Review

Effects of Type 1 Diabetes Risk Alleles on Immune Cell Gene Expression

Ramesh Ram [1,2] **and Grant Morahan** [1,2,*]

[1] Centre for Diabetes Research, Harry Perkins Institute of Medical Research, Nedlands, WA 6009, Australia; ramesh.ram@uwa.edu.au

[2] Centre of Medical Research, University of Western Australia, Nedlands, WA 6009, Australia

[*] Correspondence: grant.morahan@uwa.edu.au; Tel.: +61-861-510-756

Abstract: Genetic studies have identified 61 variants associated with the risk of developing Type 1 Diabetes (T1D). The functions of most of the non-HLA (Human Leukocyte Antigen) genetic variants remain unknown. We found that only 16 of these risk variants could potentially be linked to a protein-coding change. Therefore, we investigated whether these variants affected susceptibility by regulating changes in gene expression. To do so, we examined whole transcriptome profiles of 600 samples from the Type 1 Diabetes Genetics Consortium (T1DGC). These comprised four different immune cell types (Epstein-Barr virus (EBV)-transformed B cells, either basal or after stimulation; and cluster of differentiation (CD)4+ and CD8+ T cells). Many of the T1D-associated risk variants regulated expression of either neighboring (*cis*-) or distant (*trans*-) genes. In brief, 24 of the non-HLA T1D variants affected the expression of 31 nearby genes (*cis*) while 25 affected 38 distant genes (*trans*). The effects were highly significant (False Discovery Rate *p* < 0.001). In addition, we searched in public databases for expression effects of T1D single nucleotide polymorphisms (SNPs) in other immune cell types such as CD14+ monocytes, lipopolysaccharide (LPS) stimulated monocytes, and CD19+ B cells. In this paper, we review the (expression quantitative trait loci (eQTLs) associated with each of the 60 T1D variants and provide a summary of the genes impacted by T1D risk alleles in various immune cells. We then review the methodological steps involved in analyzing the function of genome wide association studies (GWAS)-identified variants, with emphasis on those affecting gene expression. We also discuss recent advancements in the methodologies and their advantages. We conclude by suggesting future study designs that will aid in the study of T1D risk variants.

Keywords: type 1 diabetes; eQTLs; B-cells; T-cells; dendritic cells

1. Introduction

Four [1–4] genome wide association studies (GWAS), a linkage study [5], and five [6–10] other studies have identified 61 genetic variants that confer risk of Type 1 Diabetes (T1D). In Figure 1a, we summarize 60 non-HLA (Human Leukocyte Antigen) T1D risk loci identified so far. GWAS typically involves (a) recruiting patient cohorts accurately diagnosed with the disease; (b) recruitment of healthy controls; (c) sample collection and genotyping of individuals, typically at 500,000 markers (Single Nucleotide Polymorphisms (SNPs)); (d) Imputation of potentially millions more SNPs; and (e) performing association analysis of each SNP with the disease or trait of interest. All SNPs that exceed stringent thresholds of significance ($p < 5 \times 10^{-8}$, suggestive $p < 1 \times 10^{-5}$) are deemed to be associated with the disease. It is understood that known covariates (e.g., age, sex) and unknown confounders (e.g., relatedness) could limit gene discovery, however, statistical approaches such as principal component analysis (PCA) [11], clustering analysis [12], and linear mixed models [13] have helped overcome such problems. Validation cohorts and studies in different populations provide confirmation and improve confidence of gene-disease associations. However, all SNPs identified

above genomic significance thresholds are treated as important. The results of GWAS are commonly summarized as Manhattan plots where the X-axis is the genomic coordinates and the Y-axis is the negative logarithm of the associated *p*-value for each SNP. Among the SNPs that exceed the significance threshold, each risk locus is defined by a lead SNP with maximum association signal (lowest *p*-value) and all SNPs in ±0.5 Mb of the lead SNP are excluded. Testing the genes nearest to the lead SNPs potentially reveals a link between the locus and the disease, even though this may not be the best method for establishing candidacy due to lack of evidence linking the lead SNP to the target gene. Nevertheless, in GWAS of T1D, lead SNPs were found adjacent to genes like *HLA* [14], *PTPN22* [15], *INS* [16], *CTLA4* [17], and *IL2RA* (also known as *CD25*) [18] that have all been reported since the pre-GWAS era.
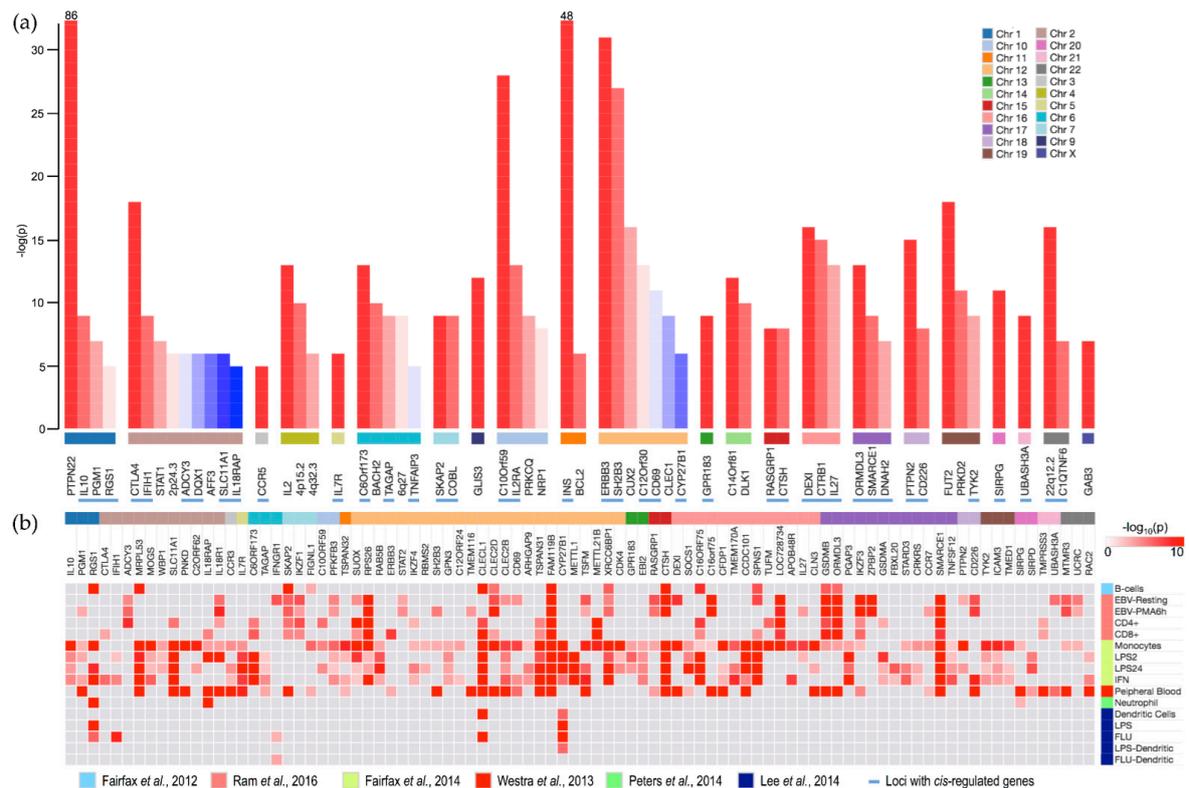

Figure 1. Summary of Type 1 Diabetes (T1D) loci and candidate *cis*-regulated genes in immune cells: (**a**) Non-HLA (human leukocyte antigen) T1D risk genes along with reported genome wide association studies (GWAS) association *p*-values; (**b**) *Cis*-regulated genes that were validated in more than one cell type. The grey cells indicate no significant regulation found. Shades of red represent the negative $\log_{10}(P)$ association. The loci and *cis*-regulated genes are grouped according to chromosome. All loci with *cis*-regulated genes are highlighted by blue dashes.

