# 5 Tissue-dependent causal regulatory effects

Gene expression studies performed so far on multiple human tissues support an extensive level of eQTL tissue-specificity. Comparisons across dissimilar cell-types documented first the considerable tissue dependency of *cis* regulatory variation. As such, overlaying LCL and cortical tissue eQTLs resulted in barely any overlap (Myers, Gibbs et al. 2007), the comparison of adipose and blood expression patterns in two Icelandic cohorts reported that 50% of the detected *cis* eQTLs were shared (Emilsson, Thorleifsson et al. 2008) and a study overlapping eQTLs from autopsy-derived cortical tissue and peripheral blood mononucleated cells revealed less than 50% sharing (Heinzen, Ge et al. 2008). In the previous chapter (Chapter 4), I further explored the complexity of eQTL tissue-specificity in LCLs, skin and fat, estimating a relatively low proportion of shared eQTL effects (~30%). Moreover, eQTLs do not display significant tissue-specific properties only among cell-types with substantially different cellular functions. The study from our lab overlapping regulatory variants in transformed B-cells (LCLs), fibroblasts and primary T-cells provided evidence that even cell-types as closely related as B-cells and T-cells share only a minority of *cis* eQTLs (<15%) (Dimas, Deutsch et al. 2009).

A similarly high emphasis on tissue-dependency has been put in the context of other complex traits, including disease phenotypes. Different diseases manifest themselves in different organs and have different tissues as primary targets of pathology (connective tissue diseases, muscle diseases, etc.). However, the tissues where diseases are manifested are not necessarily informative of the cell-type where the causal mechanism leading to disease progression occurs. Perhaps one of the best illustrative examples in this sense is the one of MC4R (melanocortin-4 receptor), a gene in the vicinity of which several common variants associated with fat mass, weight and obesity risk were discovered (Loos, Lindgren et al. 2008). Rare functional mutations in MC4R, known to cause monogenic severe childhood-onset obesity (Vaisse, Clement et al. 1998; Yeo, Farooqi et al. 1998) and further functional evidence from murine models (Huszar, Lynch et al. 1997) indicated that MC4R is also responsible for common obesity. Most likely, the susceptibility loci act by disrupting the expression of MC4R. However, this hypothesis

remains hard to verify, given that MC4R is almost exclusively expressed in the brain and found at very low levels in most of the currently available expression datasets. The brain specific expression pattern of an obesity susceptibility gene is a clear example of the fact that predicting a tissue's relevance to disease is complicated by the discrepancy between the place of its manifestation and where the causal mechanism initiates.

In this chapter, I explore the role of tissue-dependency in predicting causal regulatory effects for GWAS loci. Specifically, I apply the RTC methodology described in Chapters 2 and 3 on the three cell-types (B cells, T cells and fibroblasts) available from the GenCord (because of time and availability constraints, this analysis was performed on the GenCord rather than the MuTHER resource, which is also of very high interest and constitutes a future project in the lab; see Methods). I show that finding regulatory variants and corresponding differentially expressed genes underlying complex disease associations is highly dependent on the nature of the cell-type tested. The results confirm previously suspected candidate loci and offer new functional insights into disease aetiology by revealing novel differentially regulated susceptibility genes.

## 5.1 RTC score distribution by tissue

To detect likely causal regulatory effects for GWAS signals across multiple tissues I used expression and genotypic data derived from the 75 GenCord European individuals. The ~400,000 genotyped SNPs were imputed first on HapMap 2 in order to increase power to detect associations with expression. After imputation and QC filtering (see Methods), 1,428,314 SNPs with MAF > 5% were available for analysis. Transcript level measurements in transformed B-cells, fibroblasts and primary T-cells for probes mapping uniquely to 15,596 Ensembl genes (Methods) were tested for association with SNP genotypes. I considered SNP–gene associations with a nominal SRC p-value < $10^{-4}$ as eQTLs. This nominal threshold corresponds roughly to a 0.05 permutation threshold (estimation from multiple eQTL analyses on different datasets run in our group). At this significance level, 1139 genes were detected to have at least one eQTL in B-cells, 1098 genes in T-cells and 1157 genes in fibroblasts. I overlapped these sets of eQTLs with GWAS results in each tissue separately. For this purpose, I revisited the NHGRI GWAS catalogue (Hindorff, Sethupathy et al. 2009) and downloaded the most recent list at that time of SNPs associated with complex traits (accessed 12.04.2010). 1750 SNPs were

retrieved, associated with 248 traits. I further mapped both the GenCord eQTLs and NHGRI GWAS SNPs to recombination hotspot intervals (McVean, Myers et al. 2004) as described in Chapter 2. The 1750 GWAS SNPs mapped to 1400 intervals, suggesting that some intervals harbour multiple susceptibility loci often associated with multiple traits, as in the case of the well-known 8q24 gene desert region, where variant associations with breast, prostate and colorectal cancer have been reported (Ghoussaini, Song et al. 2008; Al Olama, Kote-Jarai et al. 2009). The overlap of recombination hotspot intervals where at least one GWAS SNP and one eQTL co-localize resulted in 106, 111 and 105 intervals in B-cells, T-cells and fibroblasts respectively. A subset of these intervals contained more than one disease GWAS locus. Each interval was tested in *cis* under the RTC framework (see Methods) for every disease association reported by NHGRI (in total 149, 150 and 144 interval-disease combinations were tested in B-cells, T-cells and fibroblasts – Table 5.1).

