genes

MDPI

Review

Functional Genetics to Understand the Etiology of Autoimmunity

Hiroaki Hatano and Kazuyoshi Ishigaki *

Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan
* Correspondence: kazuyoshi.ishigaki@riken.jp

Abstract: Common variants strongly influence the risk of human autoimmunity. Two categories of variants contribute substantially to the risk: (i) coding variants of *HLA* genes and (ii) non-coding variants at the non-*HLA* loci. We recently developed a novel analytic pipeline of T cell receptor (TCR) repertoire to understand how *HLA* coding variants influence the risk. We identified that the risk variants increase the frequency of auto-reactive T cells. In addition, to understand how non-coding variants contribute to the risk, the researchers conducted integrative analyses using expression quantitative trait loci (eQTL) and splicing quantitative trait loci (sQTL) and demonstrated that the risk non-coding variants dysregulate specific genes' expression and splicing. These studies provided novel insight into the immunological consequences of two major genetic risks, and we will introduce these research achievements in detail in this review.

Keywords: V2F; immunogenetics; TCR; eQTL; sQTL

check for updates

Citation: Hatano, H.; Ishigaki, K. Functional Genetics to Understand the Etiology of Autoimmunity. *Genes* **2023**, *14*, 572. https://doi.org/10.3390/genes14030572

Academic Editors: Katsushi Tokunaga and Yuki Hitomi

Received: 19 January 2023
Revised: 16 February 2023
Accepted: 23 February 2023
Published: 24 February 2023

1. Introduction

The genome-wide association study (GWAS) aims to detect associations between germline genetic variants and human phenotypes. The GWAS has no reverse causation: the phenotype cannot affect the variant. Therefore, the GWAS is one of a few studies that can assess the causal mechanism of human diseases. Over the past ten years, large-scale GWASs for autoimmune diseases have successfully detected hundreds of risk variants, exemplified by studies for rheumatoid arthritis (RA) [1,2] and systemic lupus erythematosus (SLE) [3–5]. However, the primary GWAS outputs are just a group of statistics of genome-wide variants. To extract biological information from GWAS results, we first need to extensively conduct genetic studies that connect variants to function (V2F). We then can infer the causal mechanisms of human autoimmunity by integrating GWAS and V2F study results (Figure 1). In this review, we provide various V2F studies and show how such study contributed to a better understanding of human autoimmunity etiologies.

1.1. Genetic Risk by HLA Coding Variants

The most outstanding characteristic of the GWAS for autoimmune diseases is the striking associations at the major histocompatibility complex (MHC) region, reflecting coding variants of *HLA* genes (Figure 2). Previous studies reported the *HLA* genes' risk and protective amino acid polymorphisms [6–10]. For example, the risk of RA is strongly associated with *HLA-DRB1**0401 in European ancestries and *HLA-DRB1**0405 in East Asian ancestries; and *HLA-DRB1**1501 has been associated with multiple sclerosis (MS). Using sophisticated analytical strategies, researchers fine-mapped the MHC associations and demonstrated that a few amino acid positions of *HLA* genes account for most associations at the MHC locus. Raychaudhuri et al. reported amino acid polymorphisms at position 13 (or position 11 in strong linkage disequilibrium (LD) with position 13), 71, 74 of *HLA-DRB1*, position 9 of *HLA-B*, and position 9 of *HLA-DPB1*, which almost completely explain the MHC association to RA [11]. Intriguingly, all positions are located in peptide-binding

grooves of *HLA* genes. Hu et al. conducted a similar analysis, and the top hit was found at position 57 of *HLA-DQB1*, followed by positions 13 and 71 of *HLA-DRB1* [12].

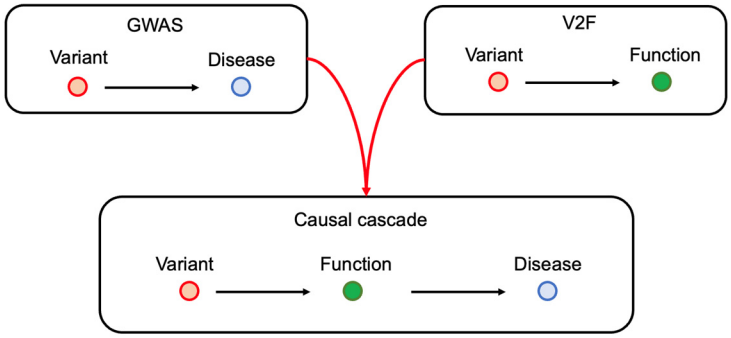

Figure 1. The V2F study illuminates the causal cascade of autoimmunity. The GWAS connects variants and diseases. V2F studies link variants and function. The function can be any immune-related phenotypes. The most studied and feasible phenotype is gene expression levels in immune cells. By combining GWAS and V2F outputs, we can draw the causal cascade.

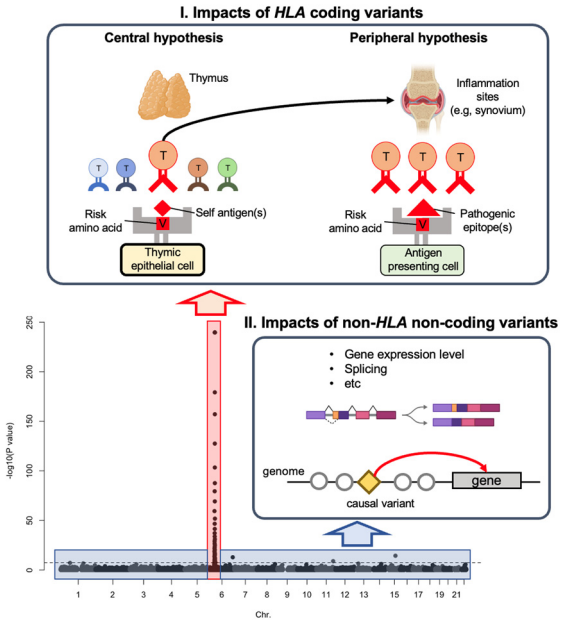

Figure 2. Two genetic risk categories of autoimmunity. In this review article, we introduced two categories of genetic risk of autoimmunity. One of our recent studies suggested that *HLA* coding variants influence thymic selection, modify TCR repertoire, and increase the auto-reactive immune response (the "central hypothesis") [13]. Other researchers reported that *HLA* coding variants influence the binding affinity of pathogenic epitopes and enhance immune reactions against them (the "peripheral hypothesis"). On the other hand, non-*HLA* non-coding variants are enriched in the regulatory region and probably influence gene expression and splicing. We used *p*-values in our recent multi-ancestry of RA-GWAS for the bottom Manhattan plot [1]. We used images of the thymus, joint, and splice isoforms from BioRender (https://biorender.com/, accessed on 2 January 2023).