The T1D risk loci are interconnected (directly or indirectly) via networks; perturbations of these networks presumably cause disease. This could arise from the inability (or decreased ability) of the products of disease genes to interact with other gene products when such interactions are required. This may be caused by factors including: (a) functional variants; (b) gene expression changes; and (c) epigenetic aberrations such as differential methylation. Regarding effects on the protein sequence of the gene, amino acid changes caused by missense SNPs could either be tolerated or damaging; in both cases the protein's function could be impacted. Similarly, the risk SNPs may affect transcription rates of genes.

To formally establish disease gene candidacy, we performed a study [19] systematically evaluating each of 60 T1D risk loci. This revealed genes affected by missense SNPs and genes whose expression levels were perturbed with respect to T1D SNPs. We found that only 16 of the risk variants could potentially be linked to a protein-coding change. Expression effects were tested using 600 samples from

the Type 1 Diabetes Genetics Consortium (T1DGC) comprising four cell types (a) EBV (Epstein-Barr virus) resting cells (b) PMA (Phorbol 12-myristate 13-acetate) stimulated EBV cells (c) cluster of differentiation (CD)4+ T-cells, and (d) CD8+ T-cells. In these analyses, each T1D SNP was tested against the whole transcriptome as individual traits and the association *p*-values were derived. The genes were categorized based on their proximity to the corresponding T1D SNPs as: (a) *cis*-regulated genes i.e., those within ±1 MB of the SNP and the remaining (b) *trans*-regulated genes. The false discovery rate (FDR) method [20] was employed to allow for multiple testing; genes with *FDR* < 0.05 were deemed significant. The aims of this analysis were to identify genes that were up or down regulated according to the risk SNP genotypes and to identify pathways and networks affected. Similar expression studies [21–26] have analyzed the effect of SNPs (inclusive of T1D SNPs) on expression of genes in eight cell types: (a) monocytes; (b) interferon-γ (IFN) stimulated monocytes; (c) lipopolysaccharide (LPS) stimulated monocytes; (d) B-cells; (e) neutrophils; (f) dendritic cells; (g) LPSs; and (h) FLU (Influenza virus) infected dendritic cells. A large meta-analysis of gene expression effects in peripheral blood was also presented [25]. The associations in peripheral blood serve as validation for associations found in other cell types. Individually, these studies had reported adjusted (FDR) and unadjusted *p*-values for the SNP-gene associations that were deemed significant by them using suitable FDR thresholds. We searched in these databases for gene expression effects of only T1D SNPs. Based on the evidence available, we compared *cis*- and *trans*- regulated genes of each one of the T1D SNPs in various immune cells and peripheral blood.

The results are summarized as genes regulated in multiple cells and cell-type specific genes. FDR corrections remain specific to each study and expression quantitative trait loci (eQTLs) deemed significant in each study were retrieved from their corresponding supplementary tables. In these supplementary tables, we looked for reported eQTLs involving either lead T1D SNPs or any SNPs in high linkage disequilibrium (LD) (i.e., r^2 > 0.8) with the lead T1D SNPs. The list of LD SNPs was obtained using LDlink [27]. Genes whose expression is associated with T1D SNPs in more than one cell type are considered validated. However, for cell-type specific genes, only eQTLs reported below arbitrary *p*-value thresholds of $p < 1 \times 10^{-4}$ (*cis*-regulated) and $p < 1 \times 10^{-8}$ (*trans*-regulated) were investigated due to their lack of validation in other cells; the excluded eQTLs reported outside this P range are nevertheless significant (as per the study-specific FDR thresholds). The lack of overlap of these eQTLs in multiple cells can always be due to lack of power, however, cell-specificity cannot be ruled out. Since genes in HLA have been studied elsewhere, the function of most of the non-HLA genetic variants remains unknown, thus, we focus only on the non-HLA T1D SNPs. Enrichment analysis was performed using the Molecular Signature Database (MsigDB) [28] to describe the possible biological mechanism of the identified *cis*- and *trans*- regulated genes.

2. Gene Expression Studies in Immune Cells

The T1D risk SNPs are summarized in Figure 1a and reviewed in (refer to Table 1 in ref. [5,19]). There has also been a flurry of publications [19,21–26,29–33] investigating the role of risk SNPs regulating gene expression. Of these, seven studies [19,21–26] focused on cells involved in the immune system and autoimmune diseases, hence, they are reviewed here. Firstly, we performed expression analysis in four immune cells [19]: EBV transformed lymphoblastoid cell lines (LCLs) (resting and PMA stimulated) and CD4+ and CD8+ T-cells of 600 subjects from the Type 1 Diabetes Genetics Consortium cohort (T1DGC). The cohort comprised families of affected siblings. We identified 76 genes *cis*-regulated by T1D SNPs with FDR < 0.05 (min *p* < 0.0008). Thirteen genes were *cis*-regulated in CD4+ and CD8+ cells, while 11 were *cis*-regulated in EBV and T-cells. In addition, we also identified 37 *trans*-regulated genes (min $p < 1 \times 10^{-8}$). Secondly, Fairfax and colleagues reported the expression effects in CD19+ B-cells and CD14+ monocytes from 239 healthy volunteers of European ancestry [21]. Among the list of significant associations reported, in the context of T1D SNPs there were 18 *cis*-regulated genes and 3 *trans*-regulated genes. Subsequently, they also reported expression analysis in CD19+ Monocytes in 432 healthy volunteers in [22] where monocytes were exposed to either interferon-γ or lipopolysaccharide

(LPS) for a short (2-h) or long (24-h) duration. The study concluded that the majority of the *cis* genes identified were condition-specific. Among these results, some 90 genes were *cis*-regulated in the context of T1D SNPs and only five genes were *trans*-regulated. A third study [23] reported expression effects in dendritic cells isolated from the peripheral blood monocytes (PBMCs) of 560 healthy individuals of mixed ancestry (European, Asian, or African-American) from the PhenoGenetic cohort. Dendritic cells were stimulated with LPS, influenza virus, or interferon-*β*. Among the 385 *cis*-eQTLs identified in this study, seven were associated with T1D SNPs. Fourthly, Peters et al. [24] reported eQTL effects in primary immune cells (CD14+ monocytes, CD19+ B-cells, CD4+, and CD8+ T-cells) in 180 subjects. A novel joint analysis found *cis*-eQTLs that were cell-type and condition-specific. Twenty of the *cis*-regulated genes reported in [19] were associated with T1D SNPs. Lastly, Westra et al. [25,26] performed a large scale meta-analysis of expression eQTLs in peripheral blood and reported *cis*- and *trans*- genetic associations using a less stringent threshold of FDR < 0.5, finding 77 *cis*- and 127 *trans*-regulated genes in association with T1D SNPs. We combined the results from the above studies such that each record would contain: SNP identifier (rsID), T1D locus name, gene name, unadjusted *p*-value, cell-type, and publication reference. These data are summarized in Supplementary Materials Table S1.