| | #Intervals | #Interval-diseases |
|---|---|---|
| **B-cells** | 106 | 149 |
| **T-cells** | 111 | 150 |
| **Fibro** | 105 | 144 |
| **Shared** | 26 | 53 |

**Table 5.1. Number of nonredundant recombination hotspot intervals with co-localizing GenCord eQTLs and GWAS SNPs.** Unique count of intervals (#Intervals) and interval-disease combinations (#Intervals with multiple disease GWAS loci - #Interval-diseases) per tissue tested for causal regulatory effects with the RTC. A subset of intervals (Shared) harbour eQTLs detected in all three tissues.

In each of the three tissues, an overrepresentation of GWAS regulatory candidates with high RTC scores is observed (Figure 5.1). I detect SNP-gene associations passing the 0.9 RTC threshold for 41 interval-disease combinations of the total 149 tested in B-cells, 36 high scoring interval-disease combinations of 150 tested in T-cells and 29 out of 144 tested in fibroblasts. The overall distribution of scores differs across tissues, with a more noticeable similarity between B-cells and T-cells. This similarity between the two cell-types in contrast to the different pattern of RTC scores in fibroblasts is even more apparent when focusing on the shared subset of 53 interval-disease combinations tested in all three tissues (Figure 5.2).

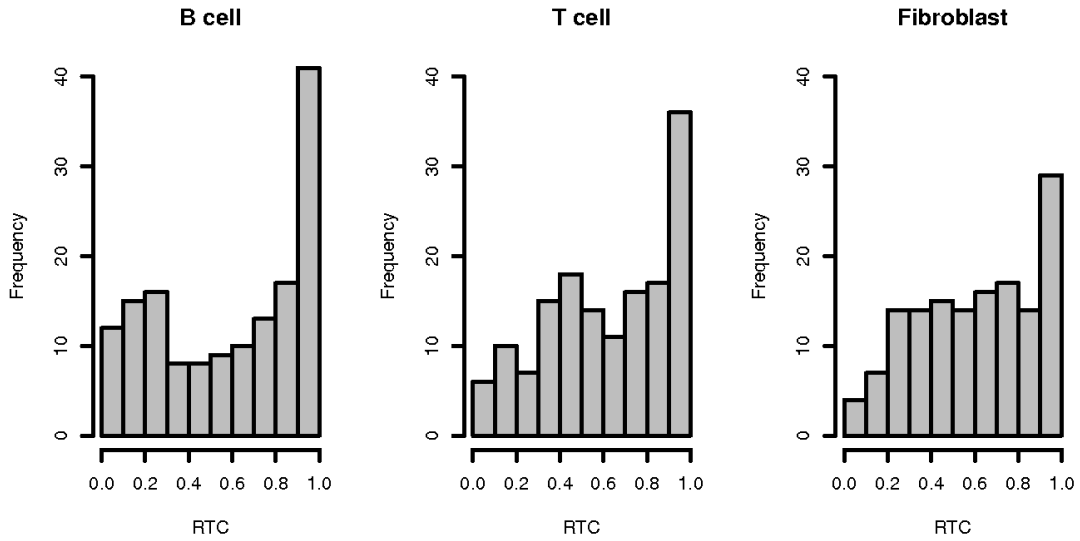


**Figure 5.1. Distribution of RTC scores across tissues.** The best score per GWAS SNP per interval-disease combination is plotted for all tested intervals. An enrichment of high RTC scoring candidates (≥ 0.9) is present in each tissue, corresponding to the subset of GWAS SNPs likely explained by regulatory effects.