The canonical function of *HLA* genes is to present antigenic peptides to T cell receptors (TCR). Since the genetic risks are accumulated at the amino acid positions within peptide-binding grooves of *HLA* genes, we need to design V2F studies in the context of three players: *HLA*, antigenic peptides, and TCR. The etiological importance of this context is also supported by the fact that risk variants have been identified around genes encoding TCR signaling molecules. A good example is a missense variant of *PTPN22* (rs2476601; W620R), which is the top hit in RA-GWAS outside of the MHC region and shows pleiotropic

associations for multiple autoimmune diseases [1,3,14]. *PTPN22* plays a key role in TCR signaling and inhibits T cell activation by dephosphorylating substrates involved in TCR signaling. Additionally, the genes implicated in RA GWAS were enriched for the TCR signaling pathway [15].

TCR is an "eye" of T cells, distinguishing self and foreign antigens. TCR can recognize antigenic peptides only when the peptides are presented on the HLA molecules. T cell initiates antigen-specific immune reactions involving multiple immune cell populations. Dysregulation of antigen-specific immunity is a hallmark of autoimmune diseases because we observe specific autoantibodies in the serum of autoimmune disease patients, e.g., anti-citrullinated peptide antibodies (ACPA) for RA and anti-double strand DNA antibodies for SLE [16,17]. In addition to the genetic evidence, this immunological evidence also supports the critical roles of *HLA*, antigenic peptides, and TCR in the pathology of autoimmunity. Since TCR signaling is a crucial factor for T cell development, activation, and differentiation, the antigenic peptide-HLA complex continuously influences T cells throughout the entire life cycle of T cells [18]. Therefore, V2F studies need to aim at different T cell developmental phases.

Historically, researchers have been conducting V2F studies mainly focusing on *HLA* and antigenic peptides by testing each *HLA* allele's binding affinity to the pathogenic epitopes. The antigen-binding groove of the *HLA* class II molecule possesses several binding pockets accommodating the side chains of the antigenic peptides; the pockets with strong interaction are P1, P4, P6, P7, and P9 [19]. The idea is that when the pathogenic epitopes are more frequently presented to T cells in the peripheral tissues (e.g., inflammation sites and regional lymph nodes), the risk of developing autoimmunity should increase, which was introduced as the "peripheral hypothesis" of the *HLA* genetic risk in our recent article [13] (Figure 2).

In RA for instance, the high binding affinity of citrullinated epitopes, the most established pathogenic epitopes in RA, is found for the *HLA-DRB1* proteins encoded by the risk alleles. Scally et al. reported the structural basis of how the risk *HLA-DB1* alleles enhance an autoimmune reaction to the citrullinated epitopes, focusing on two *HLA* risk alleles (*HLA-DRB1*04:01* and *04:04*) with an electropositive P4 antigen-binding pocket and a protective allele (*HLA-DRB1**04:02) with an electronegative P4 pocket [20]. They demonstrated that *HLA-DRB1**04:01/04 with the positive P4 pocket favors citrulline (no net charge) but disfavors arginine (positively charged), whereas *HLA-DRB1**04:01/04 with the negative P4 pocket favors arginine. They also provided in-depth mass spectrometry analyses of the peptide repertoire bound to each *HLA-DR* allele and identified substantially different binding motifs, especially at the P4 pocket, where arginine was depleted in the risk alleles while tolerated in the protective allele. Hill et al. reported *HLA-DRB1**0401 transgenic mice immunized with cartilage proteoglycan aggrecan epitopes with arginine at P4 and those with citrulline at P4 [21]. They demonstrated that the arginine to citrulline conversion at P4 significantly increases peptide-HLA affinity and leads to activating $CD4^+$ T cells in their transgenic mice.

Similarly, other studies also suggested the importance of the high binding affinity of the HLA risk alleles to the pathogenic epitopes in other autoimmune diseases such as type 1 diabetes (T1D) [22] and celiac disease [23].

Notably, the previous studies investigating the "peripheral hypothesis" did not consider how TCR repertoire is constructed before T cells encounter the molecular complex of *HLA* and pathogenic epitopes. T cells differentiate and mature in the thymus, where TCR is generated by random recombination. Thymic immature T cells randomly select and combine one TCR component gene from multiple candidates for each of V, D (only for β chain), and J gene while randomly adding or deleting several nucleotides at the junctional region of these component genes. This junctional region is called complementarity determining region 3 (CDR3). Due to these random processes, each T cell has a unique CDR3 sequence, which is a "fingerprint" of the T cell, and each human has a strikingly diverse repertoire of CDR3. Since CDR3 directly contacts with antigenic epitopes presented

on the HLA molecule, the various CDR3 sequence patterns enable the immune system to recognize a wide range of antigens.

Reasonably, these random processes generate many non-functional TCRs that cannot interact with self-HLA molecules. Since TCR is an essential molecule for T cells, the thymus needs to select cells with functional TCR, called positive selection. Naturally, many of the T cells selected in this way are autoreactive, at least to some extent. To prevent autoimmunity, the thymus must eliminate T cells with TCR showing strong reactivity to autoantigens, called negative selection. These thymic selections drastically alter TCR repertoire, and most importantly, the peptide-HLA molecular complex has a critical role in these processes. Therefore, HLA risk alleles may affect the thymic selection and modify the TCR repertoire enhancing the autoreactivity, which is the "central hypothesis" of the HLA genetic risk [13] (Figure 2).