3. Candidate Gene Identification

This section is sub-divided into *cis*-regulated genes and *trans*-regulated genes.

3.1. Cis-Regulated Genes

Most studies considered *cis*-regulated genes as being located within the ±1 MB of the SNP position. Our search for genetic interaction associated with T1D SNPs in seven cell types [19,21–26] revealed 90 genes that showed significant (FDR < 0.05) association in more than one cell type and 34 genes that were cell type-specific. In Figure 1b, we show the 90 genes and their association with 40 T1D loci where unadjusted association *p*-values extracted from corresponding studies are indicated as a heatmap. These associations may be considered validated as they were replicated in multiple cell types. Interestingly, eleven of the 90 genes were located in chromosome band 12q13 (commonly referred to as the ERBB3 locus [34]). Additionally, cytogenetic bands 17q21 (*ORMDL3*) and 16p11 (*IL27*) included six and five *cis*-regulated genes, respectively. The remaining risk loci were associated with expression of fewer genes. Enrichment analyses were performed in the Canonical pathways of KEGG (www.genome.jp/kegg/) and Reactome (www.reactome.org) databases. In doing so, eight genes (*IFNGR1*, *IL7R*, *IL18R1*, *CCR3*, *CCR7*, *IL10*, *IL18RAP*, and *TNFSF12*) were associated with the cytokine-cytokine receptor interaction pathway (KEGG) ($p = 3.82 \times 10^{-8}$) while 12 genes (*IFNGR1*, *IL7R*, *SOCS1*, *TYK2*, *STAT2*, *PTPN2*, *CTLA4*, *RASGRP1*, *CTSH*, *ICAM3*, *CD226*, and *IFIH1*) were associated with immune system (Reactome) pathways ($p = 1.55 \times 10^{-7}$). The remaining genes were tested for enrichment in the Gene Ontology (GO) database. An additional 11 genes (*GPR183*, *IKZF3*, *RAC2*, *TSPAN32*, *SKAP2*, *IL27*, *SIRPG*, *SLC11A1*, *UBASH3A*, *CLEC2D*, and *CLEC2B*) were associated with GO regulation of "immune system process" (GO:0002682) ($p = 4.02 \times 10^{-18}$). In addition to immune system or response terms, significant (FDR < 0.05) enrichment was observed for "cell activation", "cell proliferation", "leucocyte/lymphocyte activation", and "defense response".

Only *cis*-genes reported with high confidence ($p < 1 \times 10^{-4}$) were selected for further analysis, irrespective of their reported FDR *p*-values. There were 34 such genes, the majority of which showed association only in peripheral blood cells. Cell-type specific *cis*-regulated genes are summarized in Figure 2.

Upon combining the list of cell-specific genes with ubiquitously expressed genes (*n* = 124 genes), there was an increase in the number of *cis*-regulated genes (from 3 to 7) associated with the 12q24 (*SH2B3*) locus. Enrichment analyses showed 40 of the 124 *cis*-genes were related to the immune system. Among the cell-type specific genes, *OAS2* (EBV-resting), *PTEN*, and *ITGB1* (CD8+) were involved in immune system pathways (Reactome) ($p = 1.51 \times 10^{-8}$). The genes *IFNGR1*, *SOCS1,* and *PTPN2* affected regulation of the interferon gamma (*IFNG*) signaling pathway (Reactome). We performed further separate analysis of non-immune related genes. In doing so, we found that *SULTA1*, *SULTA2*,

and *SUOX* affected the Sulphur metabolism pathway (KEGG) ($p = 4.4 \times 10^{-6}$). In a study [35] into a possible bacterial role in developing T1D, sequencing of fecal samples showed a higher number of reads mapping to sulfur metabolism in T1D cases compared to that of controls, suggesting interaction of microbiome and sulphur metabolism genes.
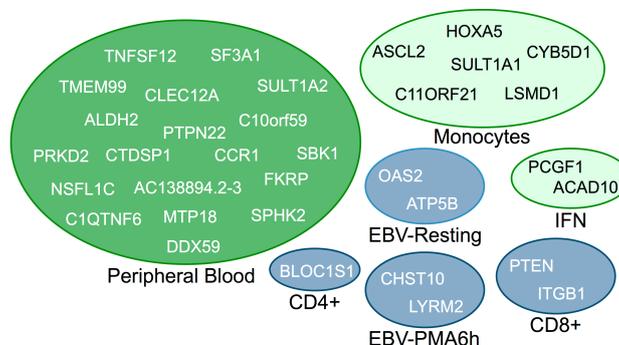

Figure 2. *Cis*-regulated genes that are cell-type specific. IFN denotes interferon-γ stimulated monocyte cells.

Additionally, twelve of the genes (*SULTA1*, *SULTA2*, *ATP5B*, *CLN3*, *SPHK2*, *PGAP3*, *MTMR3*, *ERBB3*, *STARD3*, *APOBR*, *ACAD10*, and *ORMDL3*) were also associated with lipid metabolic processes ($p = 4.37 \times 10^{-7}$). It was shown that people with T1D also displayed disturbances in plasma lipids such as hypertriglyceridemia and low high-density lipoprotein (HDL) cholesterol [36]. These analyses suggest a novel role of non-immune related genes mediating T1D.

3.2. Trans-Regulated Genes

Genes whose expression varied significantly in association with T1D SNP alleles, but were located more than 1 Mb from the SNP may be considered as *trans*-regulated genes. First, we searched for *trans*-regulated genes that are validated by their observations in multiple cell types (or studies) with FDR < 0.05. In searching for cell specific *trans*-regulated genes, we used a stringent threshold of $p < 1 \times 10^{-8}$ to deem the results as significant. There were ten such genes and their negative log(P) values of the associations in different cells are summarized in Table 1.

Table 1. List of *Trans*-regulated genes found with association in multiple cell types.

T1D SNP	Locus	Chr	Gene	B	EBV	PMA	CD4	CD8	Mono	LPS2	LPS24	IFN	Blood
rs7111341	*INS*	11p15.5	*AQP9*				6.0	5.6					
			IP6K2	14.2	18.2	11.9	17.8	20.7	16.9	7.7	13.3	10.0	
			LAP3P2	51.7			14.6	14.4	13.9				
rs11171739	*ERBB3*	12q13.2	*MAFG-AS1*		6.9	5.6	17.0	19.9					
			MIR130A			36.2	27.8	25.7	21.8				
			MIR1471		6.9	5.6	8.5	13.0					
			DPF2					6.8					12.0
rs3184504	*SH2B3*	12q24.12	*UBE2L6*		7.9	4.7							7.9
			STAT1		5.5								7.5
rs17696736	*C12Orf30*		*UBE2L6*		5.7								5.3

SNP—single nucleotide polymorphisms; B—CD19+ B cells; EBV—Epstein-Barr virus resting; PMA—Phorbol 12-myristate 13-acetate stimulated EBV; Mono—monocytes; LPS2, LPS24—lipopolysaccharide stimulated monocytes; IFN—interferon-γ stimulated monocytes.