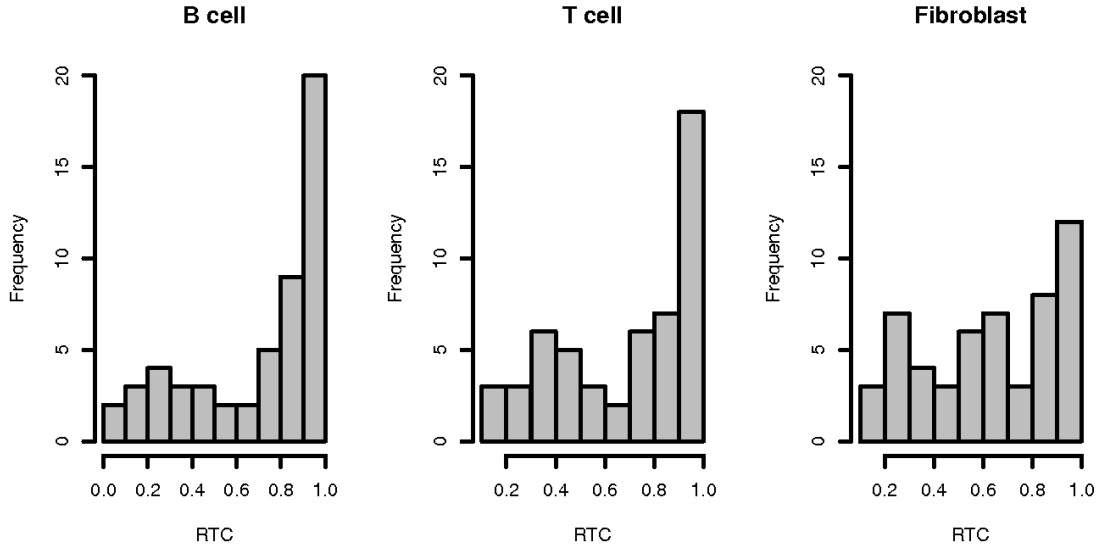


**Figure 5.2. Distribution of RTC scores across tissues for shared intervals-disease combinations.** The best score per GWAS SNP per interval-disease combination is plotted only for the 53 shared interval-disease combinations tested in all three tissues. A marked difference in distribution of scores can be observed, especially between fibroblasts and the more functionally similar B-cells and T-cells.

The differential RTC score distribution across the three tissues reflects the biological similarities of the tested cell-types. B-cells and T-cells are two classes of lymphocytes, the white blood cells involved in the body's adaptive immune response (Alberts 2002). Fibroblasts on the other hand, are a type of connective tissue cells secreting the

extracellular matrix and collagen and playing an essential role in wound healing. Nevertheless, despite being much more functionally related, B-cells and T-cells carry out two distinct types of immune responses whose disruption may thus have different phenotypic consequences. B-cells participate in the antibody immune response whereby following their activation by foreign antigens, they secrete antibodies. These can then circulate in the bloodstream, bind the antigens that stimulated their production and thus inhibit the detrimental action of viruses or microbial toxins on the cell. In most mammals, B-cells are produced in the bone marrow (Alley 1987). T-cells, the other major class of lymphocytes, are produced in the thymus and are responsible for cell-mediated immune responses (Spits 2002). After their activation, T-cells react directly against a foreign antigen, for example by killing a virus-infected host cell displaying the respective antigens on its surface. T-cells also assist other cells in immunologic processes such as macrophage activation or differentiation of B-cells into plasma cells (McHeyzer-Williams, Pelletier et al. 2009) .

Our group documented substantial differences in regulation of gene expression in the three tissues, the authors finding that 69 to 80% of the genetic regulatory effects are cell-type specific (Dimas, Deutsch et al. 2009). RTC results on the same dataset reflect the high extent of *cis* eQTL tissue-specificity (Table 5.2). As such, of the total 78 nonredundant interval-disease combinations with confident evidence of causal regulatory effects (RTC score ≥ 0.9), only 5 (6.4%) were shared across all three tissues. The pairwise tissue overlap of RTC results mirrors as expected the biological properties of the cell-types compared. Specifically, B-cells and T-cells share more GWAS relevant regulatory effects (16.7% of the total) compared to any of the other pairwise combinations (2 shared intervals between B-cells and fibroblasts and 3 examples common to T-cells and fibroblasts). Most of the confident RTC results (around 70%) are cell-type specific. This shows that as predicted, detecting causal regulatory effects for complex trait associations is highly tissue-dependent.

| | | #Interval-Disease (RTC ≥ 0.9) | % Total |
|---|---|---|---|
| **3 tissues** | B cell - T cell - Fibro | 5 | 6.4 |
| | | | |
| **2 tissues only** | B cell - T cell | 13 | 16.7 |
| | B cell - Fibro | 2 | 2.6 |
| | T cell - Fibro | 3 | 3.8 |
| | | | |
| **1 tissue only** | B cell | 21 | 26.9 |
| | T cell | 15 | 19.2 |
| | Fibro | 19 | 24.3 |
| | | | |
| **Total RTC ≥ 0.9** | B cell | 41 | |
| | T cell | 36 | |
| | Fibro | 29 | |
| **Union of total RTC ≥ 0.9** | | 78 | |

**Table 5.2. Tissue shared and tissue-specific interval-disease combinations with high RTC score.** RTC results with a score ≥ 0.9 are overlapped and compared across tissues. The significant tissue-specific component (~70% of results are found in one tissue only) implies that predicting causal regulatory effects for GWAS loci is highly dependent on the tissue where expression is determined.

In the following sections, I present the best RTC results for each tissue and focus on some interesting biological examples.

## 5.2 B-cell results

Table 5.3 summarizes the most confident *cis* results in B-cells ordered by RTC score. I detect SNP-gene combinations passing the 0.9 RTC threshold for 41 interval-disease combinations of the 149 tested. Among the 41 confident signals, 21 (51%) are only found in B-cells at this score level. Candidate genes already suspected to have a role in disease susceptibility are confirmed by the RTC and additional candidates revealed. Within the same recombination hotspot interval on chromosome 8 (chr8:11374002-11504000), two significant genome-wide SNP associations exist, one with systemic lupus erythematosus (Hom, Graham et al. 2008)(rs13277113) and the other one with rheumatoid arthritis (rs2736340) (Gregersen, Amos et al. 2009). The C8orf13-BLK locus

