REVIEW

Identification of rheumatoid arthritis causal genes using functional genomics

James Ding1 | Gisela Orozco1,2

1Arthritis Research UK Centre for Genetics and Genomics, Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK
2NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

Correspondence
Gisela Orozco, Arthritis Research UK Centre for Genetics and Genomics, The University of Manchester, Manchester, UK.
Email: gisela.orozco@manchester.ac.uk

Funding information
Wellcome Trust, Grant/Award Number: 207491/Z/17/Z; Arthritis Research UK, Grant/Award Number: 21348, 21754

Abstract
Over the past decade, genome-wide association studies have contributed a wealth of knowledge to our understanding of polygenic disorders such as rheumatoid arthritis. As the size of sample cohorts has improved so too have the computational and experimental methods used to robustly define variants associated with disease susceptibility. The challenge now remains to translate these findings into improved understanding of disease aetiology and patient care. Whilst much of the focus of translating the findings of genome-wide association studies has been on global analysis of all variants identified, careful functional study of individual disease susceptibility loci will be required in order to refine our understanding of how individual variants contribute to disease risk. Here, we present the argument behind such an approach and describe some of the novel tools being used to investigate risk loci. This includes the use of chromosomal conformation capture techniques and modifications of the CRISPR-Cas9 system, with several examples of their implementation being described.

1 INTRODUCTION

Rheumatoid arthritis (RA) is a common autoimmune disease with an estimated global prevalence of approximately 25 million.1 The disease is characterized by the inflammation of multiple synovial joints and has a chronic course. Pain and reduced mobility associated with this inflammation severely impact on patients’ quality of life and, if uncontrolled, can cause irreversible damage to joints, often resulting in disability.2

The prevalence of RA is significantly higher amongst women and increases with age,3 such that the prevalence of RA is projected to increase. This is set to amplify the therapeutic costs and high rates of work disability,2 combining to create a major health care and socioeconomic burden.

Current therapeutic options include combinations of disease modifying anti-rheumatic drugs (DMARDs), such as methotrexate, and biologic drugs, such as tumour necrosis factor (TNF) inhibitors. However, the efficacy of these treatments varies widely amongst RA patients and there is still an unmet need for novel, affordable therapies.4 A more detailed understanding of RA pathogenesis and how it is affected by the mechanism of action for these drugs is likely to improve our ability to predict therapeutic response, ensuring that patients receive optimal treatment within the window of opportunity that exists between the emergence of preclinical features and the onset of joint damage.

2 RHEUMATOID ARTHRITIS AETIOLOGY

The role of autoimmunity in RA is highlighted by the prevalence of autoimmune antibodies in patients’ sera. These include rheumatoid factor, which recognizes the fragment crystallizable region of immunoglobulin G and is present in 70%-80% of established RA,5 and anti-citrullinated protein antibodies (ACPAs), present in 50%-70% of
RA patients. Clinical evidence, along with the association of different environmental and genetic factors with ACPA-positive or -negative RA, suggests that ACPA status may discriminate between different forms of RA.

It is thought that ACPA-positive RA emerges as a result of untolerated citrullination of proteins. This results in systemic autoimmune activation; it is thought that RA is manifested in synovial joints as they contain many proteins that normally undergo citrullination. Additional detail, along with how this model applies to ACPA-negative RA is still unclear.

3 | RHEUMATOID ARTHRITIS GENETICS

In keeping with other polygenic disorders, RA susceptibility is modulated by a multitude of loci with low penetrance. These loci combine to account for approximately 60% of the variation in disease risk. To date, genome-wide association (GWA) studies based on data from tens of thousands of patients have helped to identify approximately 100 loci that are robustly associated with disease susceptibility. The reproducibility of many of these associations has been confirmed in multiple populations, with stringent statistical methods being used to ensure that associations are not erroneously reported, as a result of confounding or statistical methods being used to ensure that associations exist, including the dysregulation of chromosomal accessibility. Autoimmunity SNPs are enriched amongst DNAse I hypersensitivity sites (sites of accessible chromatin), and thousands of SNPs have been demonstrated to impact chromatin accessibility (DNase I sensitivity quantitative trait loci). Many bioinformatics tools exist to explore the potential effects of individual SNPs. For example, RegulomeDB collates an array of data including information relating to transcription factor (TF) binding, chromatin accessibility and chromatin state. It is, however, important to recognize that in reality these SNPs do not exist in isolation and it is possible that multiple SNPs within an LD block have a cumulative effect. A SNP may be considered more likely to be causal if it overlaps with a DNAse I hypersensitivity site or TF binding site, or if it disrupts a TF binding motif.