Interestingly, the insulin locus SNP showed significant association with *trans*-regulated gene aquaporin-9 (*AQP9*) in both CD4 and CD8 positive cells. *AQP9* is down regulated by insulin in obese type 2 diabetes mellitus (T2DM) patients [37]. There were four genes and two micro-RNAs that were associated with the *ERBB3* (12q13.2) T1D locus. In particular, *IP6K2* (Inositol hexakisphosphate kinase-2) showed consistent association with 12q13.2 in all cells except peripheral blood. *IP6K2* is abundantly present in pancreatic beta cells and may be involved in regulation of insulin exocytosis.

Novel genes *LAP3P2*, a Leucine Aminopeptidase 3, and *MAFG-AS1*, a *MAF* gene, showed association in multiple cell types. *MAF* genes such as *MAFA* are known to promote pancreatic development and insulin transcription. The *trans*-regulation of *DPF2*, a zinc finger gene in CD4+ T-cells, was confirmed in peripheral blood. *UBE2L6*, a ubiquitin gene, and *STAT1* were *trans*-regulated by the same T1D SNPs located in the 12q24 locus, as also confirmed in peripheral blood [26]. Another ubiquitin gene, *UBASH3A*, is a well-known candidate gene for the 21q22.3 T1D locus [1]. In [38], it was found that the ubiquitin system could be associated with insulin signaling and might be affected in diabetes.

In cell-type specific *trans*-regulated genes, we identified 12 genes in EBV-B cells, seven genes in peripheral blood, two genes in CD4+ cells, and one gene in monocytes. The list of these genes is provided in Table 2, sorted according to association *p*-value along with the regulating T1D SNP. Enrichment analysis of *trans*-regulated genes alone showed three of the genes (*IRF8*, *ID2*, and *CCL5*) were validated targets of *c-MYC* transcriptional repression. In doing Gene Ontology enrichment, we found that ten of the *trans*-regulated genes were associated with "positive regulation of biosynthetic process".

Table 2. List of cell-type specific *trans*-regulated genes.

EBV Cells		Peripheral Blood	
SNP	***Trans*-Gene**	**SNP**	***Trans*-Gene**
rs1990760	*LOC643997*	rs1701704	*CCL5*
rs10499194	*TUBB6*	rs1701704	*CRLF3*
rs7804356	*SLC39A8*	rs2058660	*CYP2C19*
rs12251307	*DERA*	rs3184504	*FOS*
rs947474	*MEIS2*	rs3184504	*GBP4*
rs3842727	*ID2*	rs11171739	*MIF*
rs1738074	*IRF8*	rs3184504	*NALP12*
rs1265565	*NCOA7*	**CD4+**	
rs12908309	*FAHD1*	rs1265565	*ZMYM5*
rs2290400	*TEX9*	rs11711054	*GRAMD1B*
rs763361	*P2RY11*	**Monocytes**	
rs7221109	*EIF5A*	rs11171739	*KCTD11*

4. Alternative Methods for Studying eQTLs Associated with Disease SNPs

Recent studies [19,21–26] indicate that T1D-associated SNPs are likely to be eQTLs. Associations at T1D risk loci were consistent with eQTLs (both *cis* and *trans*) in three relevant immune-cell populations studied [19]: EBV—lymphoblastoid cell lines (LCLs) (unstimulated and stimulated) and CD4+ and CD8+ T-cells. In this review, we performed a comparison of the eQTLs in additional cells such as CD14+ monocytes (unstimulated and LPS- and IFN-gamma stimulated) [21,22], CD19+ B-cells [21], fluorescence-activated cell sorting (FACS)-purified neutrophils [23], and dendritic cells (unstimulated and stimulated) [24] to carefully examine the expression effects of T1D risk alleles on immune gene expression. The eQTL results were also compared against results reported in whole blood [25,26]. For each T1D locus, variants in high LD ($r^2 > 0.8$) with any of the lead T1D variants were also examined for overlap with eQTLs. The authors of each of these studies employed suitable thresholds to determine true-positive *cis* and *trans*-eQTLs. The analysis we performed was a simple one-to-one comparison of the eQTL association results (assuming they are true) between the cell-types aimed at compiling a summary of validated target genes by virtue of their co-occurrence in more than one cell-type or study. The eQTLs that could not be validated in multiple cells were assumed to be cell-specific. The techniques used in this analysis have advantages and disadvantages. In this section we explore the techniques by providing step-by-step methods where we discuss some alternative approaches that have, at times, better advantages for deducing eQTL associations with disease associated variants.

4.1. Prediction of Functionality of Disease Associated Variants

The first step in the analysis of disease associated SNPs is to test whether or not the SNPs have an impact on the function of the proteins encoded by the relevant genes [19,39]. This is commonly achieved by means of prediction algorithms that determine whether or not a variant is deleterious or benign by means of a score. There are two types of predictions available: protein-based and more recently nucleotide-based. For several years, protein-based sorting intolerant from tolerant (SIFT) [40] and Polymorphism Phenotyping 2 (PolyPhen-2) [41] were the only prediction methods which were applied to non-synonymous coding region SNPs to identify whether or not the amino acid changes were deleterious.

More recently, with advancements such as Encyclopedia of DNA Elements (ENCODE) [42] and Functional Annotation of Mouse 5 (FANTOM5) [43], tools such as Combined Annotation–Dependent Depletion (CADD) [44], Deleterious Annotation of genetic variants (DANN) [45], Functional Analysis Through Hidden Markov Models (FATHMM) [46], and Linear Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (LINSIGHT) [47] have been developed to predict functionality of variants outside the coding regions. These tools are based on supervised machine learning algorithms using a well-characterized training dataset. The prediction is usually given as a score, typically between 0 and 1, where scores >0.8 usually can be treated as functional. PredictSNP [48] and PredictSNP2 [49] are convenient web-based meta-predictors incorporating all available methods in arriving at a consensus prediction where a required variant can be searched using the corresponding rsIDs. Using these tools, it is possible to determine whether the identified lead GWAS SNP can be characterized prior to performing eQTL association testing.

4.2. Gene Expression Quantification

In earlier eQTL studies [19–23], whole-transcriptome profiling was performed on Illumina's HT-12v4 bead arrays. Briefly, the processing of the expression data involved quantile or Robust Spline (RSN) normalization [50], quality control, and filtering. Probes with a detection *p*-value of <0.01 (Illumina GenomeStudio Software, Illumina, San Diego, CA, USA) in at least 5% of the samples were retained for further eQTL analysis. Popularly, software such as ReMOAT [51] (Re-annotation and Mapping of Oligonucleotide Array Technologies) were used to assess probe quality and probes considered as "bad" are removed. Alternatively, probe sequences were tested for unique alignment to the transcriptome as described in [19] for consideration for eQTL association testing, and probes whose sequences contained SNPs were filtered out. These stringent probe-mapping strategies were employed to filter out false-positives due to primer-polymorphisms and cross-hybridizations. In the case of latter studies such as [24–26] using RNA-seq for transcriptome profiling, the processing steps differ. Briefly a (variance stability transformation) VST-normalization is applied to the read counts obtained from the mapped RNA-seq data, which is then regularized log (rlog) transformed using R package DESeq2 [52] prior to further analysis.