| GWAS SNP | Complex Trait | Gene | RTC | Chr |
|---|---|---|---|---|
| rs10903129 | Cholesterol, total | TMEM50A | 0.96 | 1 |
| rs7512898 | Electrocardiographic conduction measures | TNNT2 | 0.91 | 1 |
| rs13160562 | Alcohol dependence | ERAP1 | 1 | 5 |
| rs7731657 | Fasting plasma glucose | CDC42SE2 | 0.96 | 5 |
| rs9272346 | Type 1 diabetes | HLA-DQB1 | 1 | 6 |
| rs3135388 | Multiple sclerosis | HLA-DRB5 | 1 | 6 |
| rs2517713 | Nasopharyngeal carcinoma | HLA-A | 0.99 | 6 |
| rs9272219 | Schizophrenia | HLA-DQA2 | 0.99 | 6 |
| rs3129934 | Multiple sclerosis | HLA-DRB5 | 0.98 | 6 |
| rs8321 | AIDS progression | HLA-G | 0.98 | 6 |
| rs2227139 | Hematological parameters | HLA-DRB5 | 0.97 | 6 |
| rs2523393 | Multiple sclerosis | TRIM27 | 0.96 | 6 |
| rs9268480 | Ulcerative colitis | HLA-DQB1 | 0.95 | 6 |
| rs9264942 | HIV-1 control | NFKBIL1 | 0.95 | 6 |
| rs2187668 | Systemic lupus erythematosus | HLA-DQA2 | 0.94 | 6 |
| rs2269426 | Plasma eosinophil count | HLA-DQA2 | 0.93 | 6 |
| rs13437082 | Height | C6orf48 | 0.93 | 6 |
| rs7743761 | Ankylosing spondylitis | HLA-C | 0.93 | 6 |
| rs9461688 | Protein quantitative trait loci | C6orf48 | 0.90 | 6 |
| rs2237349 | Attention deficit hyperactivity disorder | CREB5 | 0.99 | 7 |
| rs17145738 | Triglycerides | BCL7B | 0.94 | 7 |
| rs13277113 | Systemic lupus erythematosus | C8orf13 | 1 | 8 |
| rs2736340 | Rheumatoid arthritis | C8orf13 | 1 | 8 |
| rs216345 | Bipolar disorder | NUDT2 | 1 | 9 |
| rs10781500 | Ulcerative colitis | CARD9 | 1 | 9 |
| rs4130590 | Bipolar disorder | SLC2A8 | 0.99 | 9 |
| rs7871764 | Height | NUDT2 | 0.94 | 9 |
| rs1927702 | Body mass index | C9orf52 | 0.92 | 9 |
| rs4977574 | Myocardial infarction (early onset) | CDKN2A | 0.91 | 9 |
| rs7481311 | Weight | LIN7C | 1 | 11 |
| rs5215 | Type 2 diabetes | C11orf58 | 0.96 | 11 |
| rs11602954 | Mean platelet volume | ATHL1 | 0.95 | 11 |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 |
| rs8020441 | Cognitive performance | ATP5S | 1 | 14 |
| rs10133111 | Brain imaging in schizophrenia (interaction) | HSP90AA2 | 0.96 | 14 |
| rs748404 | Lung cancer | TGM5 | 0.93 | 15 |
| rs2290400 | Type 1 diabetes | GSDML | 0.97 | 17 |
| rs199533 | Parkinsons disease | ILMN_25445 | 0.93 | 17 |
| rs2014572 | Hyperactive-impulsive symptoms | VN1R1 | 0.92 | 19 |
| rs6060369 | Height | UQCC | 0.99 | 20 |
| rs5751901 | Protein quantitative trait loci | GGT4P | 1 | 22 |

**Table 5.3. Candidate B-cell results.** Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and B-cell expression data from the GenCord.

has already been associated to lupus and confirmed in the initial RTC analysis on LCLs derived from HapMap 3 CEU individuals (Nica, Montgomery et al. 2010). The rheumatoid arthritis SNP identified more recently, scores equally high (RTC = 1) with the same gene of unknown function, *C8orf13*. Furthermore, the two SNPs are in very high LD ($r^2$ = 0.95, D' = 0.99) suggesting that most likely they are tagging the same functional variant, which plays an important role in the two autoimmune diseases.

The RTC applied on GenCord B-cells recovers another suspected autoimmune disease effect, namely the implication of *CARD9* (caspase recruitment domain family, member 9) in ulcerative colitis risk (Zhernakova, Festen et al. 2008). The GWAS SNP rs10781500 on chromosome 9 scores best (RTC = 1) with this gene, which is a very plausible candidate for inflammatory bowel diseases (including ulcerative colitis). *CARD9* was shown to be essential in the process of stimulating the innate immune signalling by intracellular and extracellular pathogens (Underhill and Shimada 2007). Studies in mice documented the role of *CARD9* in contributing to cytokine production via MAPK activation (Hsu, Zhang et al. 2007) or alternatively, leading to NF-κβ activation through the syk-CARD9 interaction (Hara, Ishihara et al. 2007). Given these, disrupting *CARD9* signalling by modifying the gene's expression levels appears to be a very probable mechanism leading to a deficient immune response predisposing to disease. Nevertheless, the authors of the GWAS study highlight that the extended 120 kb haplotype where the susceptibility SNP resides includes in addition to *CARD9* a few other genes that cannot be confidently excluded, namely *GPSM1*, *PSM1*, *LOC728489*, *SNAPC4*, *SDCCAG3*, *PMPCA*, *INPP5E* and *KIAA0310* (Zhernakova, Festen et al. 2008). Interestingly, for one of these genes, *INPP5E* (inositol polyphosphate-5-phosphatase), the RTC method provides supporting evidence in T-cells (RTC score = 0.9) and a modest effect in fibroblasts (RTC score = 0.76) for the same hotspot interval (chr9:138377986-138526984) where the GWAS SNP rs10781500 resides. *INPP5E* has also been shown to mediate cell responses to various stimulations (Kong, Speed et al. 2000). Both genes at the 9q34.3 locus would therefore merit further investigation for determining the causative roles behind the disease association.

