412	0.98	6
rs11083846	Chronic lymphocytic leukemia	SLC8A2	0.98	19
rs13277113	Systemic lupus erythematosus	C8orf13	0.97	8
rs9272346	Type 1 diabetes	HLA-DRB1	0.96	6
rs12324805	Body mass index	STARD5	0.96	15
rs3764261	HDL cholesterol	MT1H	0.96	16
rs3135388	Multiple sclerosis	HLA-DRB5	0.96	6
rs3814219	Endothelial function traits	FAM26B	0.95	10
rs12708716	Type 1 diabetes	ILMN_32084	0.95	16
rs2269426	Plasma eosinophil count	HLA-DRB1	0.95	6
rs10769908	Body mass index	C11orf17	0.94	11
rs4130590	Bipolar disorder	ILMN_17339	0.94	9
rs7216389	Asthma	ORMDL3	0.94	17
rs3796619	Recombination rate (males)	CRIPAK	0.93	4
rs1748195	Triglycerides	DOCK7	0.93	1
rs2903692	Type 1 diabetes	ILMN_32084	0.93	16
rs3197999	Crohn's disease	SLC38A3	0.92	3
rs9858542	Crohn's disease	SLC38A3	0.92	3
rs6441961	Celiac disease	LIMD1	0.92	3
rs660895	Rheumatoid arthritis	PSMB9	0.91	6
rs9652490	Essential tremor	ILMN_111363	0.91	15
rs1397048	Hemostatic factors	OR8H2	0.91	11
rs3825932	Type 1 diabetes	CTSH	0.91	15
rs2395185	Ulcerative colitis	ILMN_29412	0.9	6

Table 3.1. Candidate *cis* results. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on 976 GWAS SNPs from NHGRI and HapMap 3 CEU expression data in LCLs. The higher the score, the more likely it is that the GWAS SNP and the eQTL for the gene shown are tagging the same functional variant.

An interesting example of a novel *cis* regulatory mediated effect is the one for Crohn's disease with gene *SLC38A3*, member 3 of the solute carrier family 38. Independent studies detected significant Crohn's associations of two SNPs in the same hotspot interval on chromosome 3: rs3197999 (Barrett, Hansoul et al. 2008), a non-synonymous SNP in gene *MST1* and rs9858542 (Parkes, Barrett et al. 2007; WTCCC 2007), a synonymous SNP in nearby gene *BSN*. Suggestive literature evidence supports the role of *MST1* in Crohn's pathogenesis: the protein encoded by *MST1* (macrophage-stimulating protein – MSP) and its receptor MST1R are reportedly involved in macrophage chemotaxis and activation (Leonard and Skeel 1976) and have a role also in regulating inflammatory responses following pro-inflammatory signals (Morrison, Wilson et al. 2004). These lines of evidence, in addition to the disease associated non-synonymous SNP made *MST1* the most attractive candidate gene out of the many present in that region (Goyette, Lefebvre et al. 2008). However, the data presented here supports an additional regulatory component underlying the susceptibility locus. For both GWAS SNPs, *SLC38A3* is the highest scoring candidate in the region (RTC score: 0.92). Interestingly, this is functionally similar to another Crohn's susceptibility gene *SLC22A5* confirmed with the RTC method (RTC score = 1) and also encoding a sodium dependent multi-pass membrane protein (solute carrier family protein). The observed direction of effect is the same for both genes (eQTLs associate with low expression levels) as in previous expression datasets (Barrett, Hansoul et al. 2008) and suggests a possible involvement of this gene family in the disease. This is in agreement with recent studies reporting that disease causative genes are functionally more closely related (Franke, van Bakel et al. 2006).

Overrepresentation of immunity-related results

The tissue under investigation is LCLs so it is expected that GWAS signals of immunity related traits (comprising here autoimmune disorders and diseases of the immune system e.g. AIDS progression) more likely show an overlap with eQTLs. In order to evaluate the relevance of the presented results, I analyzed the distributions of the best RTC scores per GWAS SNP stratified by the immunity relatedness of the complex trait they associate with (Figure 3.7).