4.3. Batch Effect Correction and Removing Unwanted Variations

In both RNA-seq and microarrays, it is commonly known that batch effects and their influence on normalization can result in spurious findings. Approaches (e.g., ComBat method [53]) have been proposed to remove unwanted variation caused by differences between batches of samples. Principal variation component analysis (PVCA) [54] has also been used to detect and correct batch effects where the Principal Components (PCs) attributed to batching can be subtracted by regression and residual expression calculated for further analysis. Some additional methods have been proposed based on the use of spike-in negative control probes such as SQN (subset quantile normalization) [55] and RUV (remove unwanted variation) [56] to remove other unknown variations that limit the rate of eQTL detection. Furthermore, unknown hidden variables can be also be detected by surrogate variable analysis (SVA) as described by Leek et al. [57]. SVA method is implemented in the R package "sva" [58]. Sample label mix-ups are another common problem that can often cause reduction in power to detect eQTLs. In [59], the authors introduce a

method "MixupMapper" to correct such errors by comparing actual versus predicted gene expression for genes with very strong *cis*-eQTLs where the expressions are predicted solely based on SNPs.

4.4. Identify and Remove Known and Hidden Confounding Factors in the Normalized Expression Data

It is common practice to remove confounding factors from expression data as this limits the detection of true-positive eQTLs. To remove the effects of covariates such as age, sex, and HLA types, the R package "pedigreemm" [60] performs a mixed-effect modeling, accounting for relatedness between samples. This also allows for the calculation of the residual gene expression. To adjust for hidden factors, the residuals can be subjected to a Bayesian framework known as Probabilistic Estimation of Expression Residuals (PEER) [61] and the residual expression levels can be derived again after subtracting the estimated hidden factor contributions. This technique is popular, but an alternative approach is to use expression-derived PCs [62] to remove non-genetic expression variation. The significant ($p < 0.05$) PCs in the Tracy-Wisdom test [63,64] can be used as covariates during association testing or can be regressed out from the expression data. Some studies used arbitrarily either three, five, or up to ten PCs as covariates to remove hidden factors, but others (e.g., [21,22]) repeat eQTL analysis with varying numbers of PCs to identify the optimum number of PCs that maximize the number of eQTLs identified above suitable significance thresholds.

4.5. Expression Quantitative Trait Locus (eQTL) Analysis

eQTLs can have effects in *cis* and in *trans*. *Cis*-effects are taken when there are differences in expression levels of genes within 1 Mb of the associated SNP; *trans*-associations arise from relevant SNPs affecting expression of more distant genes, including genes on other chromosomes. MatrixEQTL [65] is currently the most popular eQTL analysis tool. It is a fast additive linear regression model for performing these tests for *cis* and *trans* associations separately. If the PCs and other covariates have not already been removed from the expression sets, then they can be used as covariates inside MatrixEQTL. Additively recoded (0,1,2) SNP data is used along with an expression matrix to perform the tests. Although there are options in MatrixEQTL to test expression against three genotypes per SNP, additive recoding detects eQTLs better. In the case of testing GWAS SNPs, it is better practice if the SNPs are additively recoded by the risk allele rather than the reference allele such that the beta coefficients provide the direction of expression regulation in relation to the risk allele. This information can be particularly useful for downstream pathway and enrichment analysis as well as to compare effect directions in different cell types. In a large meta-analysis study, Westra et al. [25,26] developed a pipeline where QTLs were determined by using Spearman rank correlation on genotype dosages in each cohort. Then, a meta-analysis was performed to combine the results by a weighted z-score method.

4.6. Statistical Significance and Permutation Analysis

Generally, all eQTL findings at a Benjamini-Hochberg adjustment false discovery rate (FDR) [20] under 0.001 are considered significant (as in [19]). FDR adjustment is separately applied for *cis*- and *trans*-associations. In some studies [23,24], probe-variant pairs with adjusted *p*-values less than 0.05 were deemed significant. However, this simple correction procedure is confounded by LD between SNPs tested, correlation between probes, and differences in minor allele frequencies (MAF) between the SNPs. To address these issues, permutation based strategies were introduced [25] to correct for multiple testing. For this purpose, the eQTL analyses are repeated up to 10,000 times using a R "*q*-value package" [66] with permuted sample labels and null *p*-values derived. The *q*-values are then derived over the null *p*-values, and eQTL associations with *q*-values < 0.05 are generally deemed significant.

4.7. Colocalization: Overlap between eQTL for a Gene and GWAS SNPs for Disease

There are a number of tools available to link the eQTL to a disease associated variant. For example, Sherlok [67] calculates a SNP-level Bayes factor using observed GWAS and eQTL *p*-values of SNPs to determine the likelihood that expression changes in the gene mediate disease risk as opposed to

the gene not being related to disease. Alternatively, co-localization directly evaluates whether two associations (GWAS and eQTL) in the same locus, observed in different cohorts, were due to the same underlying effect. The R package "coloc" [68] is a well-calibrated Bayesian framework that considers spatial similarities in association data across sets of SNPs; "gwas-pw" [69] is a similar method with the addition of hierarchical priors, and it optimizes model parameters; HEIDI/SMR15 [70], applies Mendelian randomization between traits.

4.8. Joint eQTL Analysis for Multiple Cell Types/Tissues

In this paper, we compared the eQTLs of T1D in different immune cells by comparing the results reported in different studies. This technique has its disadvantages where there is a possibility that cell-specific eQTLs were not detected in other cells due to incomplete power. Several approaches [71,72] have been developed to perform joint eQTL analysis in multiple cells or tissues primarily using a Bayesian framework to overcome these issues. These techniques were recently applied to perform analysis on 45 tissues [73], and data has been made available in the GTEx online portal (http://www.gtexportal.org). The GTEx data can be accessed to identify the eQTLs associated with disease associated SNPs as well as their proxy (high LD) SNPs. GTEx data can be also useful in performing tissue specific enrichment analysis, whereby the most affected tissue for a given list of disease-associated SNPs can be identified by simply counting the number of eQTL instances identified per tissue.

4.9. Imputation of Gene Expression Profiles

PrediXcan [74] is a recently developed method aimed at saving the costs of expensive transcripome sequencing. It can impute transcriptome-wide expression profiles for Caucasian samples based on reference transcriptome datasets from large studies such as Genotype-Tissue Expression (GTEx) [73], Genetic European Variation in Health and Disease (GEUVADIS) [75], and Depression Genes and Networks (DGN) [76] where both expression and SNP data are available. In the current release of PredictDB (a database associated with the tool), the authors of PrediXscan only included genes that had a false discovery rate ≤5% (for example 11,553 autosomal genes in whole blood) based on the elastic net models used to generate the SNP weights. The quality of the transcriptome imputation depends highly on the number of SNPs included in the gene expression prediction model: the more SNPs, the better the imputation quality. Prediction models are available for many tissues to allow evaluation of eQTL associations with diseases before undertaking actual gene expression measurements.

4.10. Chromatin Conformation Capture and Linking GWAS SNPs to Target Genes

Chromatin conformation capture (3C) and variants of this approach (4C, 5C, Hi-C, and ChIA-PET) probe long-range interactions by utilizing formaldehyde-directed cross-linking of genomic modules that are close in physical space [77]. A recent paper [78] provided high-resolution analysis of interactions involving almost all annotated promoters (Fantom5 [43], ENCODE [42]) in 17 human primary blood cell types. Links were identified between disease-associated variants with their putative target genes by integrating chromatin-interaction with population genetics data. The data can be accessed via an online portal: www.chicp.org.