The association of the *UQCC* (ubiquinol-cytochrome c reductase complex chaperone) locus on chromosome 20 with human height is one of the most robust signals for this trait, repeatedly replicated across multiple studies (Gudbjartsson, Walters et al. 2008;

Sanna, Jackson et al. 2008; Soranzo, Rivadeneira et al. 2009). The gene encodes a transmembrane protein and evidence from studies in mouse embryonic stem cells shows that the gene is down regulated in the presence of FGF2 (Vetter and Wurst 2001), which acts together with bone morphogenic proteins and *Hox* gene products to initiate and promote skeleton growth (Sanna, Jackson et al. 2008). The RTC successfully recovers the causal regulatory effect at this convincing gene locus with a score of 0.99 in B-cells only.

## 5.3 T-cell results

Using gene expression data from T-cells, I detect SNP-gene associations passing the 0.9 RTC score threshold for 36 interval-gene combinations of the 150 tested (discoveries sorted by RTC are presented in Table 5.4). Of these, results for 15 recombination hotspot intervals (41.7%) are restricted to T-cells, denoting once more the important role of the tissue-type in detecting disease relevant regulatory effects. As expected, the RTC method reveals candidate genes for a variety of autoimmune conditions, but also other interesting traits, some already flagged in the literature.

A recent GWAS identified significant associations at the 5q31 locus with osteoporosis risk (Guo, Tan et al. 2010). The strongest associated SNP, rs13182402 maps within the gene *ALDH7A1* (aldehyde dehydrogenase 7 family, member A1), that plays a major role in degrading and detoxifying acetaldehyde generated by alcohol metabolism and lipid peroxidation. Acetaldehyde was shown to inhibit osteoblast proliferation and result in decreased bone formation in murine and human bone marrow cultures (Giuliani, Girasole et al. 1999), making it thus a plausible candidate for explaining the disease association. We provide further evidence for this hypothesis by detecting a strong likely causal regulatory effect (RTC score = 1) for the same GWAS SNP with *ALDH7A1*. The effect is only detectable in T-cells and it suggests that rs13182402 predisposes to osteoporosis by substantially affecting the expression of *ALDH7A1*.

Similarly, the RTC recovers and confirms the implication of another suspected susceptibility gene, *GPR22* (G protein-coupled receptor 22) in osteoarthritis (Kerkhof, Lories et al. 2010). The most significant associated GWAS SNP, rs3815148 maps in an intronic region of a very large gene, *COG5* (component of oligomeric golgi complex 5) spanning ~3.6 Mb on the reverse strand of chromosome 7. However, the same SNP is

| GWAS SNP | Complex Trait | Gene | RTC | Chr |
|---|---|---|---|---|
| rs3890745 | Rheumatoid arthritis | MMEL1 | 0.98 | 1 |
| rs6435862 | Neuroblastoma (high-risk) | BARD1 | 0.97 | 2 |
| rs10495928 | Hemoglobin | SOCS5 | 0.91 | 2 |
| rs3772130 | Cognitive performance | IQCB1 | 1 | 3 |
| rs13182402 | Osteoporosis | ALDH7A1 | 1 | 5 |
| rs7731657 | Fasting plasma glucose | CDC42SE2 | 0.99 | 5 |
| rs13160562 | Alcohol dependence | ERAP1 | 0.91 | 5 |
| rs1321311 | Electrocardiographic traits | FGD2 | 1 | 6 |
| rs3135388 | Multiple sclerosis | HLA-DRB5 | 1 | 6 |
| rs2517713 | Nasopharyngeal carcinoma | HLA-G | 1 | 6 |
| rs9272346 | Type 1 diabetes | HLA-DQB1 | 0.99 | 6 |
| rs9264942 | HIV-1 control | EHMT2 | 0.99 | 6 |
| rs3129934 | Multiple sclerosis | HLA-DRB5 | 0.98 | 6 |
| rs13437082 | Height | HLA-C | 0.97 | 6 |
| rs2523393 | Multiple sclerosis | HLA-G | 0.97 | 6 |
| rs10484554 | Psoriasis | HLA-C | 0.96 | 6 |
| rs13194053 | Schizophrenia | BTN3A2 | 0.96 | 6 |
| rs9461688 | Protein quantitative trait loci | HLA-C | 0.96 | 6 |
| rs12216125 | Serum markers of iron status | BTN3A2 | 0.95 | 6 |
| rs9268480 | Ulcerative colitis | HLA-DQB1 | 0.95 | 6 |
| rs2227139 | Hematological parameters | HLA-DRB5 | 0.94 | 6 |
| rs8321 | AIDS progression | HLA-A | 0.92 | 6 |
| rs741301 | Diabetic nephropathy | GPR141 | 0.98 | 7 |
| rs3815148 | Osteoarthritis | GPR22 | 0.98 | 7 |
| rs4130590 | Bipolar disorder | SH2D3C | 0.97 | 9 |
| rs1927702 | Body mass index | BNC2 | 0.93 | 9 |
| rs10781500 | Ulcerative colitis | INPP5E | 0.90 | 9 |
| rs703842 | Multiple sclerosis | FAM119B | 1 | 12 |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 |
| rs1402279 | Smoking behavior | OSBPL8 | 0.92 | 12 |
| rs1378942 | Diastolic blood pressure | C15orf39 | 0.91 | 15 |
| rs4785763 | Melanoma | CDK10 | 0.99 | 16 |
| rs11648785 | Tanning | CDK10 | 0.98 | 16 |
| rs199533 | Parkinsons disease | NSF | 0.91 | 17 |
| rs8099917 | Response to Hepatitis C treatment | PSMC4 | 0.95 | 19 |
| rs5751614 | Height | BCR | 0.97 | 22 |