Motivated by this idea, we recently conducted the first genetic study testing associations between *HLA* alleles and TCR-CDR3 amino acid compositions, named cdr3-QTL [13]. In our research question, the TCR-CDR3 data (the response variable) are sequence data, and the *HLA* genotypes (the explanatory variable) are multi-allelic. Hence, the classical linear models were not feasible in this study. Therefore, we developed a novel analytical pipeline. First, we transformed CDR3 sequence data into a group of quantitative traits: a 20-dimensional vector with each component representing the usage frequency of each amino acid at a specific CDR3 position. We next transformed multi-allelic *HLA* genotype data (for example, m alleles) into a multi-dimensional vector, with each component representing the count of each *HLA* allele at a specific HLA position. We then applied a multivariate multiple linear regression model (MMLM) to detect associations between the CDR3 and *HLA* vectors, assessing the significance with the multivariate analysis of variance (MANOVA) test. Intuitively, this MMLM model estimates the correlation between CDR3 amino acid composition at a CDR3 position and all *HLA* alleles at an HLA position.

We applied our cdr3-QTL pipeline to publicly available TCR repertoire data of whole T cells from 628 healthy donors [24]. We demonstrated the strongest association at the amino acid position 13 of *HLA-DRB1*, the position with the strongest associations for the RA risk, and the 2nd strongest associations for T1D risk. These cdr3-QTL signals were successfully replicated in naïve $CD4^+$ T cell TCR repertoire (number of donors = 169), and the signals were attenuated when we included clonally expanded T cell fraction. Therefore, the cdr3-QTL signals probably reflect thymic T cell selection rather than T cell selection during peripheral memory formation. Since the exact *HLA* position showed the most robust associations both for autoimmunity and CDR3 amino acid compositions, the *HLA* genetic risk is probably mediated by the thymic TCR-CDR3 selection dysregulated by *HLA* risk alleles.

In addition, we further conducted in-depth analyses to identify specific CDR3 patterns associated with *HLA* risks. We found several disease-specific patterns. RA and T1D *HLA* risk alleles increase acidic amino acid and decrease basic amino acid at the center of CDR3, linking the CDR3 negative charge and the genetic risk. In contrast, celiac disease *HLA* risk alleles increase hydrophobic amino acid at the center of CDR3. Previous studies showed that both amino acid charge and hydrophobicity of CDR3 influence antigen specificity [25]. Therefore, we hypothesized that accumulating these CDR3 amino acid patterns increases the T cell reactivity to pathogenic epitopes. We confirmed the possibility of this hypothesis by analyzing TCR sequence datasets derived from T cell subsets showing reactivity to several pathogenic epitopes: gluten-specific TCRs from celiac disease patients and citrullinated peptide-specific TCRs from RA patients. In summary, our study demonstrated striking associations between the *HLA* alleles and TCR-CDR3 amino acid compositions, providing novel genetic evidence supporting the "central hypothesis".

1.2. Genetic Risk by Non-HLA Non-Coding Variants

In contrast to the *HLA* genes with a limited number of high-impact risk variants, non-*HLA* genes have numerous low-impact risk variants [26]. Specifically, the risk variants of

non-*HLA* genes are enriched in the regulatory regions of relevant immune cell subsets. For example, the RA risk variants are enriched in the active regulatory regions of $CD4^+$ T cell lineages, such as regulatory T cells [27,28]. Therefore, researchers have been conducting V2F studies to elucidate how variants affect the gene regulatory machinery in a cell type-specific manner (Figure 2).

The most straightforward scenario of the risk variant etiology is that they affect gene expression, i.e., expression of quantitative trait loci (eQTL). Therefore, researchers have conducted large-scale eQTL studies of immune cell subsets trying to illuminate the risk variant's mechanisms, e.g., for which gene(s) and in which cell subset(s) the risk variants exert gene regulatory functions. The first wave of such research effort includes The Immune Variation (ImmVar) project, aiming to map the extent of variation in immune function in healthy human subjects [29]. Among multiple accompanying studies, Raj et al. conducted an eQTL study using purified $CD4^+$ T cells and monocytes of 461 healthy donors, linking RA risk variants with T cell-specific eQTLs and Alzheimer's disease risk variants with monocyte-specific eQTLs [30].

As the eQTL study platform matured, researchers started aiming to obtain a landscape of immune cell-specific eQTL across various immune cell subsets. We conducted an eQTL study using six immune cell subsets from 105 healthy donors [31]. Schmiedel et al. used 13 immune cell subsets isolated from 106 healthy donors [32] (DICE project). These research efforts were followed by our latest study that used 28 distinct immune cell subsets from 416 donors [33] (ImmuNexUT project). This study found several cell type-specific eQTLs colocalized with risk variants. For example, we observed the eQTL effect on ARHGAP31 only in plasmablasts, and the eQTL signal showed strong colocalization with a GWAS signal of SLE.

The intriguing characteristic of the ImmuNexUT project is that 337 among 416 donors were patients diagnosed with ten categories of immune-mediated diseases (IMD). Therefore, we were able to investigate the context-dependent eQTLs, e.g., how immune alterations in IMD patients affect eQTL effect size magnitude. We searched for genes whose expression level interacts with the eQTL effect; we called such genes "proxy genes" (pGenes). We successfully identified 37,875 significant pGene-eQTL interactions (FDR < 0.05). Furthermore, we found that pGenes were significantly overlapped with IFN signature genes, suggesting IFN has a pivotal role in the gene regulatory machinery in IMD patients. In addition, we found the enrichment of context-dependent eQTLs in GWAS top signals compared with all immune cell eQTLs.

Since the cell type specificity of eQTL signals is the key factor to elucidate the genetic etiology of complex traits, one of the most promising directions of eQTL research is arguably the single-cell level analysis as in other research fields. Monique et al. reported the first single-cell eQTL study using peripheral blood mononuclear cells (PBMC). Although the study scale is relatively limited (~25,000 PBMCs from 45 donors), they successfully showed the feasibility of a single-cell eQTL study, which produces very sparse expression data with many dropouts. They used the "pseudo-bulk approach" to mitigate this issue. They first conducted clustering to identify cell groups with similar expression profiles and created one expression data by integrating all cells within the same group (the data structure at this stage is essentially identical to that of bulk eQTL studies) and finally tested associating between genotypes and pseudo-bulk expression data. The pseudo-bulk approach is efficient and flexible. For example, this approach enables us to deploy previously established analytical pipelines for bulk eQTL, e.g., normalization, association tests, and the detection of cell type specificity of eQTL signals.

Using the pseudo-bulk approach, Perez et al. conducted a large-scale single-cell eQTL study using around 1.2 million PBMCs from 162 SLE cases and 99 healthy controls [34]. Among 3331 genes with at least one cis-eQTL in a cell type (FDR < 0.05), they identified 535 genes with at least one cell type-specific cis-eQTL. In addition, they reported several examples of colocalizations between single-cell eQTL and SLE-GWAS signals. One example is *ORMDL3*, a regulator of sphingolipid biosynthesis and ubiquitously expressed across

cell types but showed eQTL-GWAS colocalization specifically in B cells, $CD8^+$ T cells, and plasmacytoid dendritic cells with sufficiently high posterior probabilities (>90%).

Since the single-cell eQTL study is a relatively new field, its analytical strategy has not yet been fully matured. A promising alternative approach is the association test preserving single-cell resolution data structure. Three major hurdles for this approach are the sparsity in the expression data, the multiple repeated measurements from a donor, and a substantial amount of experimental noise. Assuming the expression count data follows a Poisson distribution, we can mitigate all hurdles using a Poisson mixed effects (PME) model with appropriate covariates to adjust confounding factors for every single cell and a random effect term for repeated sampling of a single donor. Indeed, Nathan et al. successfully deploy a PME model to a large-scale single-cell dataset comprising more than 500,000 unstimulated memory T cells from 259 donors [35]. This study demonstrated the utility of the PME model single-cell eQTL analysis to detect the cell-state dependency of eQTL effects. Using this model, they successfully showed that risk variants of autoimmunity were enriched in cell-state-dependent eQTLs (e.g., ORMDL3 and CTLA4 loci), indicating that cell-state context is crucial to understanding the genetic etiology of autoimmunity.

Although previous eQTL studies substantially contributed to a better understanding of autoimmunity pathology, these studies have primarily focused on the quantitative aspect of gene expression. However, its qualitative aspect is also critical for cellular biology and the immune system. RNA splicing is crucial to enhance the complexity of protein sequences and functions, and almost all genes have splicing isoforms. Therefore, splicing quantitative trait loci (sQTL) may illuminate autoimmunity pathology not explained by eQTL alone. sQTL analytical strategy is much more complicated than eQTL; we summarized sQTL methods used in the previous studies (Table 1).

One of the apparent challenges in splice isoform quantification is that most RNA-seq platforms are short-read sequencing. Typically, the read pair only covers a few hundred bases at most, whereas the median length of mRNAs is around 3000 base pairs [36]. Therefore, we cannot directly capture the entire isoform structure in most cases using short-read sequencing. On the other hand, we can directly capture splice junctions even using short-read sequencing.

Table 1. Software for splicing isoform quantifications.

Software	Year	Method	Annotation	Novel Isoform Detection	Features
LeafCutter [37]	2018	Event	Not required	Yes	Focused on the variation in "intron" splicing. Used in many sQTL studies. Computationally efficient and accurate at detecting splicing events.
DEXSeq [38]	2012	Event	Required	No	Focused on differentially used exons. Analyzes replicate RNA-seq data.
rMATS [39]	2014	Event	Required	Yes	Accounts for sampling uncertainty and variability.
SUPPA2 [40]	2015	Event	Required	No	High accuracy at low sequencing depth and short read length.
MAJIQ [41]	2016	Event	Required	Yes	Designed to detect "complex" splice variations (e.g., alternative splice site and intron retention)
Cufflinks [42]	2012	Isoform	Not required	Yes	Early-phase software developed in 2010. A transcriptome assembler (it can estimate novel isoform structures). A successor software (stringTie) has already been developed.
StringTie2 [43]	2019	Isoform	Not required	Yes	Capable of assembling both short and long reads. Higher accuracy for assembling complicated isoforms (those with many exons) than Cufflinks.

Table 1. *Cont.*

Software	Year	Method	Annotation	Novel Isoform Detection	Features
RSEM [44]	2011	Isoform	Required	No	Available for organisms lacking sequenced genomes. Computationally intensive.
Salmon [45]	2017	Isoform	Required	No	Fast quantification due to alignment-free quantification. Accounts for sample-specific bias.
Kallisto [46]	2016	Isoform	Required	No	Fast quantification due to alignment-free quantification. Pseudoaligns the reads to the reference avoiding alignment of individual bases.

Year, the year of publication; method, the method of splice isoform quantification (either of splice event- or isoform-level quantification); annotation, requirements of annotation files (e.g., GTF file); novel isoform detection and the ability to detect novel splice isoform(s).

Leafcutter is a leading software widely used for splicing event detection [37]. Leafcutter extracts introns from reads that span between two exons from each sample integrates these across samples and defines a group of introns that share at least one splice site as an intron cluster. Leafcutter then calculates an intron excision ratio for each sample. Changes in this ratio provide a quantitative view of splicing changes. Leafcutter has been used in numerous studies, particularly in sQTL studies [47].

Leafcutter has multiple advantages over other splicing detection methods. Leafcutter does not require an existing annotation file, allowing for identifying novel splicing events. In addition, while other methods for quantifying exons (DEXSeq [38], rMATS [39], and SUPPA2 [40]) are unstable due to ambiguity in assigning reads that map to multiple isoforms of a gene, Leafcutter solves this problem by quantifying introns instead of exons. On the other hand, Leafcutter has several disadvantages. We cannot directly compare the Leafcutter results from different datasets because the definition of intron clusters is dataset-dependent. In addition, relating splicing events to transcript-level quantification is often tricky.

Instead of detecting splicing events at the exon junctions, we can computationally estimate the abundance of full-length transcripts from short-read sequence data (e.g., RSEM [44] and Cufflinks [42]), although the accuracy is relatively low. We can use these estimates to test the associations between the isoform usage ratio and genetic variants. For example, we used Cufflinks in our previous study and found an intriguing sQTL signal; rs10466829, a multiple sclerosis risk variant, showed an sQTL effect on *CLECL1* without noticeable eQTL effect in B cells [31]. This unique pattern (sQTL without eQTL) reflects that the expression of two major isoforms of *CLECL1* (NM_001267701 and NM_172004) were oppositely correlated with the risk variant. These isoforms differ only in the five amino acid residues at the extracellular domain of *CLECL1*. As exemplified by this result, sQTL studies can narrow candidate molecular etiology to specific molecule positions.

Inaccurate isoform quantification is partially caused by incomplete reference datasets we use for isoform quantification [48]. For example, some disease-causing isoforms have incomplete coding sequences in the GENCODE annotation [49]. Furthermore, even if all constituent exons are identified, complete isoform reconstruction from short-read data remains challenging [50].

In contrast to short-read sequencing, long-read sequencing techniques can generate reads of 10 kb or more and sequence full-length isoforms [51]. In one of our recent studies, we obtained a full isoform picture of the *PADI4* gene using long-read sequencing and found a novel non-functional splicing isoform lacking a functional domain [1]. With this updated *PADI4* isoform reference data, we re-analyzed one of our short-read sequencing datasets (Ref. [31]) and quantified *PADI4* isoform abundance. The splicing QTL signal for this novel *PADI4* isoform colocalized with the RA-GWAS signal [1]. This research direction

is currently expanding. Inamo et al. performed long-read sequencing to create a complete isoform reference panel of fine-sorted immune cells, which improves the quality of future sQTL studies using immune cells (https://www.biorxiv.org/content/10.1101/2022.09.13.507708v1, accessed on 2 January 2023).

One of the most challenging and scientifically intriguing questions researchers have been asking is the cell type or tissue specificity of genetic effects. The GTEx v8 project shows that cis-sQTLs were significantly more tissue-specific than cis-eQTLs when considering all mapped cis-QTLs [52]. However, this pattern is reversed when considering only those cis-QTLs where the gene or splicing event is quantified in all tissues. This observation indicates that splicing measures are more tissue-specific than gene expression; in contrast, genetic regulation on splicing tends to be more shared, which suggests that it might be better to use the same cell types or tissues to investigate the effect of splicing on traits.

2. Discussion

As we introduced in this manuscript, V2F studies successfully identified several candidate causal mechanisms of the risk variants. However, many risk variants remain functionally characterized. Chun et al. evaluated how much of the autoimmunity risk variants can be explained by eQTLs discovered in the previous studies analyzing three major immune subpopulations [53]. To this end, they developed a new analytical method called joint likelihood mapping (JLIM) and found that eQTL signals only account for around 25% of the risk loci. Although sQTL can explain an additional fraction of heritability independent from eQTL, the gain in the ratio is relatively limited [54]. To further evaluate the eQTL-mediated autoimmunity genetic risk, Yao et al. developed a sophisticated method called mediated expression score regression (MESC) that accounts for genome-wide GWAS and eQTL signals [55]. They applied MESC to GWAS results for Crohn's disease and eQTL results obtained in immune cells and found that gene expression levels mediated only around 20% of heritability. If we assume all non-coding risk variants possess eQTL or sQTL in specific immune cell types (although we admit this is an over-simplified scenario), these results suggest that the previous QTL projects have failed to detect such QTL signals; we call this "missing QTL".

How can we solve the missing QTL problem? The straightforward approach will be diversifying the cellular conditions (e.g., various stimulatory conditions) where we test eQTL and sQTL. In addition, we can use single-cell transcriptomes to improve cellular resolution. However, of course, the culprit may be other molecular phenotypes, not expression and splicing, such as RNA editing [56] and other omics (e.g., metabolomics). Large-scale functional genomic experiments may not be a single solution. For example, the recent rapid progress of machine learning technologies started to solve the regulatory codes in our genome [57,58], i.e., we can partially infer the variant's function solely based on the genomic sequence patterns around that variant. At the moment, we have not yet reached a conclusion about what the best approach is to maximize biological information extracted from GWAS outputs. In any case, we need to scale up V2F studies further.

Author Contributions: Conceptualization, K.I.; writing—original draft preparation, H.H.; writing—review and editing, K.I.; visualization, H.H.; supervision, K.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ishigaki, K.; Sakaue, S.; Terao, C.; Luo, Y.; Sonehara, K.; Yamaguchi, K.; Amariuta, T.; Too, C.L.; Laufer, V.A.; Scott, I.C.; et al. Multi-Ancestry Genome-Wide Association Analyses Identify Novel Genetic Mechanisms in Rheumatoid Arthritis. *Nat. Genet.* **2022**, *54*, 1640–1651. [CrossRef]
2. Okada, Y.; Wu, D.; Trynka, G.; Raj, T.; Terao, C.; Ikari, K.; Kochi, Y.; Ohmura, K.; Suzuki, A.; Yoshida, S.; et al. Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery. *Nature* **2014**, *506*, 376–381. [CrossRef] [PubMed]
3. Bentham, J.; Morris, D.L.; Graham, D.S.C.; Pinder, C.L.; Tombleson, P.; Behrens, T.W.; Martín, J.; Fairfax, B.P.; Knight, J.C.; Chen, L.; et al. Genetic Association Analyses Implicate Aberrant Regulation of Innate and Adaptive Immunity Genes in the Pathogenesis of Systemic Lupus Erythematosus. *Nat. Genet.* **2015**, *47*, 1457–1464. [CrossRef] [PubMed]
4. Langefeld, C.D.; Ainsworth, H.C.; Cunninghame Graham, D.S.; Kelly, J.A.; Comeau, M.E.; Marion, M.C.; Howard, T.D.; Ramos, P.S.; Croker, J.A.; Morris, D.L.; et al. Transancestral Mapping and Genetic Load in Systemic Lupus Erythematosus. *Nat. Commun.* **2017**, *8*, 16021. [CrossRef] [PubMed]
5. Yin, X.; Kim, K.; Suetsugu, H.; Bang, S.-Y.; Wen, L.; Koido, M.; Ha, E.; Liu, L.; Sakamoto, Y.; Jo, S.; et al. Meta-Analysis of 208370 East Asians Identifies 113 Susceptibility Loci for Systemic Lupus Erythematosus. *Ann. Rheum. Dis.* **2021**, *80*, 632–640. [CrossRef] [PubMed]
6. Okada, Y.; Kim, K.; Han, B.; Pillai, N.E.; Ong, R.T.-H.; Saw, W.-Y.; Luo, M.; Jiang, L.; Yin, J.; Bang, S.-Y.; et al. Risk for ACPA-Positive Rheumatoid Arthritis Is Driven by Shared HLA Amino Acid Polymorphisms in Asian and European Populations. *Hum. Mol. Genet.* **2014**, *23*, 6916–6926. [CrossRef] [PubMed]
7. Han, B.; Diogo, D.; Eyre, S.; Kallberg, H.; Zhernakova, A.; Bowes, J.; Padyukov, L.; Okada, Y.; González-Gay, M.A.; Rantapää-Dahlqvist, S.; et al. Fine Mapping Seronegative and Seropositive Rheumatoid Arthritis to Shared and Distinct HLA Alleles by Adjusting for the Effects of Heterogeneity. *Am. J. Hum. Genet.* **2014**, *94*, 522–532. [CrossRef]
8. Fries, J.F.; Wolfe, F.; Apple, R.; Erlich, H.; Bugawan, T.; Holmes, T.; Bruce, B. HLA-DRB1 Genotype Associations in 793 White Patients from a Rheumatoid Arthritis Inception Cohort: Frequency, Severity, and Treatment Bias. *Arthritis Rheum.* **2002**, *46*, 2320–2329. [CrossRef]
9. de Vries, N.; Tijssen, H.; van Riel, P.L.C.M.; van de Putte, L.B.A. Reshaping the Shared Epitope Hypothesis: HLA-Associated Risk for Rheumatoid Arthritis Is Encoded by Amino Acid Substitutions at Positions 67-74 of the HLA-DRB1 Molecule. *Arthritis Rheum.* **2002**, *46*, 921–928. [CrossRef]
10. Alcina, A.; Abad-Grau, M.D.M.; Fedetz, M.; Izquierdo, G.; Lucas, M.; Fernández, O.; Ndagire, D.; Catalá-Rabasa, A.; Ruiz, A.; Gayán, J.; et al. Multiple Sclerosis Risk Variant HLA-DRB1*1501 Associates with High Expression of DRB1 Gene in Different Human Populations. *PLoS ONE* **2012**, *7*, e29819. [CrossRef]
11. Raychaudhuri, S.; Sandor, C.; Stahl, E.A.; Freudenberg, J.; Lee, H.-S.; Jia, X.; Alfredsson, L.; Padyukov, L.; Klareskog, L.; Worthington, J.; et al. Five Amino Acids in Three HLA Proteins Explain Most of the Association between MHC and Seropositive Rheumatoid Arthritis. *Nat. Genet.* **2012**, *44*, 291–296. [CrossRef] [PubMed]
12. Hu, X.; Deutsch, A.J.; Lenz, T.L.; Onengut-Gumuscu, S.; Han, B.; Chen, W.-M.; Howson, J.M.M.; Todd, J.A.; de Bakker, P.I.W.; Rich, S.S.; et al. Additive and Interaction Effects at Three Amino Acid Positions in HLA-DQ and HLA-DR Molecules Drive Type 1 Diabetes Risk. *Nat. Genet.* **2015**, *47*, 898–905. [CrossRef] [PubMed]
13. Ishigaki, K.; Lagattuta, K.A.; Luo, Y.; James, E.A.; Buckner, J.H.; Raychaudhuri, S. HLA Autoimmune Risk Alleles Restrict the Hypervariable Region of T Cell Receptors. *Nat. Genet.* **2022**, *54*, 393–402. [CrossRef]
14. Acosta-Herrera, M.; Kerick, M.; González-Serna, D.; Myositis Genetics Consortium; Scleroderma Genetics Consortium; Wijmenga, C.; Franke, A.; Gregersen, P.K.; Padyukov, L.; Worthington, J.; et al. Genome-Wide Meta-Analysis Reveals Shared New Loci in Systemic Seropositive Rheumatic Diseases. *Ann. Rheum. Dis.* **2019**, *78*, 311–319. [CrossRef] [PubMed]
15. 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]
16. Aggarwal, R.; Liao, K.; Nair, R.; Ringold, S.; Costenbader, K.H. Anti-Citrullinated Peptide Antibody Assays and Their Role in the Diagnosis of Rheumatoid Arthritis. *Arthritis Rheum.* **2009**, *61*, 1472–1483. [CrossRef]
17. Fox, B.J.; Hockley, J.; Rigsby, P.; Dolman, C.; Meroni, P.L.; Rönnelid, J. A WHO Reference Reagent for Lupus (Anti-DsDNA) Antibodies: International Collaborative Study to Evaluate a Candidate Preparation. *Ann. Rheum. Dis.* **2019**, *78*, 1677–1680. [CrossRef]
18. Germain, R.N. T-Cell Development and the CD4-CD8 Lineage Decision. *Nat. Rev. Immunol.* **2002**, *2*, 309–322. [CrossRef]
19. Rossjohn, J.; Gras, S.; Miles, J.J.; Turner, S.J.; Godfrey, D.I.; McCluskey, J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol.* **2015**, *33*, 169–200. [CrossRef]
20. Scally, S.W.; Petersen, J.; Law, S.C.; Dudek, N.L.; Nel, H.J.; Loh, K.L.; Wijeyewickrema, L.C.; Eckle, S.B.G.; van Heemst, J.; Pike, R.N.; et al. A Molecular Basis for the Association of the HLA-DRB1 Locus, Citrullination, and Rheumatoid Arthritis. *J. Exp. Med.* **2013**, *210*, 2569–2582. [CrossRef]
21. Hill, J.A.; Southwood, S.; Sette, A.; Jevnikar, A.M.; Bell, D.A.; Cairns, E. Cutting Edge: The Conversion of Arginine to Citrulline Allows for a High-Affinity Peptide Interaction with the Rheumatoid Arthritis-Associated HLA-DRB1*0401 MHC Class II Molecule. *J. Immunol.* **2003**, *171*, 538–541. [CrossRef] [PubMed]

22. Kwok, W.W.; Domeier, M.L.; Raymond, F.C.; Byers, P.; Nepom, G.T. Allele-Specific Motifs Characterize HLA-DQ Interactions with a Diabetes-Associated Peptide Derived from Glutamic Acid Decarboxylase. *J. Immunol.* **1996**, *156*, 2171–2177. [CrossRef]
23. Jabri, B.; Sollid, L.M. T Cells in Celiac Disease. *J. Immunol.* **2017**, *198*, 3005–3014. [CrossRef] [PubMed]
24. Emerson, R.O.; DeWitt, W.S.; Vignali, M.; Gravley, J.; Hu, J.K.; Osborne, E.J.; Desmarais, C.; Klinger, M.; Carlson, C.S.; Hansen, J.A.; et al. Immunosequencing Identifies Signatures of Cytomegalovirus Exposure History and HLA-Mediated Effects on the T Cell Repertoire. *Nat. Genet.* **2017**, *49*, 659–665. [CrossRef] [PubMed]
25. Dash, P.; Fiore-Gartland, A.J.; Hertz, T.; Wang, G.C.; Sharma, S.; Souquette, A.; Crawford, J.C.; Clemens, E.B.; Nguyen, T.H.O.; Kedzierska, K.; et al. Quantifiable Predictive Features Define Epitope-Specific T Cell Receptor Repertoires. *Nature* **2017**, *547*, 89–93. [CrossRef] [PubMed]
26. O'Connor, L.J.; Schoech, A.P.; Hormozdiari, F.; Gazal, S.; Patterson, N.; Price, A.L. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* **2019**, *105*, 456–476. [CrossRef]
27. Ishigaki, K.; Akiyama, M.; Kanai, M.; Takahashi, A.; Kawakami, E.; Sugishita, H.; Sakaue, S.; Matoba, N.; Low, S.-K.; Okada, Y.; et al. Large-Scale Genome-Wide Association Study in a Japanese Population Identifies Novel Susceptibility Loci across Different Diseases. *Nat. Genet.* **2020**, *52*, 669–679. [CrossRef] [PubMed]
28. Trynka, G.; Sandor, C.; Han, B.; Xu, H.; Stranger, B.E.; Liu, X.S.; Raychaudhuri, S. Chromatin Marks Identify Critical Cell Types for Fine Mapping Complex Trait Variants. *Nat. Genet.* **2013**, *45*, 124–130. [CrossRef] [PubMed]
29. De Jager, P.L.; Hacohen, N.; Mathis, D.; Regev, A.; Stranger, B.E.; Benoist, C. ImmVar Project: Insights and Design Considerations for Future Studies of "Healthy" Immune Variation. *Semin. Immunol.* **2015**, *27*, 51–57. [CrossRef]
30. Raj, T.; Rothamel, K.; Mostafavi, S.; Ye, C.; Lee, M.N.; Replogle, J.M.; Feng, T.; Lee, M.; Asinovski, N.; Frohlich, I.; et al. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science* **2014**, *344*, 519–523. [CrossRef] [PubMed]
31. Ishigaki, K.; Kochi, Y.; Suzuki, A.; Tsuchida, Y.; Tsuchiya, H.; Sumitomo, S.; Yamaguchi, K.; Nagafuchi, Y.; Nakachi, S.; Kato, R.; et al. Polygenic Burdens on Cell-Specific Pathways Underlie the Risk of Rheumatoid Arthritis. *Nat. Genet.* **2017**, *49*, 1120–1125. [CrossRef]
32. Schmiedel, B.J.; Singh, D.; Madrigal, A.; Valdovino-Gonzalez, A.G.; White, B.M.; Zapardiel-Gonzalo, J.; Ha, B.; Altay, G.; Greenbaum, J.A.; McVicker, G.; et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **2018**, *175*, 1701–1715. [CrossRef] [PubMed]
33. Ota, M.; Nagafuchi, Y.; Hatano, H.; Ishigaki, K.; Terao, C.; Takeshima, Y.; Yanaoka, H.; Kobayashi, S.; Okubo, M.; Shirai, H.; et al. Dynamic Landscape of Immune Cell-Specific Gene Regulation in Immune-Mediated Diseases. *Cell* **2021**, *184*, 3006–3021. [CrossRef]
34. Perez, R.K.; Gordon, M.G.; Subramaniam, M.; Kim, M.C.; Hartoularos, G.C.; Targ, S.; Sun, Y.; Ogorodnikov, A.; Bueno, R.; Lu, A.; et al. Single-Cell RNA-Seq Reveals Cell Type-Specific Molecular and Genetic Associations to Lupus. *Science* **2022**, *376*, eabf1970. [CrossRef] [PubMed]
35. Nathan, A.; Asgari, S.; Ishigaki, K.; Valencia, C.; Amariuta, T.; Luo, Y.; Beynor, J.I.; Baglaenko, Y.; Suliman, S.; Price, A.L.; et al. Single-Cell EQTL Models Reveal Dynamic T Cell State Dependence of Disease Loci. *Nature* **2022**, *606*, 120–128. [CrossRef] [PubMed]
36. Piovesan, A.; Antonaros, F.; Vitale, L.; Strippoli, P.; Pelleri, M.C.; Caracausi, M. Human Protein-Coding Genes and Gene Feature Statistics in 2019. *BMC Res. Notes* **2019**, *12*, 315. [CrossRef] [PubMed]
37. Li, Y.I.; Knowles, D.A.; Humphrey, J.; Barbeira, A.N.; Dickinson, S.P.; Im, H.K.; Pritchard, J.K. Annotation-Free Quantification of RNA Splicing Using LeafCutter. *Nat. Genet.* **2018**, *50*, 151–158. [CrossRef] [PubMed]
38. Anders, S.; Reyes, A.; Huber, W. Detecting Differential Usage of Exons from RNA-Seq Data. *Genome Res.* **2012**, *22*, 2008–2017. [CrossRef]
39. Shen, S.; Park, J.W.; Lu, Z.-X.; Lin, L.; Henry, M.D.; Wu, Y.N.; Zhou, Q.; Xing, Y. RMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E5593-601. [CrossRef]
40. Trincado, J.L.; Entizne, J.C.; Hysenaj, G.; Singh, B.; Skalic, M.; Elliott, D.J.; Eyras, E. SUPPA2: Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions. *Genome Biol.* **2018**, *19*, 40. [CrossRef]
41. Vaquero-Garcia, J.; Barrera, A.; Gazzara, M.R.; González-Vallinas, J.; Lahens, N.F.; Hogenesch, J.B.; Lynch, K.W.; Barash, Y. A New View of Transcriptome Complexity and Regulation through the Lens of Local Splicing Variations. *Elife* **2016**, *5*, e11752. [CrossRef] [PubMed]
42. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [CrossRef] [PubMed]
43. Kovaka, S.; Zimin, A.V.; Pertea, G.M.; Razaghi, R.; Salzberg, S.L.; Pertea, M. Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2. *Genome Biol.* **2019**, *20*, 278. [CrossRef] [PubMed]
44. Li, B.; Dewey, C.N. RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome. *BMC Bioinformatics* **2011**, *12*, 323. [CrossRef] [PubMed]
45. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef] [PubMed]
46. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-Optimal Probabilistic RNA-Seq Quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef]