4.11. Pathway, Network, and Enrichment Analysis

After identification of target genes regulated by the disease variants, pathway and enrichment analysis follow to provide insights into potential biological mechanisms the genes might be involved in. These analyses can be performed on a cell-type specific basis or on combined tissues. There are several tools available to perform these analyses. DAVID v6.8 [79] is a very popular enrichment analysis tool that has been extensively applied to discover pathways for a given set of genes. Typically, annotation terms meeting Benjamini-Hochberg $p < 0.05$ (adjusted for the number of terms) are considered significant. Gene-set collections available in the Molecular Signature Databases (MSigDB) [28] of the Broad Institute can be also tested for enrichment using their web interface (http://software.broadinstitute.org/gsea/

msigdb/annotate.jsp). Again an FDR adjusted *p*-value is computed for all terms for measuring significance. Both DAVID and MSigDB provide options to test Gene Ontology (GO) terminologies in addition to canonical pathways; GO is useful in identifying significant biological, cellular, and molecular functions associated with the gene lists. If the direction of eQTL associations with respect to the risk allele is known, GO analysis can be separately applied to up- and down-regulated genes so as to identify functions that are either magnified or diminished in relation to susceptibility.

Other popular enrichment tools such as GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD) [80] and Genomic Regulatory Elements and Gwas Overlap algorithm (GREGOR) [81] also provide cross-verification for significantly enriched terms. Ingenuity Pathway Analysis (IPA) is a commercial software for performing enrichment analysis with special options to perform toxicity analysis and identify drugs associated with candidate genes. Network analysis allows connecting target genes with related molecules and helping identify associated functions. This type of analysis can be performed with the help of the Cytoscape-GeneMania [82] module that allows for the connection of genes based on co-expression patterns. An MCODE (molecular complex detection) plugin within Cytoscape [83] allows for identifying clusters of sub-networks within highly connected networks, functions of which can be separately defined within Cytoscape itself. Such types of analysis will help to shed insights into the biological functions underlying susceptibility.

5. Conclusions

Type 1 Diabetes is a complex genetic disease with reports of over 60 loci that increase a person's risk of developing the disease. Most of the risk loci do not mediate disease susceptibility via missense changes to coding regions [19]. We investigated the role of risk SNPs in affecting gene expression in various immune related cells, combining the results of expression studies conducted in B-cells, monocytes, dendritic cells, EBV-transformed B-cells, and CD4+ and CD8+ T-cells. Our results revealed that there were 90 *cis*-regulated genes and ten *trans*-regulated genes that were evidenced in multiple cells (or studies). In addition, there were 34 and 22 highly significant cell-specific *cis*- and *trans*-regulated genes, respectively. We provided a methodology for the identification of eQTLs associated with T1D SNPs in different immune cells and suggested alternative methods to improve and overcome the statistical limitations (power) of the techniques used. We have also discussed several advancements in the area of linking GWAS SNPs to functional variants using next generation sequencing techniques. Future study designs for testing T1D GWAS variants will need to incorporate these advancements and find ways to integrate data from various sources as discussed in the paper to further improve our understanding of the disease susceptibility.

Author Contributions: Ramesh Ram analyzed the data; Ramesh Ram and Grant Morahan wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1.　Barrett, J.C.; Clayton, D.G.; Concannon, P.; Akolkar, B.; Cooper, J.D.; Erlich, H.A.; Julier, C.; Morahan, G.; Nerup, J.; Nierras, C.; et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **2009**, *41*, 703–707. [CrossRef] [PubMed]

2.　Onengut-Gumuscu, S.; Chen, W.M.; Burren, O.; Cooper, N.J.; Quinlan, A.R.; Mychaleckyj, J.C.; Farber, E.; Bonnie, J.K.; Szpak, M.; Schofield, E.; et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **2015**, *47*, 381–386. [CrossRef] [PubMed]

3. Bradfield, J.P.; Qu, H.Q.; Wang, K.; Zhang, H.; Sleiman, P.M.; Kim, C.E.; Mentch, F.D.; Qiu, H.; Glessner, J.T.; Thomas, K.A.; et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **2011**, *7*, e1002293. [CrossRef] [PubMed]

4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **2007**, *447*, 661–678.

5. Morahan, G.; Mehta, M.; James, I.; Chen, W.M.; Akolkar, B.; Erlich, H.A.; Hilner, J.E.; Julier, C.; Nerup, J.; Nierras, C.; et al. Tests for Genetic Interactions in Type 1 Diabetes: Linkage and Stratification Analyses of 4422 Affected Sib-Pairs. *Diabetes* **2011**, *60*, 1030–1040. [CrossRef] [PubMed]

6. Cooper, J.D.; Smyth, D.J.; Smiles, A.M.; Plagnol, V.; Walker, N.M.; Allen, J.E.; Downes, K.; Barrett, J.C.; Healy, B.C.; Mychaleckyj, J.C.; et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **2008**, *40*, 1399–1401. [CrossRef] [PubMed]

7. Todd, J.A.; Walker, N.M.; Cooper, J.D.; Smyth, D.J.; Downes, K.; Plagnol, V.; Bailey, R.; Nejentsev, S.; Field, S.F.; Payne, F.; et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **2007**, *39*, 857–864. [CrossRef] [PubMed]

8. Smyth, D.J.; Plagnol, V.; Walker, N.M.; Cooper, J.D.; Downes, K.; Yang, J.H.; Howson, J.M.; Stevens, H.; McManus, R.; Wijmenga, C.; et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **2008**, *359*, 2767–2777. [CrossRef] [PubMed]

9. Yang, J.H.; Downes, K.; Howson, J.M.; Nutland, S.; Stevens, H.E.; Walker, N.M.; Todd, J.A. Evidence of association with type 1 diabetes in the SLC11A1 gene region. *BMC Med. Genet.* **2011**, *12*, 59. [CrossRef] [PubMed]

10. Swafford, A.D.; Howson, J.M.; Davison, L.J.; Wallace, C.; Smyth, D.J.; Schuilenburg, H.; Maisuria-Armer, M.; Mistry, T.; Lenardo, M.J.; Todd, J.A. An allele of IKZF1 (Ikaros) conferring susceptibility to childhood acute lymphoblastic leukemia protects against type 1 diabetes. *Diabetes* **2011**, *60*, 1041–1044. [CrossRef] [PubMed]

11. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. [CrossRef] [PubMed]

12. Alexander, D.H.; Novembre, J.; Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **2009**, *19*, 1655–1664. [CrossRef] [PubMed]

13. Kang, H.M.; Zaitlen, N.A.; Wade, C.M.; Kirby, A.; Heckerman, D.; Daly, M.J.; Eskin, E. Efficient control of population structure in model organism association mapping. *Genetics* **2008**, *178*, 1709–1723. [CrossRef] [PubMed]

14. Singal, D.P.; Blajchman, M.A. Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes* **1973**, *22*, 429–432. [CrossRef] [PubMed]

15. Vang, T.; Miletic, A.V.; Bottini, N.; Mustelin, T. Protein tyrosine phosphatase PTPN22 in human autoimmunity. *Autoimmunity* **2007**, *40*, 453–461. [CrossRef] [PubMed]