**Table 5.4. Candidate T-cell results.** Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and T-cell expression data from the GenCord.

associated with differential expression levels of a nearby gene on the forward strand encoding a G protein-coupled receptor (*GPR22*). The RTC framework provides compelling evidence that the GWAS SNP acts by modulating *GPR22* expression, as revealed by the 0.98 RTC score on the interval tested (chr7: 106574287-107092285). Results from experimental work in mouse models strengthen the potential role of *GPR22* in osteoarthritis pathology. For example, immunohistochemistry experiments showed that the GPR22 protein was absent in normal mouse articular cartilage or synovium, but GPR22-positive chondrocytes (the only cell-type found in cartilage) were detected in osteophytes (bony projections usually formed along joints) in instability-induced osteoarthritis and in the upper layers of the articular cartilage of mouse knee joints challenged with papain or albumin treatment (Kerkhof, Lories et al. 2010). All these lines of evidence point towards *GPR22* as having a causative disease role rather than *COG5*, the gene where the GWAS SNP resides.

Finally, another interesting result of the RTC analysis is the one linking the *FAM119B* (family with sequence similarity 119, member B) gene to multiple sclerosis susceptibility. Multiple sclerosis is an immune-mediated disorder whereby the body's own immune system attacks and damages the myelin sheaths around the axons in the brain and spinal cord. This severe disease of the central nervous system is characterized by myelin loss, chronic inflammation, axonal and oligodendrocyte pathology, and progressive neurological dysfunction (Oksenberg and Baranzini 2010). Although mechanisms involved in the disease process are relatively well described, very little is known about what causes the disease. Recently, a handful of GWAS studies revealed a subset of well-replicating associations, which remain largely elusive with respect to the genes whose activity they disrupt. One significant risk associated locus has been found on chromosome 12q13–14 in a gene-dense region of very high LD (2009). The most significant SNP in this region (rs703842) maps to the 3' UTR of the *METTL1* (methyltransferase-like protein 1) gene, 1.76 kb upstream of the *CYP27B1* (cytochrome P450 family 27 subfamily B) gene. While not being able to confidently exclude the other genes in the 12q13-14 haplotype (17 genes in total), the authors propose *CYP27B1* as the strongest causative candidate based on current genetic, immunological and epidemiological evidence. *CYP27B1* encodes an enzyme which hydroxylates 25-hydroxyvitamin D into its bioactive form, $1,25(OH)_2D$. This, along with the vitamin D

endocrine system, have been shown to play an important role in the prevention of disease onset and progression of autoimmune conditions modelled in the mouse (Lemire and Archer 1991). Furthermore, the link between vitamin D deficiency and increased multiple sclerosis incidence (van der Mei, Ponsonby et al. 2007) as well as the association between common variants in *CYP27B1* and risk of Type 1 diabetes (Bailey, Cooper et al. 2007) - also an autoimmune condition - advocate for the functional role of this gene in disease. With the RTC method, an additional candidate susceptibility gene in the 12q13-14 haplotype is discovered. Having a maximum RTC score with the same risk SNP rs703842 (RTC = 1) both in T-cells and fibroblasts, *FAM119B* appears as a noteworthy likely causal gene with a regulatory effect. Unfortunately, little is known about the function of this gene. However, a recent study analyzing the whole blood mRNA transcriptome of 99 untreated multiple sclerosis patients supports the RTC discovery (Gandhi, McKay et al. 2010). The authors find evidence for specific dysregulation of T-cell pathways in the trait pathogenesis. Of the 17 genes at the 12q13-14 locus, they can quantify expression data in leukocytes for 13 genes and one of them, *FAM119B* is expressed at significantly lower levels in the susceptibility haplotype (P-value < $10^{-14}$). This is yet another example of a non-intuitive disease susceptibility candidate, as the gene where the risk GWAS SNP maps to has no functional relevance to the trait. Instead, the GWAS SNP affects another proximal gene and the RTC methodology is helpful in discerning between such cases.