I observe a significant overrepresentation of high-scoring genes (RTC ≥ 0.9) for immunity related traits compared to non-immunity related ones (Fisher's Exact Test, P-value = 0.0125) (Fraser and Xie 2009). This suggests that the scoring scheme predicts regulatory effects of the relevant phenotypes. In addition, we observed that for GWAS signals with RTC score ≥ 0.9, only 10% of the nearest gene to the GWAS SNP was also the eQTL gene. These however, correspond as expected to instances when the eQTL gene is also the nearest gene to the eQTL itself. If that is not the case, the inference of relevance of a gene simply based on its proximity to the GWAS SNP is not informative.

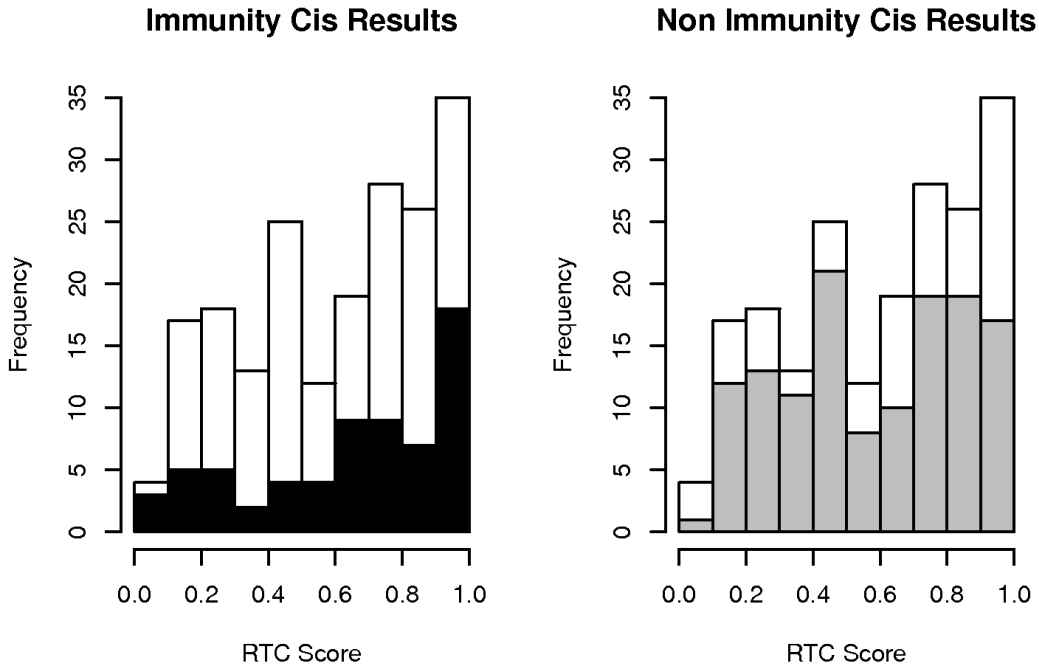

Figure 3.7. Overrepresentation of immunity-related high RTC scoring *cis* signals. Distribution of best RTC scores per GWAS SNP stratified by immunity relatedness. Histogram contains results from the analysis of 130 hotspot intervals with colocalizing disease SNPs and *cis* eQTLs. We observe a significant overrepresentation of high-scoring (RTC ≥ 0.9) candidate genes (black bars) for immunity related complex traits compared to non-immunity related ones (grey bars) (Fisher's Exact Test, P-value = 0.0125).

3.6 *Trans* results

Even if the causal SNP is not *cis*-regulatory, using gene expression to determine its downstream targets, coupled with information about the biological pathways these targets act in could help interpret the primary GWAS effect.