16. Pugliese, A.; Zeller, M.; Fernandez, A., Jr.; Zalcberg, L.J.; Bartlett, R.J.; Ricordi, C.; Pietropaolo, M.; Eisenbarth, G.S.; Bennett, S.T.; Patel, D.D. The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nat. Genet.* **1997**, *15*, 293–297. [CrossRef] [PubMed]

17. Nistico, L.; Buzzetti, R.; Pritchard, L.E.; Van der Auwera, B.; Giovannini, C.; Bosi, E.; Larrad, M.T.; Rios, M.S.; Chow, C.C.; Cockram, C.S.; et al. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.* **1996**, *5*, 1075–1080. [CrossRef] [PubMed]

18. Lowe, C.E.; Cooper, J.D.; Brusko, T.; Walker, N.M.; Smyth, D.J.; Bailey, R.; Bourget, K.; Plagnol, V.; Field, S.; Atkinson, M.; et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.* **2007**, *39*, 1074–1082. [CrossRef] [PubMed]

19. Ram, R.; Mehta, M.; Nguyen, Q.T.; Larma, I.; Boehm, B.O.; Pociot, F.; Concannon, P.; Morahan, G. Systematic Evaluation of Genes and Genetic Variants Associated with Type 1 Diabetes Susceptibility. *J. Immunol.* **2016**, *196*, 3043–3053. [CrossRef] [PubMed]

20. Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **2001**, *29*, 1165–1188.

21. Fairfax, B.P.; Humburg, P.; Makino, S.; Naranbhai, V.; Wong, D.; Lau, E.; Jostins, L.; Plant, K.; Andrews, R.; McGee, C.; et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **2014**, *343*, 1246949. [CrossRef] [PubMed]

22. Fairfax, B.P.; Makino, S.; Radhakrishnan, J.; Plant, K.; Leslie, S.; Dilthey, A.; Ellis, P.; Langford, C.; Vannberg, F.O.; Knight, J.C. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **2012**, *44*, 502–510. [CrossRef] [PubMed]

23. Lee, M.N.; Ye, C.; Villani, A.C.; Raj, T.; Li, W.; Eisenhaure, T.M.; Imboywa, S.H.; Chipendo, P.I.; Ran, F.A.; Slowikowski, K.; et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **2014**, *343*, 1246980. [CrossRef] [PubMed]

24. Peters, J.E.; Lyons, P.A.; Lee, J.C.; Richard, A.C.; Fortune, M.D.; Newcombe, P.J.; Richardson, S.; Smith, K.G. Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. *PLoS Genet.* **2016**, *12*, e1005908. [CrossRef] [PubMed]

25. Westra, H.J.; Arends, D.; Esko, T.; Peters, M.J.; Schurmann, C.; Schramm, K.; Kettunen, J.; Yaghootkar, H.; Fairfax, B.P.; Tserel, L.; et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **2014**, *11*, e1005223.

26. Westra, H.J.; Peters, M.J.; Esko, T.; Yaghootkar, H.; Schurmann, C.; Kettunen, J.; Christiansen, M.W.; Fairfax, B.P.; Schramm, K.; Powell, J.E.; et al. Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **2013**, *45*, 1238–1243. [CrossRef] [PubMed]

27. Machiela, M.J.; Chanock, S.J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **2015**, *31*, 3555–3557. [CrossRef] [PubMed]

28. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]

29. Lessard, C.J.; Li, H.; Adrianto, I.; Ice, J.A.; Rasmussen, A.; Grundahl, K.M.; Kelly, J.A.; Dozmorov, M.G.; Miceli-Richard, C.; Bowman, S.; et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nat. Genet.* **2013**, *45*, 1284–1292. [CrossRef] [PubMed]

30. Corradin, O.; Saiakhova, A.; Akhtar-Zaidi, B.; Myeroff, L.; Willis, J.; Cowper-Sal, L.R.; Lupien, M.; Markowitz, S.; Scacheri, P.C. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **2014**, *24*, 1–13. [CrossRef] [PubMed]

31. Farh, K.K.; Marson, A.; Zhu, J.; Kleinewietfeld, M.; Housley, W.J.; Beki, S.; Shoresh, N.; Whitton, H.; Ryan, R.J.H.; Shishkin, A.A.; et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **2015**, *518*, 337–343. [CrossRef] [PubMed]

32. Walsh, A.M.; Whitaker, J.W.; Huang, C.C.; Cherkas, Y.; Lamberth, S.L.; Brodmerkel, C.; Curran, M.E.; Dobrin, R. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. *Genome Biol.* **2016**, *17*, 79. [CrossRef] [PubMed]

33. Tewhey, R.; Kotliar, D.; Park, D.S.; Liu, B.; Winnicki, S.; Reilly, S.K.; Andersen, K.G.; Mikkelsen, T.S.; Lander, E.S.; Schaffner, S.F.; et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **2016**, *165*, 1519–1529. [CrossRef] [PubMed]

34. Hakonarson, H.; Qu, H.Q.; Bradfield, J.P.; Marchand, L.; Kim, C.E.; Glessner, J.T.; Grabs, R.; Casalunovo, T.; Taback, S.P.; Frackelton, E.C.; et al. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **2008**, *57*, 1143–1146. [CrossRef] [PubMed]

35. Brown, C.T.; Davis-Richardson, A.G.; Giongo, A.; Gano, K.A.; Crabb, D.B.; Mukherjee, N.; Casella, G.; Drew, J.C.; Ilonen, J.; Knip, M.; et al. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS ONE* **2011**, *6*, e25792. [CrossRef] [PubMed]

36. Guy, J.; Ogden, L.; Wadwa, R.P.; Hamman, R.F.; Mayer-Davis, E.J.; Liese, A.D.; D'Agostino, R., Jr.; Marcovina, S.; Dabelea, D. Lipid and lipoprotein profiles in youth with and without type 1 diabetes: The SEARCH for Diabetes in Youth case-control study. *Diabetes Care* **2009**, *32*, 416–420. [CrossRef] [PubMed]

37. Catalán, V.; Gómez-Ambrosi, J.; Pastor, C.; Rotellar, F.; Silva, C.; Rodríguez, A.; Gil, M.J.; Cienfuegos, J.A.; Salvador, J.; Vendrell, J.; et al. Influence of morbid obesity and insulin resistance on gene expression levels of AQP7 in visceral adipose tissue and AQP9 in liver. *Obes. Surg.* **2008**, *18*, 695–701. [CrossRef] [PubMed]

38. Wing, S.S. The UPS in diabetes and obesity. *BMC Biochem.* **2008**, *9*, S6. [CrossRef] [PubMed]

39. Ram, R.; Morahan, G. Using Systems Genetics to Understanding the Etiology of Complex Disease. *Methods Mol. Biol.* **2017**, *1488*, 597–606. [PubMed]

40. Kumar, P.; Henikoff, S.; Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **2009**, *4*, 1073–1081. [CrossRef] [PubMed]

41. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* **2010**, *7*, 248–249. [CrossRef] [PubMed]

42. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.