## 5.4 Fibroblast results

Finally, fibroblast expression data was also used to test for potential explanatory disease effects via regulatory mechanisms. Of the 144 recombination hotspot intervals harbouring nonredundant disease risk loci, SNP-gene associations passing the 0.9 RTC threshold were detected for 29 interval-disease combinations. Of these, 19 (65%) were confined to this tissue, reiterating the RTC tissue-specificity observed previously also in B-cells and T-cells. The most confident discoveries, sorted by RTC are listed in Table 5.5.

Notably, two signals for multiple sclerosis score a maximum RTC of 1 in this cell-type. One of the two (rs703842 associated with *FAM119B* on chromosome 12q13-14) has been described in the previous section as a plausible candidate supported additionally by

recent evidence from the analysis of T-cell expression in untreated individuals with the disease. The common high score (RTC = 1) attributed to the same SNP-gene association in both T-cells and fibroblasts indicates the likelihood of potentially shared disease-relevant biological properties between the two tissues. Interestingly, both leukocytes and fibroblasts produce type I interferons, ubiquitous cytokines released following exposure to a stimulus (pathogen or tumour cell) to trigger an immune response (Meyer 2009). More specifically, fibroblasts produce interferon (INF) beta, a type I interferon that has been used extensively over the past decades as an effective first-line therapy against relapsing-remitting multiple sclerosis (Zhang and Markovic-Plese 2010). This common function of interferons - stimulating or inhibiting a variety of genes involved in immunity-related conditions - might partially explain the surprising usefulness of fibroblast expression in explaining associations with multiple sclerosis.

The SNP rs744166 on chromosome 17q21.1 maps to the first intron of the *STAT3* (signal transducer and activator of transcription 3) gene and has been associated to multiple sclerosis by studying a high-risk isolated Finnish population (Jakkula, Leppa et al. 2010). The role of *STAT3* in disease predisposition seems very likely, given its suspected implication in another autoimmune disorder (Crohn's disease (Barrett, Hansoul et al. 2008)) and the evidence from mouse studies where targeted deletion of the gene in CD4+ T-cells prevented the development of experimental autoimmune encephalomyelitis, the murine model of multiple sclerosis (Liu, Lee et al. 2008). However, another study in an independent Spanish population investigated the role of common variants in *STAT3* in multiple sclerosis and the two clinical subtypes of inflammatory bowel disease, ulcerative colitis and Crohn's disease (Cenit, Alcina et al. 2010). While *STAT3* polymorphisms confirmed the gene's implication in colitis and Crohn's, the authors found no evidence for a major role of this gene in multiple sclerosis. The RTC pinpoints the existence of a regulatory effect in another proximal gene, also a member of the STAT family of transcription factors - *STAT5* (signal transducer and activator of transcription 5A). rs744166 scores an RTC of 1 with this gene in fibroblasts only, making it also an interesting candidate. Functional studies in mouse have shown that *STAT5* mediates the antiapoptotic effects of methylprednisolone (a synthetic glucocorticoid agonist used widely for the clinical therapy of spinal cord injuries and multiple sclerosis) on oligodendrocytes (Xu, Chen et al. 2009). Overexpression of an activated form of *STAT5* prevents oligodendrocyte cell death whereas knocking down this gene leads to

| GWAS SNP | Complex Trait | Gene | RTC | Chr |
|---|---|---|---|---|
| rs6537837 | Major depressive disorder | UBL4B | 1 | 1 |
| rs1390401 | Height | JMJD4 | 0.92 | 1 |
| rs3197999 | Crohns disease | WDR6 | 0.92 | 3 |
| rs4380451 | Bipolar disorder | CMTM7 | 0.91 | 3 |
| rs4143832 | Plasma eosinophil count | HSPA4 | 0.97 | 5 |
| rs2517713 | Nasopharyngeal carcinoma | HLA-A | 1 | 6 |
| rs2523393 | Multiple sclerosis | HLA-F | 0.99 | 6 |
| rs7743761 | Ankylosing spondylitis | IER3 | 0.99 | 6 |
| rs3130340 | Bone mineral density (spine) | HLA-DMB | 0.97 | 6 |
| rs3131379 | Systemic lupus erythematosus | HLA-DMB | 0.96 | 6 |
| rs9461688 | Protein quantitative trait loci | IER3 | 0.96 | 6 |
| rs742132 | Serum uric acid | HIST1H4C | 0.95 | 6 |
| rs9264942 | HIV-1 control | IER3 | 0.95 | 6 |
| rs9469220 | Crohns disease | HLA-DMA | 0.94 | 6 |
| rs12191877 | Psoriasis | HLA-C | 0.93 | 6 |
| rs3131296 | Schizophrenia | HLA-DMB | 0.92 | 6 |
| rs198846 | Hemoglobin | HIST1H2BH | 0.92 | 6 |
| rs703842 | Multiple sclerosis | FAM119B | 1 | 12 |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 |
| rs1994090 | Parkinsons disease | C12orf4 | 0.98 | 12 |
| rs10444502 | Biochemical measures | RFC5 | 0.92 | 12 |
| rs8020441 | Cognitive performance | CDKL1 | 0.92 | 14 |
| rs3825932 | Type 1 diabetes | CTSH | 0.95 | 15 |
| rs744166 | Multiple sclerosis | STAT5A | 1 | 17 |
| rs758642 | Smoking behavior | OR1A1 | 1 | 17 |
| rs8073783 | Conduct disorder (interaction) | KIF2B | 0.98 | 17 |
| rs2191566 | Acute lymphoblastic leukemia (childhood) | ZNF155 | 0.91 | 19 |
| rs1555322 | Attention deficit hyperactivity disorder | TRPC4AP | 0.94 | 20 |
| rs5751614 | Height | BCR | 0.97 | 22 |

**Table 5.5. Candidate fibroblast results.** Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and fibroblast expression data from the GenCord.

the loss of the protective effect. Thus, both genes (*STAT3* and *STAT5*) could be further considered as potential key determinants of disease risk. Additional studies of the interactions between each of the two gene products with other transcription factors might provide a better insight into the disease mechanisms characteristic for multiple sclerosis or other autoimmune traits.

Finally, for a subset of the GWAS SNPs tested in this chapter (N = 9) the RTC reveals identical gene candidates independently in at least two of the GenCord tissues (Table 5.6). Such consistent examples deserve special attention in future functional studies focusing on the respective genomic regions.

| GWAS SNP | Complex Trait | Gene | RTC | Chr | Tissue |
|---|---|---|---|---|---|
| rs13160562 | Alcohol dependence | ERAP1 | 1 | 5 | B-cells |
| rs13160562 | Alcohol dependence | ERAP1 | 0.91 | 5 | T-cells |
| rs7731657 | Fasting plasma glucose | CDC42SE2 | 0.99 | 5 | T-cells |
| rs7731657 | Fasting plasma glucose | CDC42SE2 | 0.96 | 5 | B-cells |
| rs2227139 | Hematological parameters | HLA-DRB5 | 0.97 | 6 | B-cells |
| rs2227139 | Hematological parameters | HLA-DRB5 | 0.94 | 6 | T-cells |
| rs3129934 | Multiple sclerosis | HLA-DRB5 | 0.98 | 6 | B-cells |
| rs3129934 | Multiple sclerosis | HLA-DRB5 | 0.98 | 6 | T-cells |
| rs3135388 | Multiple sclerosis | HLA-DRB5 | 1 | 6 | B-cells |
| rs3135388 | Multiple sclerosis | HLA-DRB5 | 1 | 6 | T-cells |
| rs9268480 | Ulcerative colitis | HLA-DQB1 | 0.95 | 6 | B-cells |
| rs9268480 | Ulcerative colitis | HLA-DQB1 | 0.95 | 6 | T-cells |
| rs9272346 | Type 1 diabetes | HLA-DQB1 | 1 | 6 | B-cells |
| rs9272346 | Type 1 diabetes | HLA-DQB1 | 0.99 | 6 | T-cells |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 | B-cells |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 | T-cells |
| rs11171739 | Type 1 diabetes | RPS26 | 0.98 | 12 | Fibro |
| rs5751614 | Height | BCR | 0.97 | 22 | T-cells |
| rs5751614 | Height | BCR | 0.97 | 22 | Fibro |

**Table 5.6. RTC signals consistent across at least two tissues.** Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects consistent in at least two GenCord tissues.

## 5.5 Conclusions

In this chapter I explored the tissue-dependent value of gene expression variation in predicting candidate disease genes. The RTC methodology was applied to expression data in B-cells, T-cells and fibroblasts derived from 75 Swiss individuals and GWAS data

from the NHGRI catalogue. As expected given the demonstrated high extent of *cis* eQTL tissue-specificity and the tissue-restricted manifestation of diseases, the results of the RTC analysis are overall highly cell-type specific. Of the total number of confident discoveries passing the 0.9 RTC score threshold, roughly 70% of the predicted effects are found only in one cell-type. In each of the three tissues, this corresponds to approximately 50% of all confident effects being specific per cell-type. The distribution of the scores and the pairwise comparisons of RTC discoveries mimic the biological properties of the tissues tested, whereby B-cells and T-cells share as expected proportionally more causal regulatory effects, many for immunity-related conditions. Each of the three cell-types permits the discovery of candidate disease genes whose differentiated regulation is affected by GWAS SNPs. The RTC confirms previously suspected expression mediated disease effects but also facilitates the informative prioritization of novel candidate causal genes, some already having plausible functional justification from experimental studies.

I highlight the risks of misinformed candidate gene prediction by relying solely on genetic distance criteria and give examples where the RTC can help distinguish likely causal effects from genes coincidentally residing closest to the GWAS SNP loci. Finally, the data suggests that establishing relevance of a cell-type to a complex trait is not trivial. The current knowledge about disease biology is generally limited and thus, predicting candidate disease tissues could be as unsuccessful as candidate gene approaches have proved to be for complex diseases. Additionally, the estimated fair amount (~30%) of tissue shared regulatory variation should encourage the interrogation of any available cell-type for potential regulatory disease effects. In this sense, I conclude by presenting a few unexpected associations revealed by the RTC, which could offer new insights into disease aetiology.