I investigate this hypothesis in the topmost 50 GWAS intervals ordered by their *trans* eQTL significance. For each interval, I apply the RTC scoring scheme on the subset of genes in the whole genome with a notable effect in *trans* (SRC nominal P-value < 10^{-5}). These signals amount to a total of 552 genes. I obtain SNP-gene combinations passing the 0.9 RTC score threshold for 24 of the 50 tested intervals (corresponding to a total of 85 genes). Six of these intervals contain GWAS SNPs associated with immunity related traits (Table 3.2).

While not statistically significant - unsurprisingly given that only a small subset of the total GWAS intervals is tested - these examples support the usefulness of the *trans* approach. As hypothesized, for the same complex trait associated SNP, several potential candidate genes in *trans* can be discovered throughout the genome. Some of these are biologically plausible results and merit further investigation. However, many *trans* candidates are hard to interpret at this stage given their incomplete annotation and further functional studies will need to be performed for validation.

Table 3.2. Candidate *trans* results. Candidate *trans* genes likely involved in the same biological pathways, relevant to the GWAS SNPs (GWAS SNP and the genes it affects in *trans* often reside on different chromosomes, as indicated in the SNP Chr and Genes Chr fields respectively). Signals related to the same hotspot interval separated by a horizontal line. Regulatory *trans* effects RTC applied in *trans* on 976 GWAS SNPs from NHGRI and HapMap 3 CEU expression data in LCLs. Table contains only the confident results (RTC Score ≥ 0.9) for the six immunity related intervals.

GWAS SNP	Complex Trait	Genes	RTC	SNP Chr	Genes Chr
rs2251746	Serum IgE levels	SLC25A18	0.99	1	22
rs983332	Response to TNF antagonists	RGS16, IGSF3	0.97	1	1
rs983332	Response to TNF antagonists	C17orf58	0.97	1	17
rs653178	Celiac disease	PAX8, DOK1	1	12	2
rs17696736	Type 1 diabetes	PAX8, DOK1	0.98	12	2
rs2542151	Crohn's,Type 1 diabetes	MMP12	1	18	11
rs2542151	Crohn's,Type 1 diabetes	SLC39A4, PSD3, AHNAK2, FAM108B1, CYP2S1, CLEC7A	0.97	18	8, 8, 14, 9, 19, 12
rs2542151	Crohn's,Type 1 diabetes	LENEP	0.91	18	1
rs3134792	Psoriasis	ADRA2C	1	6	4
rs3134792	Psoriasis	DPEP1, ARHGEF3	0.99	6	16, 3
rs1265181	Psoriasis	POU5F1P1	0.96	6	8
rs1265181	Psoriasis	DPEP1	0.95	6	16
rs1265181	Psoriasis	CYP4F8, ADRA2C	0.94	6	19, 4
rs1265181	Psoriasis	RGS9	0.92	6	17
rs2395185	Ulcerative colitis	B4GALT2, ASB5	0.97	6	1, 4
rs2395185	Ulcerative colitis	STK32A	0.94	6	5
rs2395185	Ulcerative colitis	OXT	0.93	6	20
rs2395185	Ulcerative colitis	CSRP3	0.92	6	11
rs2395185	Ulcerative colitis	LGALS4	0.91	6	19
rs3135388	Multiple sclerosis	LIMS1	0.95	6	2
rs477515	Inflammatory bowel disease	B4GALT2	1	6	1
rs477515	Inflammatory bowel disease	ASB5	0.99	6	4
rs477515	Inflammatory bowel disease	STK32A	0.95	6	5
rs477515	Inflammatory bowel disease	OXT	0.94	6	20
rs477515	Inflammatory bowel disease	CSRP3	0.93	6	11
rs477515	Inflammatory bowel disease	DCHS2	0.91	6	4
rs477515	Inflammatory bowel disease	LGALS4	0.9	6	19
rs615672	Rheumatoid arthritis	DCHS2	0.99	6	4
rs6457617	Rheumatoid arthritis	SMARCD3	0.95	6	7
rs6457620	Rheumatoid arthritis	SMARCD3	0.95	6	7
rs660895	Rheumatoid arthritis	RETSAT	0.99	6	2
rs660895	Rheumatoid arthritis	CALCR	0.98	6	7
rs9268877	Ulcerative colitis	LIMS1	0.97	6	2
rs9268877	Ulcerative colitis	B4GALT2	0.94	6	1
rs9268877	Ulcerative colitis	ASB5	0.91	6	4
rs9272346	Type 1 diabetes	LIMS1	0.97	6	2
rs9272346	Type 1 diabetes	WHDC1L1	0.94	6	15
rs9272346	Type 1 diabetes	ASB5	0.93	6	4
rs9272346	Type 1 diabetes	SEMA6D, OXT, B4GALT2	0.92	6	15, 20, 1

A subset (N=15) of the hotspot intervals containing GWAS SNPs and tested in this chapter harbour both *cis* and *trans* eQTLs (as defined in Methods). For two of the 15 intervals, I detect potential explanatory regulatory effects (genes with high RTC score) in both *cis* and *trans* (Table 3.3.). It is likely that changes in expression levels of all these genes are relevant to the single common GWAS signal. Interestingly for example, the *DOCK7* (dedicator of cytokinesis 7) locus has been implicated in coronary heart disease risk (Aulchenko, Ripatti et al. 2009) and SNP variants at the *SORCS2* (sortilin-related VPS10 domain containing receptor 2) locus have been associated with hemorrhagic stroke (Yoshida, Kato et al. 2010). Both genes score a high RTC with SNP rs1748195 associated with triglyceride levels, a quantitative trait highly relevant to heart disorders. Functional verification of similar gene connections might lead to the discovery of new disease-relevant pathways.

	GWAS SNP	Complex Trait	Gene	RTC	SNP Chr	Gene Chr	Interval
cis	rs1748195	Triglycerides	DOCK7	0.93	1	1	1:62673568-62974568
trans	rs1748195	Triglycerides	SORCS2	0.9	1	4	1:62673568-62974568
cis	rs1007738	Bone mineral density (hip)	ACP2	0.88	11	11	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	CAPN12	0.98	11	19	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	SYNGR3	0.87	11	16	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	TMEM149	0.83	11	19	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	PBXIP1	0.82	11	1	11:46234001-46861001

Table 3.3. Hotspot intervals with overlapping *cis* and *trans* effects as indicated by the high RTC score. Candidate regulatory effects explaining GWAS signals were detected for two of the 15 intervals tested for both *cis* and *trans* effects.

3.7 RTC outperforms alternative correlation metrics

The power to detect significant associations between genotyped SNP proxies and a phenotype depends on the correlation between those proxies and the functional variant (Pritchard and Przeworski 2001). Just like for the simulated data, I tested whether the correlation between a GWAS SNP and its co-localizing eQTL is sufficient for predicting a shared causal effect. For both the *cis* and the *trans* analysis, I observe that the r^2 between the eQTL and the disease SNP is not a direct predictor of the RTC score, and in

several cases I predict that even pairs with low r^2 are likely tagging the same functional effect (Figure 3.8, top panel).

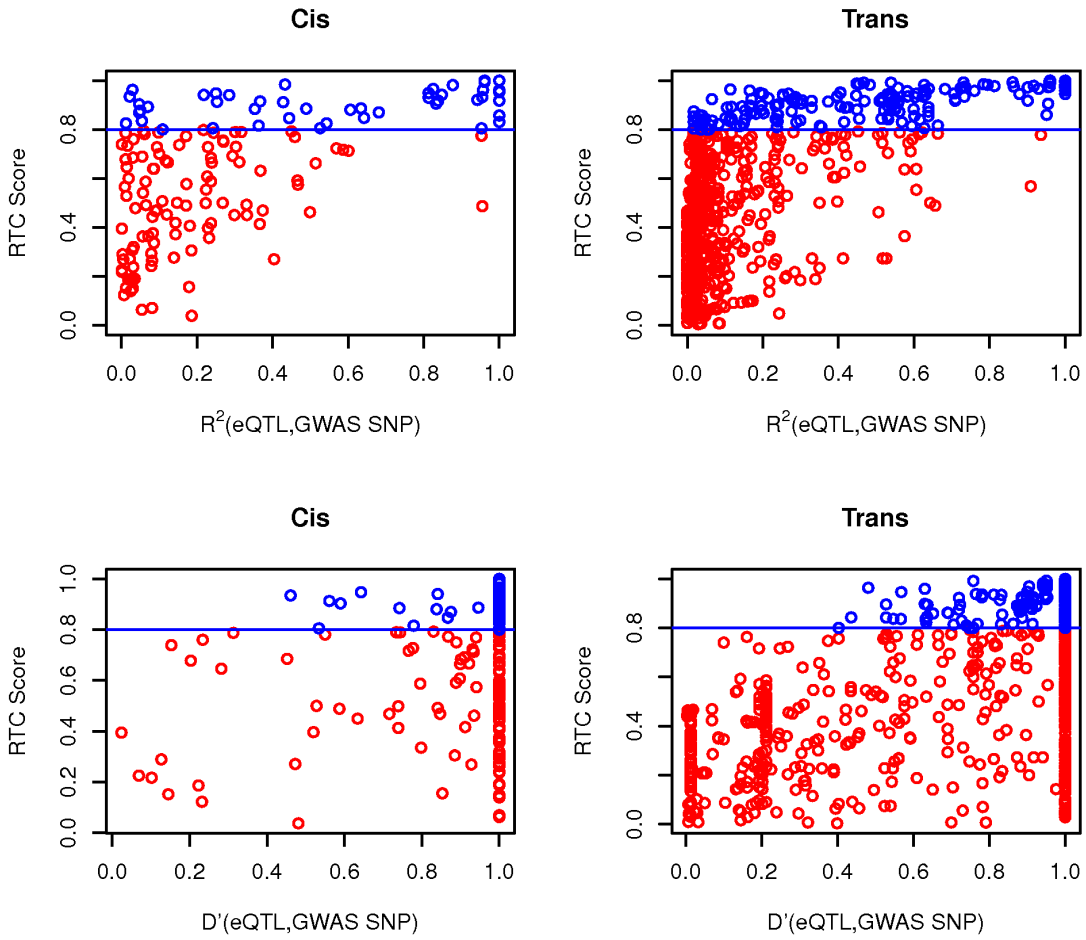

Figure 3.8. The RTC method compared to standard LD measurements in the observed data. Neither r^2 nor D' between the eQTL and the GWAS SNP are direct predictors of a high RTC score. Highlighted here are the results from the *cis* and *trans* analyses. I obtain high scoring results (RTC scores ≥ 0.8 in blue) for cases with a high correlation between the disease SNP and the eQTL as expected, but also for pairs with low statistical correlation (r^2 – top panel). As shown in the bottom panel, many of these high scoring pairs are historically correlated (D' = 1), but so are many more by chance. Additionally, high scoring pairs with low D' can be detected as well. Hence, no obvious combination of the two LD measures can predict a high RTC score.

The reason for this is that many of the high scoring pairs with poor statistical correlation (low r^2) are actually historically correlated (D'=1). Nevertheless, D' is not very informative either (Figure 3.8, bottom panel), the main problem here being that in regions with generally high D' among many SNPs, one cannot determine which of the pairs actually represents a common functional variant.

Another metric of potential predictive value is the fraction of eQTL variance explained by the dSNP. Figure 3.9 indicates the relationship between the RTC score and the fraction of explained variance at the eQTL left unexplained after the dSNP correction (ratio of linear regression adjusted R^2 after and before correction). As expected given the definition of the RTC, the highest density of good scoring results is registered for dSNPs that explain most of the eQTL variance. However, RTC outperforms the variance metric, scoring high even when less of the eQTL variance is explained by a dSNP. As such, setting a threshold on the explained variance would not be sufficiently informative either.

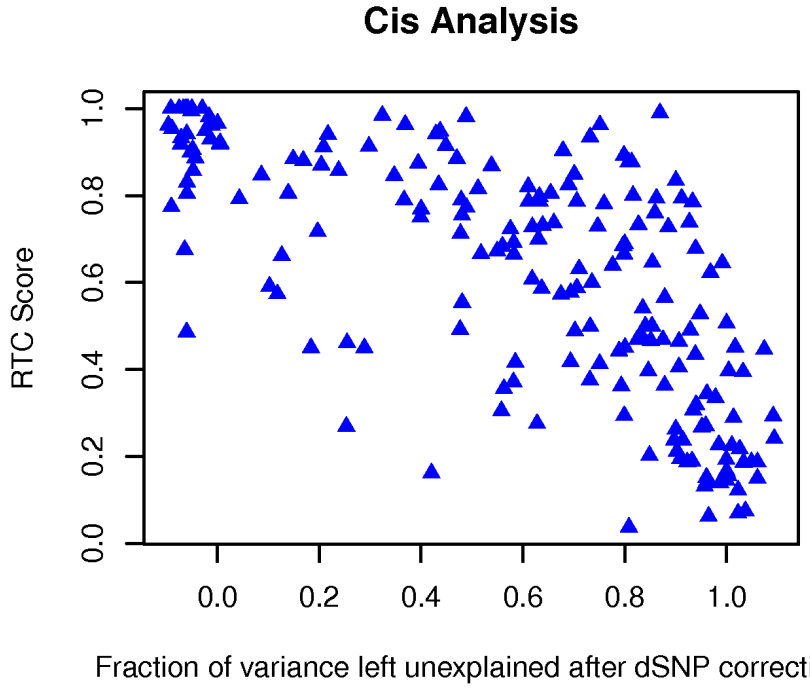

Figure 3.9. The fraction of eQTL variance explained away by the dSNP versus the RTC score. Contrasted are the LR adjusted R^2 at the eQTL after and before correction of the dSNP. It is observed that while most high scoring pairs correspond to cases of lowest variance left unexplained, solely using an arbitrary variance threshold would cause other interesting cases to be missed.

3.8 Conclusions

In this chapter, I described a newly developed empirical methodology, called Regulatory Trait Concordance (RTC). The purpose of this method is to account for local LD structure in the human genome and integrate eQTLs and GWAS results to reveal the subset of association signals that are due to *cis* eQTLs. This approach aims to help understand some of the biological mechanisms - should they be regulatory - behind the genetic associations with complex diseases. Candidate genes linked to the SNP variants

reported so far as implicated in disease susceptibility are often chosen solely based on genomic proximity criteria. The RTC enables therefore a more informed choice of candidate disease genes, based on evidence in favour of common functional regulatory effects.

Genomic regions of various LD patterns were first simulated to explore the properties of the RTC score. Simulated intervals for both cases when a single or two different causal variants exist were analyzed. Consequently, I showed that the proposed scoring scheme outperforms SNP correlation metrics, be they statistical (r^2) or historical (D'). Following the observation of a significant abundance of regulatory signals among currently published GWAS loci, I applied the method on expression data in blood-derived LCLs extracted from HapMap 3 individuals of European descent. Relevant genes under regulatory control were prioritized for each of the respective complex traits. As such, I detected several potential disease causing regulatory effects, with a strong enrichment for immunity-related conditions, consistent with the nature of the cell line tested (LCLs). Furthermore, I presented an extension of the method in *trans*, where interrogating the whole genome for downstream effects of the disease variant can be informative regarding its unknown primary biological effect.

Overall, the RTC method supports the integration of cellular phenotype associations with organismal complex traits as a way to biologically interpret the genetic determinants of these traits.