47. Castaldi, P.J.; Abood, A.; Farber, C.R.; Sheynkman, G.M. Bridging the Splicing Gap in Human Genetics with Long-Read RNA Sequencing: Finding the Protein Isoform Drivers of Disease. *Hum. Mol. Genet.* **2022**, *31*, R123–R136. [CrossRef]
48. Frankish, A.; Diekhans, M.; Jungreis, I.; Lagarde, J.; Loveland, J.E.; Mudge, J.M.; Sisu, C.; Wright, J.C.; Armstrong, J.; Barnes, I.; et al. GENCODE 2021. *Nucleic Acids Res.* **2021**, *49*, D916–D923. [CrossRef]
49. Yamaguchi, K.; Ishigaki, K.; Suzuki, A.; Tsuchida, Y.; Tsuchiya, H.; Sumitomo, S.; Nagafuchi, Y.; Miya, F.; Tsunoda, T.; Shoda, H.; et al. Splicing QTL Analysis Focusing on Coding Sequences Reveals Mechanisms for Disease Susceptibility Loci. *Nat. Commun.* **2022**, *13*, 4659. [CrossRef]
50. Steijger, T.; Abril, J.F.; Engström, P.G.; Kokocinski, F.; RGASP Consortium; Hubbard, T.J.; Guigó, R.; Harrow, J.; Bertone, P. Assessment of Transcript Reconstruction Methods for RNA-Seq. *Nat. Methods* **2013**, *10*, 1177–1184. [CrossRef]
51. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biol.* **2020**, *21*, 30. [CrossRef]
52. GTEx Consortium the GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, *369*, 1318–1330. [CrossRef]
53. Chun, S.; Casparino, A.; Patsopoulos, N.A.; Croteau-Chonka, D.C.; Raby, B.A.; De Jager, P.L.; Sunyaev, S.R.; Cotsapas, C. Limited Statistical Evidence for Shared Genetic Effects of EQTLs and Autoimmune-Disease-Associated Loci in Three Major Immune-Cell Types. *Nat. Genet.* **2017**, *49*, 600–605. [CrossRef] [PubMed]
54. Qi, T.; Wu, Y.; Fang, H.; Zhang, F.; Liu, S.; Zeng, J.; Yang, J. Genetic Control of RNA Splicing and Its Distinct Role in Complex Trait Variation. *Nat. Genet.* **2022**, *54*, 1355–1363. [CrossRef]
55. Yao, D.W.; O'Connor, L.J.; Price, A.L.; Gusev, A. Quantifying Genetic Effects on Disease Mediated by Assayed Gene Expression Levels. *Nat. Genet.* **2020**, *52*, 626–633. [CrossRef] [PubMed]
56. Li, Q.; Gloudemans, M.J.; Geisinger, J.M.; Fan, B.; Aguet, F.; Sun, T.; Ramaswami, G.; Li, Y.I.; Ma, J.-B.; Pritchard, J.K.; et al. RNA Editing Underlies Genetic Risk of Common Inflammatory Diseases. *Nature* **2022**, *608*, 569–577. [CrossRef]
57. Zhou, J.; Troyanskaya, O.G. Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nat. Methods* **2015**, *12*, 931–934. [CrossRef]
58. Zhou, J.; Theesfeld, C.L.; Yao, K.; Chen, K.M.; Wong, A.K.; Troyanskaya, O.G. Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk. *Nat. Genet.* **2018**, *50*, 1171–1179. [CrossRef] [PubMed]