43. FANTOM Consortium; the RIKEN PMI; CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **2014**, *507*, 462–470.

44. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef] [PubMed]

45. Quang, D.; Chen, Y.; Xie, X. DANN: A deep-learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **2015**, *31*, 761–763. [CrossRef] [PubMed]

46. Shihab, H.A.; Rogers, M.F.; Gough, J.; Mort, M.; Cooper, D.N.; Day, I.N.; Gaunt, T.R.; Campbell, C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **2015**, *31*, 1536–1543. [CrossRef] [PubMed]

47. Huang, Y.H.; Gulko, B.; Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **2017**, *49*, 618–624. [CrossRef] [PubMed]

48. Bendl, J.; Stourac, J.; Salanda, O.; Pavelka, A.; Wieben, E.D.; Zendulka, J.; Brezovsky, J.; Damborsky, J. PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **2014**, *10*, e1003440. [CrossRef] [PubMed]

49. Bendl, J.; Musil, M.; Štouraeč, J.; Zendulka, J.; Damborský, J.; Brezovský, J. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Comput. Biol.* **2016**, *12*, e1004962. [CrossRef] [PubMed]

50. Du, P.; Zhang, X.; Huang, C.C.; Jafari, N.; Kibbe, W.A.; Hou, L.; Lin, S.M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform.* **2010**, *11*, 587. [CrossRef] [PubMed]

51. Arloth, J.; Bader, D.M.; Röh, S.; Altmann, A. Re-Annotator: Annotation pipeline for microarray probe sequences. *PLoS ONE* **2015**, *10*, e0139516. [CrossRef] [PubMed]

52. Oshlack, A.; Robinson, M.D.; Young, M.D. From RNA-seq reads to differential expression results. *Genome Biol.* **2010**, *11*, 220. [CrossRef] [PubMed]

53. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef] [PubMed]

54. Scherer, A. Batch Effects and Noise in Microarray Experiments: Sources and Solutions. In *Wiley Series Probability Statistics*; Wiley-Blackwell: Oxford, UK, 2009; p. 272.

55. Wu, Z.; Aryee, M.J. Subset quantile normalization using negative control features. *J. Comput. Biol.* **2010**, *17*, 1385–1395. [CrossRef]

56. Risso, D.; Ngai, J.; Speed, T.P.; Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **2014**, *32*, 896–902. [CrossRef] [PubMed]

57. Leek, J.T.; Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **2007**, *3*, e161. [CrossRef] [PubMed]

58. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Fertig, E.J.; Jaffe, A.E.; Storey, J.D. sva: Surrogate Variable Analysis. In *R Package*; Version 3.25.0; R Foundation for Statistical Computing: Vienna, Austria, 2017.

59. Westra, H.J.; Jansen, R.C.; Fehrmann, R.S.; te Meerman, G.J.; van Heel, D.; Wijmenga, C.; Franke, L. MixupMapper: Correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **2011**, *27*, 2104–2111. [CrossRef] [PubMed]

60. Vazquez, A.I.; Bates, D.M.; Rosa, G.J.M.; Gianola, D.; Weigel, K.A. Technical note: An R package for fitting generalized linear mixed models in animal breeding. *J. Anim. Sci.* **2010**, *88*, 497–504. [CrossRef] [PubMed]

61. Stegle, O.; Parts, L.; Piipari, M.; Winn, J.; Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **2012**, *7*, 500–507. [CrossRef] [PubMed]

62. Ellis, S.E.; Gupta, S.; Ashar, F.N.; Bader, J.S.; West, A.B. Arking DE RNA-Seq optimization with eQTL gold standards. *BMC Genom.* **2013**, *14*, 892. [CrossRef] [PubMed]

63. Tracy, C.A.; Widom, H. Level-Spacing Distributions and the Airy Kernel. *Commun. Math. Phys.* **1994**, *159*, 151–174. [CrossRef]

64. Patterson, N.; Price, A.L.; Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2006**, *2*, e190. [CrossRef] [PubMed]

65. Shabalin, A.A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **2012**, *28*, 1353–1358. [CrossRef] [PubMed]

66. Dabney, A.; Storey, J.D.; Warnes, G.R. qvalue: Q-value estimation for false discovery rate control. In *R Package*; Version 1.20.0; R Foundation for Statistical Computing: Vienna, Austria, 2009.

67. He, X.; Fuller, C.K.; Song, Y.; Meng, Q.; Zhang, B.; Yang, X.; Li, H. Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **2013**, *92*, 667–680. [CrossRef] [PubMed]

68. Giambartolomei, C.; Vukcevic, D.; Schadt, E.E.; Franke, L.; Hingorani, A.D.; Wallace, C.; Plagnol, V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **2014**, *10*, e1004383. [CrossRef] [PubMed]

69. Pickrell, J.K.; Berisa, T.; Liu, J.Z.; Ségurel, L.; Tung, J.Y.; Hinds, D.A. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **2016**, *48*, 709–717. [CrossRef] [PubMed]

70. Zhu, Z.; Zhang, F.; Hu, H.; Bakshi, A.; Robinson, M.R.; Powell, J.E.; Montgomery, G.W.; Goddard, M.E.; Wray, N.R.; Visscher, P.M.; et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **2016**, *48*, 481–487. [CrossRef] [PubMed]

71. Petretto, E.; Bottolo, L.; Langley, S.R.; Heinig, M.; McDermott-Roe, C. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* **2010**, *6*, e1000737. [CrossRef] [PubMed]

72. Flutre, T.; Wen, X.; Pritchard, J.; Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **2013**, *9*, e1003486. [CrossRef] [PubMed]

73. The GTEx Consortium; Ardlie, K.G.; Deluca, D.S.; Segrè, A.V.; Sullivan, T.J.; Young, T.R.; Gelfand, E.T.; Trowbridge, C.A.; Maller, J.B.; Tukiainen, T. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648660.

74. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; GTEx Consortium; Nicolae, D.L.; et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **2015**, *47*, 1091–1098. [CrossRef] [PubMed]

75. Lappalainen, T.; Sammeth, M.; Friedländer, M.R.; AC't Hoen, P.A.; Monlong, J.; Rivas, M.A.; Gonzàlez-Porta, M.; Kurbatova, N.; Griebel, T.; Ferreira, G.P.; et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **2013**, *501*, 506–511. [CrossRef] [PubMed]

76. Battle, A.; Mostafavi, S.; Zhu, X.; Potash, J.B.; Weissman, M.M.; McCormick, C.; Haudenschild, C.D.; Beckman, K.B.; Shi, J.; Mei, R.; et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **2014**, *24*, 14–24. [CrossRef] [PubMed]

77. De Wit, E.; de Laat, W. A decade of 3C technologies: Insights into nuclear organization. *Genes Dev.* **2012**, *26*, 11–24. [CrossRef] [PubMed]

78. Schofield, E.C.; Carver, T.; Achuthan, P.; Freire-Pritchett, P.; Spivakov, M.; Todd, J.A.; Burren, O.S. CHiCP: A web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics* **2016**, *32*, 2511–2513. [CrossRef] [PubMed]

79. Huang, W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13. [CrossRef] [PubMed]

80. Iotchkova, V.; Ritchie, G.R.S.; Geihs, M.; Morganella, S.; Min, J.L.; Walter, K.; Timpson, N.J.; Dunham, I.; Birney, E.; Soranzo, N.; et al. GARFIELD—GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *Biorxiv* **2016**. [CrossRef]

81. Schmidt, E.M.; Zhang, J.; Zhou, W.; Chen, J.; Mohlke, K.L.; Chen, Y.E. Gregor: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **2015**, *31*, 2601–2606. [CrossRef] [PubMed]

82. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*, W214–W220. [CrossRef] [PubMed]

83. Bader, G.D.; Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef]