Take Your PICS: Moving from GWAS to Immune Function

Stephen Eyre[1,2] and Jane Worthington[1,2,*]

[1]Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK
[2]NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK
*Correspondence: jane.worthington@manchester.ac.uk
http://dx.doi.org/10.1016/j.immuni.2014.12.014

Many of the hits identified through genome-wide association studies are located outside protein-coding regions, making it difficult to define mechanism. In *Nature*, Farh et al., (2014) describe an approach to identify causal variants in autoimmune disease as first step to assigning function.

Although genome-wide association studies (GWASs) have provided insights into disease pathology and supported established theories about the importance of the role of T cells, antigen presentation, and cytokine production in autoimmune disease, the signals are often difficult to interpret, and relatively few GWAS hits have been assigned a molecular function. A common feature of GWASs is that the vast majority of signals (up to 90%) lie outside traditional protein-coding gene sequences. There is now growing evidence that these GWAS signals are enriched in cell-type-specific (Trynka et al., 2013), large active regulatory regions of the genome (Maurano et al., 2012), known as superenhancers (Hnisz et al., 2013). Based on the analysis of 21 autoimmune GWASs, a recent *Nature* paper by Farh et al. describes the development of a unique resource for assigning a probability of single nucleotide polymorphisms (SNPs) being causal in disease (Farh et al., 2014). A series of epigenetic analyses revealed that ~60% of these likely causal variants are located within stimulus and cell-type-specific enhancers, enriched for both histone modifications and the transcription of noncoding RNA.

Using 39 large-scale GWAS studies from a range of traits (21 of which were from autoimmune diseases), Farh et al. initially clustered diseases by their shared genetic loci. Unsurprisingly, this indicated a large overlap for immune-related traits, with 69% of genes for each trait shared by at least one other autoimmune disease. In addition, it revealed a higher degree of clustering between diseases within the general "autoimmune" category. This included evidence for shared genetics between type 1 diabetes and juvenile idiopathic arthritis (but, perhaps surprisingly, not rheumatoid arthritis), psoriasis, ankylosing spondylitis and Crohn's disease, and multiple sclerosis and celiac disease. The authors went on to provide an algorithm (PICS) and easy-to-access portal that allows a SNP to be assigned a probability score of being a causal variant. Currently, researchers attempting to integrate disease causal variants with functional assays select the SNPs for investigation from the GWAS Catalogue (www.genome.gov/gwastudies). This catalog provides the lead associated SNP for all loci passing the established GWAS significance threshold of $p < 5 \times 10^{-8}$. The GWAS arrays used to detect these associations are designed such that each genotyped SNP strongly correlates with a large number of ungenotyped SNPs, allowing comprehensive GWA analysis. Unfortunately, this has the downside that the lead GWAS SNP implicates a large number of ungenotyped SNPs, some a large distance from the genotyped association and unlikely to be the causal variant. Inevitably, this might lead to large numbers of potentially causal variants requiring downstream functional assays. Farh et al. developed a fine-mapping algorithm to refine the set of likely causal variants from GWAS signals. The correlation of SNPs along a chromosome is termed a haplotype, and the authors used large data sets (over 38,000 samples) to capture rare breaks (recombination events) in these haplotypes; this, along with the strength of association at each region, is used to assign each SNP in a region a probabilistic value for being disease causal. The database described by Farh et al. is, therefore, a major step forward in quickly prioritizing a SNP, or more likely group of SNPs, as being responsible for the association signal seen in disease. In fact, the authors went on to show that the "lead" SNP highlighted in the GWAS catalog described above has only a 5% chance of being causal and lies on average 14 kb from the predicted true causal SNP. Further reinforcement is provided by the illustration that the "PICS" SNPs are better at predicting functional annotation than the GWAS catalog SNPs.

Histone modifications are established markers of different chromatin states, and methylation (Me) or acetylation (Ac) of specific histones strongly correlate with promoter or enhancer position and activity (Ernst et al., 2011). For example, histone H3 lysine 3 monomethylation (H3K3Me1) is known to correlate with "poised" (open and accessible regions of the DNA, but not actively driving gene expression) enhancers, whereas H3K27Ac marks active enhancers. Farh et al. went on to examine epigenetic marks in primary human blood immune cells, including resting and stimulated CD4[+] T cell subsets, regulatory T cells, CD8[+] T cells, B cells, and monocytes, under different stimulatory conditions. They found different cell types clustered according to *cis*-regulatory element patterns, and based on clusters of enhancers, cell-type- and stimulation-specific patterns of histone modifications
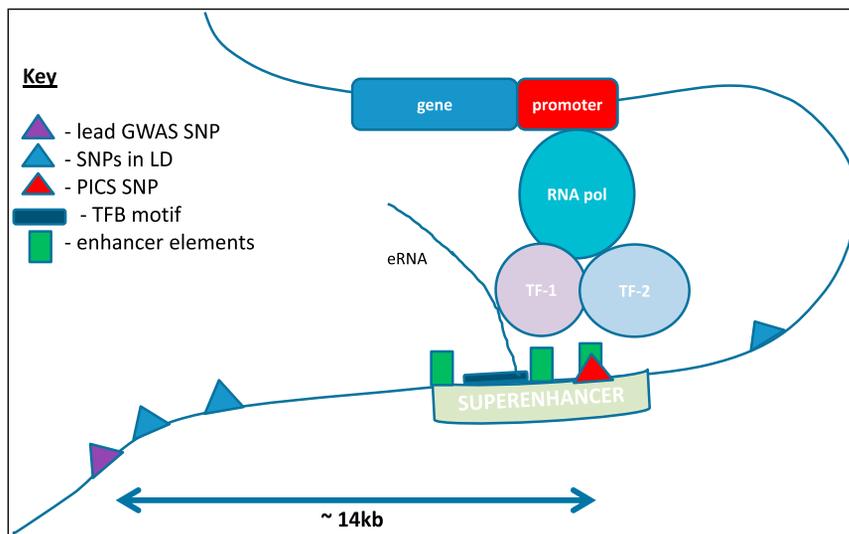
Figure 1. Schematic Diagram of a Generalized Autoimmune Disease Susceptibility Locus
Lead SNPs from GWASs (purple triangle) are correlated with multiple SNPs (blue triangles) at a suscepti-bility locus. Evidence from the paper by Farh et al. indicates that SNPs with the highest probability of being causal (PICS SNPs, red triangle) can lie some distance from the GWAS SNP (median 14 kb). They also show that 90% of PICS SNPs are found outside gene regions, whereas 60% are within enhancer regions. As indicated in the schematic, they find that PICS SNPs are enriched for residing in superenhancers, within discrete enhancer regions, and usually lie close to, but outside, transcription factor-binding motifs.

could be determined. Interestingly, the mark for active enhancers (H3K27ac) was found to be more specific for defining cell-type-specific activity than the H3K4me1 mark for poised en-hancers. Enhancer activity is known to correlate with local gene expression, and using the PICS SNPs defined by Farh et al., potentially causal autoimmune SNPs were found enriched in stimula-tory-specific enhancers, with probable disease-causing variants detected in enhancer regions 60% of the time, but in promoter regions only 8% of the time. Subsequent experiments indicated that the potentially causal variants were found in superenhancers but were confined to discrete regions (Figure 1). Specific anal-ysis of the *IL2RA* region supported the idea that different causal variants occur in different discrete regions of the super-enhancer for different diseases and might well play specific roles in different cell types. For example, PICS SNPs physi-cally close to each other in the IL2RA superenhancer are either associated with autoimmune thyroiditis or multiple sclerosis, but not both. These SNPs clus-ter in two discrete enhancer units within the superenhancer, showing a different pattern of activity in different cell sub-types. Further annotation of the causal

SNPs indicated an enrichment for enhancer RNA expression, a known marker for enhancer activity, in the impli-cated cell types, suggesting that the region that contains the PICS SNP is indeed active in gene regulation. Finally the authors found that, although causal SNPs often map very close to tran-scription factor binding motifs (within 100 bp), they rarely disrupt the motif it-self, suggesting that disease SNPs might not interfere directly with binding of transcription factors; however, as yet, the mechanism via which they operate is unknown.

The authors also report, perhaps surpris-ingly, a lack of evidence that disease causal SNPs correlate strongly with expression of genes (eQTL). As they acknowledge, this might reflect the fact that whole blood was used as the source of eQTL data, whereas disease SNPs are thought to have more subtle effects depending on the cell-type and stimulatory conditions. By mapping of causal variants to cell-type- and stimulatory-specific enhancers, the paper shows how the effects of disease causal variants might be restricted to certain cell types. It is no surprise to see autoimmune diseases associated with causal SNPs located in enhancers regions active in T cell subtypes and B cells.

This paper illustrates the possibility to take findings from GWASs through to identification of functional variants and generation of hypotheses about disease mechanisms. It sets a standard for how to use the wealth of data being generated from large-scale international efforts, such as Blueprint (Martens and Stunnenberg 2013), Encode (ENCODE Project Con-sortium, 2012), and Epigenetic Roadmap (Bernstein et al., 2010) and will transform the way in which genomic and immunoge-netic studies are performed. Informed interrogation of these combined re-sources should underpin the better design of experiments to investigate the immuno-logical consequences of disease-associ-ated genetic variation. Researchers will have improved knowledge of the genetic architecture of a disease locus—not just the lead variant but the potential causative SNP(s) that might be some genomic dis-tance away; they will also know the cell-type and stimulatory conditions most relevant to the specific gene, pathway, and disease of interest. Indeed, the au-thors provide a simple interface to allow researchers to search for SNPs and an easy-to-use portal to determine which cell-type, stimulatory condition and epige-netic mark is most informative for the gene of interest. It is now possible to determine, for example, that four SNPs each have a 23% chance of being causal for the dis-ease of interest, that they all map within a discrete unit of a superenhancer, which is active in CD4$^+$ T cells stimulated to become T helper 17 cells, and that the ge-notype leads to a 20% increase in tran-scription of a key immune gene. This infor-mation then has the capacity to inform study designs to mimic the mode of ac-tion, in primary cells or in vivo models, to illuminate the mechanism by which ge-netic variants increase the risk of disease.

The authors, then, have advanced knowledge of causal variants for autoim-mune diseases, finding that they are enriched in discrete, stimulus dependent regions of superenhancers, correlate with eRNA and histone acetylation and, although often close to transcription fac-tor motifs, rarely alter the motif itself. In addition, they provide valuable and user-friendly data resources to ensure studies attempting to translate genetic findings to functional effects do so armed with empirical data on how to improve study design.

REFERENCES

ENCODE Project Consortium (2012). Nature *489*, 57–74.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). Nat. Biotechnol. *28*, 1045–1048.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X.,

Wang, L., Issner, R., Coyne, M., et al. (2011). Nature *473*, 43–49.

Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2014). Nature. Published online October 29, 2014. http://dx.doi.org/10.1038/nature13835.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Cell *155*, 934–947.

Martens, J.H., and Stunnenberg, H.G. (2013). Haematologica *98*, 1487–1489.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Science *337*, 1190–1195.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Nat. Genet. *45*, 124–130.