When considering the functional implications of variants identified through GWA studies, it is important to recognize that they typically tag a large collection of variants that are highly correlated due to linkage disequilibrium (LD). It is possible that any of these variants, or indeed a combination of them, may be responsible for mediating the observed increase in disease susceptibility. Bayesian approaches, such as the Probabilistic Identification of Causal SNPs (PICS) have been applied in order to infer causality based on underlying haplotype structure and patterns of association at a given locus. PICS enables generalization based on the collections of risk loci; however, at a given locus the true causality of individual SNPs remains to be experimentally resolved.

RA PICS are enriched within super-enhancers, large clusters of transcriptional enhancers that drive cell identity. This same enrichment is true for RA SNPs more generally and is also true of SNPs associated with other autoimmune diseases. It is likely that SNPs contained within enhancers, or super-enhancers, affect transcriptional regulation. Many mechanisms explaining how this effect may be mediated exist, including the dysregulation of chromosomal accessibility. Autoimmunity SNPs are enriched amongst DNAse I hypersensitivity sites (sites of accessible chromatin), and thousands of SNPs have been demonstrated to impact chromatin accessibility (DNase I sensitivity quantitative trait loci).

A minority of the variants that constitute RA susceptibility loci are known to affect the amino acid sequence of protein-coding genes, such as the Human Leukocyte Antigen (HLA), or Protein tyrosine phosphatase, non-receptor type 22 (PTPN22). Variants located within the coding sequence of these genes are common to many autoimmune diseases and generally have relatively high odds ratios (OR), signifying a relatively high impact on the likelihood of developing RA; however, they are outnumbered by the vast majority of identified loci which have smaller ORs and are generally found in non-coding regions (Figure 1). Of the 46 RA susceptibility variants that have been fine mapped using the Immunochip custom array, only seven potentially causal exonic variants were identified (approximately 85% non-coding). This is in keeping with other polygenic disorders and autoimmune diseases in general, where it has been estimated that 90% of causal single nucleotide polymorphism (SNP) associations lie outside of protein-coding regions.

![FIGURE 1](image)

**FIGURE 1** European RA genetic susceptibility loci and their odds ratios as reported as part of a trans-ethnic meta-analysis (CI, confidence interval; OR, odds ratio).
The underlying data sets are populated with annotations based on sequencing data from chromatin immunoprecipitation (ChIP) and chromatin accessibility experiments. These experiments map the prevalence of various features across the genome, in a variety of different cell types. Such features include TF occupancy, histone modification occupancy and DNase/transposase accessibility. In the majority of instances, these data have been generated through international collaborative efforts, including the NIH Roadmap Epigenomics Consortium, the Functional Annotation of the Mammalian Genome and constituents of the International Human Epigenome Consortium, such as the Encyclopedia of DNA Elements (ENCODE).

Several methods have been developed to integrate functional annotation data with GWA study summary statistics, including methods that first identify features that are enriched with risk loci and then prioritize variants that overlap these features.

GWA study findings have already contributed to our understanding of RA. For example, an enrichment of susceptibility variants in enhancers active in T cells supports a role for these cells in the aetiology of RA. In addition, the most significant association between peaks of histone 3 lysine 4 trimethylation (H3K4me3), a marker of active promoters, and 31 RA SNPs was found in T helper cells.

Synovial fibroblasts (SFs) represent an additional cell type, demonstrated to play a crucial role in mediating the inflammation and irreversible damage characteristic of RA synovial joints. These innate immune cells are activated within the synovial joints of patients with RA and show progressive changes in their epigenome, based on disease progression and severity. To date, there is little genetic evidence supporting their causality in RA; however, it is clear that SFs play an important role. The paucity of genetic evidence may be due to a lack of epigenetic annotations, compared to frequently studied peripheral blood cell types, such as T helper cells.

The findings of GWA studies have already initiated the repositioning of biologic DMARDs targeting components of the Interleukin 23 pathway to treat psoriasis, psoriatic arthritis, ankylosing spondylitis and IBD. Novel therapeutic strategies for RA have also been identified, including peptidyl arginine deiminase (PAD) inhibitors and Janus kinase (JAK) inhibitors. PADs orchestrate citrullination and include PAD type IV, which overlaps an RA susceptibility locus (rs2301888). Similarly, tyrosine kinase 2 (TYK2), a member of the JAK family also overlaps an RA susceptibility locus (rs34536443).

Efforts to further translate the findings of GWA studies into improvements in our understanding of RA and improved patient care are hampered by an inability to accurately assign variants and overlapping enhancers to protein-coding genes, relevant biological processes and cell types.

## 4 Identifying Genes That Mediate the Effect of Non-Coding Variants

Whilst non-coding RNA and trans-regulatory elements should not be overlooked, most non-coding variants are thought to disrupt cis-regulatory elements, such as enhancers. Non-coding variants are typically assumed to affect the nearest protein-coding gene known to have immunological relevance; however, this has been demonstrated to be misguided in some instances. For example, C-type lectin domain family 16, member A (CLEC16A) has been historically associated with several autoimmune diseases, including type 1 diabetes (rs12708716), with disease susceptibility variants found along its length and especially within intron 19. It has recently been demonstrated that this intron forms part of a chromatin interaction with the promoter of a neighbouring gene, homolog of dexamethasone-induced protein (DEXI), and that variation in genotype at this locus correlates with DEXI expression. The CLEC16A/DEXI example outlined above highlights the utility of functional data, such as genotype-specific expression or chromosomal conformation data, in identifying the protein-coding genes through which the effect of disease-associated variants is mediated.

### 4.1 Expression Quantitative Trait Loci

Variants that correlate with the expression of a given transcript are termed expression quantitative trait loci (eQTLs) and a number of large databases of eQTLs exist, such as BloodeQTL, the Genotype-Tissue Expression (GTEx) project and the Immune Variation (ImmVar) project. GTEx demonstrated that eQTLs in relevant pathogenic tissues are significantly enriched for trait associations and may explain a substantial proportion of heritability (40%–80%). However, many susceptibility loci, such as those associated with RA, may have an effect that is too subtle to be captured in such databases, which can be derived from relatively low numbers of samples. The minimal impact of individual variants upon disease susceptibility would suggest that any discernible effect may be too small to be captured and would be highly specific to cell type or stimulatory conditions.

Whilst eQTL data are often comprised of heterogeneous cell types, such as whole blood, several studies have looked more closely at the constituent cell types. For example, Ye et al identified 39 loci that explain 25% of the variation in the response of T helper cells to in vitro stimulation. These loci overlap with disease susceptibility loci such as the RA susceptibility locus found in the interleukin 2 receptor subunit alpha gene (IL2RA). In addition, eQTLs are often specific to cell types and stimulatory conditions, with stimulation specific
eQTLs identified in CD14+ monocytes being demonstrated to be enriched amongst GWA study loci.\textsuperscript{38} Taken to its extremity, this approach of focussing on ever more tightly defined cell types and conditions has led to the identification of eQTLs using single-cell RNA sequencing (scRNA-seq) of peripheral blood mononuclear cells.\textsuperscript{39} Ultimately, the most relevant collection may be carefully defined cell types of known clinical relevance isolated from patient samples, as performed on T helper cells and B cells isolated from patients with RA.\textsuperscript{40}

### 4.2 | Chromosomal interactions

Certain aspects of the network of chromosomal interactions that occurs within the three-dimensional genome are thought to be less variable across cell types than eQTLs. For example, topologically associated domains (TADs), or regions of the genome within which interactions occur at a higher frequency than with neighbouring regions, are thought to be fairly robust across cell types.\textsuperscript{41} It may, therefore, be easier to identify potential causal genes based on TADs. However, the relative strength of individual interactions, occurring within TADs, varies across cell types, with interaction strength between the enhancers and promoters of a given gene thought to correlate with expression of that gene. This has been demonstrated for individual loci, such as the β-globin locus\textsuperscript{42}; however, it is difficult to demonstrate on a global scale due to the limited resolution of chromosomal conformation data and the confounding correlation of various other features, such as chromosomal accessibility.\textsuperscript{43}

The prevailing model of enhancer function comprises of chromosomal looping bringing promoters and enhancers into physical proximity and generating a transcriptionally permissive environment (as reviewed elsewhere\textsuperscript{41,44}). Whilst this model was originally developed based on experiments using fluorescence in situ hybridization,\textsuperscript{45,46} ligation-based methods are currently widely used to characterize interactions.\textsuperscript{43} These methods rely on fixation of nuclei, fragmentation and ligation of chromosomal fragments that are found in close physical proximity. When followed by qPCR using primers specific to the interaction being interrogated, this is commonly termed chromosomal conformation capture (3C)\textsuperscript{47}; however, it is also possible to use high-throughput sequencing to quantify all interactions, termed Hi-C.\textsuperscript{48} The number of possible interactions is vast, and as a result, Hi-C data are typically sparse and render a low-resolution picture of the chromosomal interactome, suitable for identifying TADs. High-throughput relative quantification of specific interactions can be achieved through positive selection of ligated fragments prior to sequencing. This can be achieved using a library of bait oligonucleotides (Capture Hi-C, Chi-C)\textsuperscript{49} or an antibody against a specific protein of interest (HiChIP),\textsuperscript{50} such as a TF. A number of variations of these techniques exist, with an array of names and acronyms, such as Circularized 3C (4C\textsuperscript{51}), 4C followed by high-throughput sequencing (4C-seq\textsuperscript{52}), 3C Seq\textsuperscript{53} Carbon copy 3C (5C\textsuperscript{54}), ChIP-loop\textsuperscript{55} and chromatin interaction analysis by paired-end tag sequencing (ChiA-PET\textsuperscript{56}); typically, they are categorized according to the variety of interactions characterized, for example one vs one, 3C; all vs all, Hi-C; many vs all, Chi-C\textsuperscript{57} (Figure 2).

Several studies have performed CHi-C using a bait library that captures the promoters of 89% of Ensembl protein-coding, non-coding, antisense, snRNA, miRNA and snoRNA transcripts, with only technical considerations being used to exclude 11% of promoters for these transcript types. Using this bait library, in a lymphoblastoid cell line and CD34+ haematopoietic progenitor cells, it was demonstrated that disease-associated SNPs, identified through GWA studies, are enriched amongst fragments that interact with promoters.\textsuperscript{58}

The same bait library has been used to produce high-resolution maps of promoter contacts in 17 human primary blood cell types, demonstrating that promoter interactions are cell type-specific, frequently occur between active promoters and enhancers, and that eQTLs for a given gene are enriched amongst the fragments identified as interacting with the promoter of that gene.\textsuperscript{59} Within the breast cancer field, a bait library designed to capture interactions occurring with 68 disease-associated loci, characterized through GWA studies, was used to identify putative causal genes. The authors identify 110 genes as interacting with 33 breast cancer susceptibility loci and confirm that 22 of these genes have been previously identified through eQTL studies, 32 are associated with breast cancer survival and 14 are somatically mutated in breast cancer.\textsuperscript{60}

HiChIP represents an alternative approach to mapping regulatory interactions occurring within a given cell type. This has been successfully achieved for various T helper cell subtypes, by focussing on interactions that can be immunoprecipitated with H3K27ac\textsuperscript{61} a marker for active chromatin. Of particular relevance to the identification of putative causal genes, the authors identified 2597 genes that interact with enhancers overlapping 684 autoimmune PICS, with only 14% of these genes representing the closest gene to the SNP.

The potential of CHi-C and HiChIP data for identifying putative causal genes is, therefore, well established; however, the statistical methods used to interpret Hi-C data are still being developed and further refinement of the underlying methods remains important in order to improve reproducibility.\textsuperscript{62} In addition, the costs associated with high-throughput sequencing and high cell numbers required by 3C-based methods mean that much of the data generated using techniques such as Hi-C, CHI-C and HiChIP has been generated from low numbers of samples.
PUTTING RA SUSCEPTIBILITY LOCI IN 3D CONTEXT

CHi-C has been used to characterize interactions occurring with RA susceptibility loci, and those of several related diseases, in T and B cell lines. These data provide a key insight into the 3D chromosomal landscape at these loci and can be compared with that generated from primary cells using the promoter capture bait library described earlier.

The resulting picture is understandably complex and can best be summarized using examples of some of the different interactions observed.

5.1 Disease susceptibility loci do not necessarily interact with the nearest gene and may interact with alternative genes located at a long distance

Susceptibility variants associated with RA and juvenile idiopathic arthritis and found at chromosomal position 13q14.1 (rs9603616 and rs7993214, respectively) are located within an intron of COG6 (component of oligomeric Golgi complex 6), which is required for normal Golgi function (Figure 3A). CHi-C has revealed that, in

| FIGURE 2 | Summary of a selection of chromosomal conformation capture (3C)-based techniques. Common steps and distinguishing features are highlighted. Adapted from Pombo et al. |

| One vs One | Many vs Many |
| 3C | ChIP-loop |
| • qPCR using primers to regions of interest | • 3C with immunoprecipitation based on a protein of interest |
| 5C | |
| • Annealment of a pool of primers followed by high-throughput sequencing |
| One vs All | Many vs All |
| 4C | 4C-seq |
| • Circularization, amplification and microarray analysis | • Circularization, amplification and high-throughput sequencing |
| 3C-seq | • 4C-seq, improved precision through increased fragmentation |
| All vs All | |
| Hi-C | ChIA-PET |
| • Biotinylation of ligation products, followed by pull down and high-throughput sequencing | • Hi-C with immunoprecipitation based on a protein of interest |
| Many vs All | |
| CHi-C | HiChIP |
| • Bait-specific capture of a Hi-C library | • Antibody-specific capture of a Hi-C library |
both T and B cell lines, this region interacts with the promoter of forkhead box O1 (FOXO1), found over 1 Mb away\(^7\) (whilst no consensus exists, a threshold of 1 Mb is often used to distinguish cis- and trans-regulatory effects). FOXO1 has previously been implicated in RA, as it is crucial to the survival of SFs.\(^6^4\) FOXO1 has also been demonstrated to be hypermethylated in RA SFs\(^6^5\) and therefore represents a strong candidate for mediating the increased disease susceptibility associated with the 13q14.1 risk locus.

### 5.2 Distinct susceptibility loci associated with different diseases may interact with each other, with their effects being mediated by alterations to the same genes

The interaction occurring between a type 1 diabetes susceptibility locus overlapping CLEC16A (rs12708716) and the promoter of DEXI has already been described, as observed by Davison et al.\(^3^3\) This susceptibility locus overlaps a psoriatic arthritis susceptibility locus (rs12928822) and, as well as confirming the presence this interaction in T (Jurkat E6.1) and B (GM12878) cell lines, Martin et al also observed interactions between the DEXI promoter and an RA susceptibility locus found at chromosomal position 16p13.13 (Figure 3B, rs4780401).\(^6^3\) Interactions were also observed between these three disease susceptibility loci, suggesting a common autoimmune mechanism.

### 5.3 Regions can show a complex pattern involving both of the interaction types mentioned above

6q23.3 contains two independent risk loci for RA, systemic lupus erythematosus and coeliac disease, one of which overlaps TNF alpha-induced protein 3 (TNFAIP3, rs7752903), with the second being intergenic (Figure 3C, rs6920220). In addition, an alternative selection of variants, also overlapping TNFAIP3, comprises a risk locus for psoriasis and psoriatic arthritis (rs610604). These regions show a complex network of interactions with each other and with the promoters of various protein-coding genes, including TNFAIP3, interferon gamma receptor 1 (IFNGR1) and interleukin 20 receptor subunit alpha (IL20RA).\(^6^6,6^7\) TNFAIP3 encodes the protein A20. It has been demonstrated that a TT>A polymorphism (rs35926684), associated with RA disease susceptibility and overlapping

---

**FIGURE 3** Illustrations exemplifying chromosomal interactions observed at susceptibility loci for RA and genetically similar diseases. Included are three examples that show long-range interactions between susceptibility loci and non-neighbouring putative causal genes (A), interactions between different disease susceptibility loci (B) and combinations of the aforementioned categories (C) (Adapted from ref.\(^6^3,6^6,6^7\)). Index SNPs for each locus are given in parenthesis and indicated in red. Susceptibility loci included as baits in the region-capture experiment are coloured blue, along with interactions observed (arrows) and overlapping genes. In contrast, chromosomal loci including promoters identified as interacting with susceptibility loci are coloured green, along with promoter capture interactions and associated genes. Interaction strength is conveyed by the colour intensity of associated arrows, with arrowheads identifying non-bait ends. JIA, juvenile idiopathic arthritis; PsA, psoriatic arthritis; T1D, type 1 diabetes.
TNFAIP3, results in impaired binding of the TF NFκB (nuclear factor kappa-light-chain-enhancer of activated B cells) and reduced expression of A20. It is, however, possible that this and other variants associated with RA also affect expression of IFNGR1 and IL20RA. Interestingly increased NFκB binding has been associated with the intergenic RA risk allele in a heterozygous T cell line. The risk genotype also shows a higher frequency of interactions with IFNGR1 and IL20RA in B-lymphoblastoid cell lines and has been correlated with increased expression of IL20RA in primary T helper cells. As a receptor for the pro-inflammatory cytokine interleukin-20, IL20RA is a strong candidate for mediating increased susceptibility for RA. Similarly, Interferon-γ, which binds to IFNGR1, plays an important role in macrophage activation, MHC expression and T helper cell differentiation. In addition, in an African American cohort, IFNGR1 was found to be expressed at elevated levels in blood isolated from patients with RA, whilst the generalizability of this finding is unclear it may support the relevance of this gene in RA.

6 | INVESTIGATING THE IMPACT OF VARIANTS ON PUTATIVE CAUSAL GENES

Chromosomal interaction data can contribute to a body of evidence that aids the identification of putative causal genes. In order to demonstrate conclusively that a variant impacts the expression of a putative causal gene, it is likely that additional techniques will be necessary. The potential contribution of eQTL or other expression data has already been discussed. A further tool that is already proving itself invaluable in the field of functional genetics is the Clustered Regularly Interspersed Palindromic Repeats (CRISPR)-CRISPR-associated protein 9 (Cas9) system.

In its conventional wild-type form, the CRISPR-Cas9 system is capable of targeted generation of double-strand breaks, which are either repaired by non-homologous end-joining (NHEJ) or homology directed repair (HDR). Cycles of Cas9-mediated cleavage and accurate repair are typically escaped as a result of imperfect NHEJ, resulting in small insertions and deletions (INDELs). Alternatively, HDR can be hijacked, through the provision of an exogenous repair template in order to knock-in specific mutations.

For example, Sokhi et al recently used the CRISPR-Cas9 system to study the TT>A polymorphism found on chromosome 6q23.3 and mentioned earlier (Figure 4A). Using bacterial artificial chromosomes (BACs), the authors introduced topologically associated 6q23.3 subdomains into wild-type mice and bred these onto a Tnfaip3 null background. The resultant progeny expressed human Tnfaip3 in B cells, T cells, bone marrow-derived macrophages and SFs in a stimulation responsive manner that recapitulates the human expression pattern. A 789 bp segment that includes the enhancer containing the TT>A mutation was deleted from the BAC using the CRISPR-Cas9 system. This resulted in decreased A20 expression in the human HEK293T cell line and all of the murine cell types mentioned above. In keeping with A20-deficient mice, humanized, TNFAIP3 expressing mice harbouring this targeted deletion have an increased incidence of arthritis.

In addition to its conventional wild-type form, genetically engineered catalytically dead Cas9 (dCas9) proteins are capable of recruiting chromatin modifiers to specific genomic locations. This enables modulation of chromatin such that the impact of activating or inactivating a non-coding region can be determined. This is particularly useful for studying enhancers, where activation or inactivation can be linked to increased or decreased expression of genes which are influenced by that enhancer.

Using a fusion of dCas9 and the transcriptional activator VP64, Simeonov et al were able to demonstrate that activation of an intronic enhancer located within IL2RA increases IL2RA expression (Figure 4B). This enhancer contains a SNP associated with Crohn's disease (rs61839660), where fine mapping has resolved the association, such that it is unlikely any other variant is mediating the associated increase in disease susceptibility. Interestingly, this same variant is also associated with thyroiditis susceptibility and is protective for type 1 diabetes. Simeonov et al went on to introduce this SNP in mice, using the wild-type CRISPR-Cas9 system, and showed that this led to impaired expression of IL2RA upon T cell stimulation (Figure 4C). In addition, a 12 bp deletion, including the Crohn's disease SNP, skewed the response of naïve murine T cells to activation. Using this combination of genetic data and functional experimentation, the authors were able to convincingly demonstrate how disease susceptibility is mediated at this locus. In addition to VP64, alternative activator domains are available with varying potencies and mechanisms. It is also possible to fuse dCas9 and the transcriptionally repressive Krüppel-associated box (KRAB) domain. These techniques are often termed CRISPR activation and CRISPR interference.

The CRISPR-Cas9 system is targeted by a guideRNA (gRNA) that harbours complementarity to the target genomic site. A single gRNA is sufficient to enable targeting to a single location and typically low numbers of gRNAs are used in order to aid the interpretation of their effects. In contrast, CRISPR screens use viral libraries containing thousands of gRNAs, to deliver different single guides to different cells in a population. Typically, following transduction of viral libraries, cells are selected based on a phenotype of interest, such as expression of a particular protein (flow cytometry), or drug resistance.
FIGURE 4 Application of the CRISPR-Cas9 system to the study of RA risk loci. Graphical representations of the experimental strategies employed by Sokhi et al\textsuperscript{72} (A) and Simeonov et al\textsuperscript{73} (B-D). gRNA, guide RNA
enrichment of specific gRNAs in the selected population is assessed by high-throughput sequencing and used as a readout of the gRNAs ability to confer the phenotype of interest. This screening technique was used by Simeonov et al in combination with dCas9-VP64 to identify the IL2RA enhancer mentioned earlier, with flow cytometry used to isolate cells expressing high levels of IL2RA (Figure 4D). Further examples include the combination of flow cytometry and wild-type Cas9 to perform saturating mutagenesis of a BCL11A, known to regulate foetal haemoglobin levels, and the combination of growth selection and dCas9-KRAB to map enhancers at the MYC and GATA1 locus that regulate cellular proliferation. Where the phenotype, or gene, of interest is not amenable to flow cytometry or growth selection a reporter systems can be used, as exemplified by Klann et al at the HBE1 locus. CRISPR screens have also been successfully combined with scRNA-seq to study the impacts of individual guides on the transcriptome more generally. This approach may be more conducive to the study of multiple disease susceptibility loci in a single experiment; however, it comes with additional considerations relating to experimental design and challenges when it comes to interpreting the resulting data.

7 | LIMITATIONS AND CHALLENGES
Identification of causal genes using functional genomics represents a huge challenge in the translation of GWA study findings towards a clinically applicable understanding of RA. The potential benefits of this include the identification of new drug targets and better stratification of patients amongst available therapies. As discussed, there are a
number of novel techniques that can contribute to answering this question; however, there are also several obstacles that warrant discussion.

As illustrated in Figure 1, the median OR for an RA susceptibility locus is 1.11, meaning that, on average, an individual with a given RA susceptibility variant is only 11% more likely to develop RA than an individual who does not carry that variant. As discussed, with reference to eQTL data sets, it is likely that the effect of a given disease susceptibility variant is very small and may only be detectable under the correct conditions. This necessitates the use of sensitive techniques in cell types and stimulatory conditions that are relevant to RA. This obstacle is exacerbated, since the effect of a disease susceptibility locus may be mediated by a combination of SNPs found in LD and through a combination of protein-coding genes.

One method of overcoming the limited effect expected from individual variants is to study the effect of completely disrupting underlying regulatory elements: The effect of chromatin modifiers such as KRAB is not specific to a single nucleotide and can have an effect that spans several hundred nucleotides \(^{76}\); Similarly, whilst a SNP is likely to modulate the activity of an enhancer large deletions are likely to completely abrogate it.

The functional interrogation of individual susceptibility loci is a laborious and technically challenging task and is likely to lack the same statistical rigour and reproducibility that has become a hallmark of GWA studies. \(^{83}\) It is therefore important that experiments are carefully designed, reported and interpreted in order to best contribute to our incrementally improving understanding of RA aetiology. There is, however, an unmet need for such studies, which will undoubtedly contribute to our increasing understanding of how non-coding elements, such as enhancers, function in health and disease.

8 | CONCLUSION

Over the past decade, our ability to confidently describe variants associated with susceptibility to RA and other polygenic disorders has improved dramatically. This includes the imputation of associations for SNPs not experimentally investigated and the performance of high-density fine mapping, to refine observed associations. Recently the focus has switched to understanding how these variants confer disease susceptibility, a question that has largely emerged from the observation that most polygenic disease susceptibility loci do not have obvious, direct effects on protein-coding genes. Novel techniques and tools, such as chromosomal conformation capture, and the CRISPR-Cas9 system are well suited to investigating the effect of disease susceptibility variants and underlying non-coding elements. A workflow illustrating the position of the various experimental approaches and uses of data described above is shown in Figure 5. The implementation of such methods to study RA risk loci will contribute further evidence to the role of certain genes, pathways and cell types in RA as well as identifying new genes, pathways and cell types. It is expected that this increased understanding of disease aetiology will aid the stratification of patients for existing therapies as well as aiding in the discovery of new therapeutic targets, ultimately improving the prospects of patients with RA and other polygenic disorders.

ACKNOWLEDGMENTS

JD and GO are funded by the Wellcome Trust (Senior Research Fellowship reference number 207491/Z/17/Z) and by Arthritis Research UK (grant references numbers 21348 and 21754). Figures 2 and 4 were created through adaptation of Servier Medical Art images (https://smart.servier.com/), produced by Servier and licensed under a Creative Commons Attribution 3.0 Unported License (CC BY 3.0), allowing copying, adaptation and redistribution of content, with attribution (https://creativecommons.org/licenses/by/3.0/).

ORCID

James Ding https://orcid.org/0000-0001-7273-9646
Gisela Orozco https://orcid.org/0000-0002-3479-0448

REFERENCES

1. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545–1602.
2. Cutolo M, Kitas GD, van Riel P. Burden of disease in treated rheumatoid arthritis patients: going beyond the joint. Semin Arthritis Rheum. 2014;43(4):479-488.
3. Cross M, Smith E, Hoy D, et al. The global burden of rheumatoid arthritis: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis. 2014;73(7):1316-1322.
4. Winthrop KL, Strand V, van der Heijde D, et al. The unmet need in rheumatology: reports from the targeted therapies meeting 2017. Clin Immunol. 2018;186:87-93.
5. Barra L, Pope J, Bessette L, Harauoi B, Bykerk V. Lack of seroconversion of rheumatoid factor and anti-cyclic citrullinated peptide in patients with early inflammatory arthritis: a systematic literature review. Rheumatology. 2011;50(2):311-316.
6. van Venrooij WI, Zendman A, Pruijn G. Autoantibodies to citrullinated antigens in (early) rheumatoid arthritis. Autoimmun Rev. 2006;6(1):37-41.
7. Mclnnes IB, Schett G. The pathogenesis of rheumatoid arthritis. N Engl J Med. 2011;365(23):2205-2219.
8. MacGregor AJ, Snieder H, Rigby AS, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum. 2000;43(1):30-37.
9. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014;506(7488):376-381.

10. Eyre S, Bowes J, Diogo D, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet*. 2012;44(12):1336-1340.

11. Gibbs RA, Boerwinkle E, Doddapaneni H, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

12. Farh KK-H, Manson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343.

13. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-947.

14. Vahedi G, Kanno Y, Furumoto Y, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature*. 2015;520(7548):558-562.

15. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195.

16. Degner JF, Pai AA, Pique-Rege R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012;482(7385):390-394.

17. Boyle AP, Hong EL, Harirhan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.

18. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330.

19. Kawai H, Kasukawa T, Forrest A, Carminci P, Hayashizaki Y. The FANTOM5 collection, a data set series underpinning mammalian transcriptional atlases in diverse cell types. *Sci Data*. 2017;4:170113.

20. Stunnenberg HG, Hirst M, Almeida M, et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell*. 2016;167(5):1145-1149.

21. ENCODE Project Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306(5696):636-640.

22. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*. 2014;94(4):559-573.

23. Kichaev G, Yang W-Y, Lindstrom S, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*. 2014;10(10):e1004722.

24. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*. 2013;45(2):124-130.

25. Osplct C, Brentano F, Rengel Y, et al. Overexpression of toll-like receptors 3 and 4 in synovial tissue from patients with early rheumatoid arthritis: toll-like receptor expression in early and longstanding arthritis. *Arthritis Rheum*. 2008;58(12):3684-3692.

26. Karouzasikis E, Raza K, Kolling C, et al. Analysis of early changes in DNA methylation in synovial fi broblasts of RA patients before diagnosis. *Sci Rep*. 2018;8(1):7370.

27. Hueber W, Patel DD, Dryja T, et al. Effects of AIN457, a fully human antibody to interleukin-17A, on psoriasis, rheumatoid arthritis, and uveitis. *Sci Trans Med*. 2010;2(52):52ra72.

28. McNnes IB, Sieper J, Braun J, et al. Efficacy and safety of secukinumab, a fully human anti-interleukin-17A monoclonal antibody, in patients with moderate-to-severe psoriatic arthritis: a 24-week, randomised, double-blind, placebo-controlled, phase II proof-of-concept trial. *Ann Rheum Dis*. 2014;73(2):349-356.

29. Sieper J, Deodhar A, Marzo-Ortega H, et al. Secukinumab efficacy in anti-TNF-naïve and anti-TNF-experienced subjects with active ankylosing spondylitis: results from the MEASURE 2 Study. *Ann Rheum Dis*. 2017;76(3):571-592.

30. Kavalkowska J, Quirk A-M, Ghar F, et al. Abrogation of collagen-induced arthritis by a peptidyl arginine deiminase inhibitor is associated with modulation of T cell-mediated immune responses. *Sci Rep*. 2016;6(1):26430.

31. Subramanian V, Knight JS, Parekar S, et al. Design, synthesis, and biological evaluation of tetrazole analogs of Cl-aminidine as protein arginine deiminase inhibitors. *J Med Chem*. 2015;58(3):1337-1344.

32. Lee EB, Fleischmann R, Hall S, et al. Tofacitinib versus methotrexate in rheumatoid arthritis. *N Engl J Med*. 2014;370(25):2377-2386.

33. Davison LJ, Wallace C, Cooper JD, et al. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum Mol Genet*. 2012;21(2):322-333.

34. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*. 2013;45(10):1238-1243.

35. Gamanaz ER, Segrè AV, de van Bunt M, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease and trait-associated variation. *Nat Genet*. 2018;50(7):956-967.

36.-de Jager PL, Hacohen N, Mathis D, Regev A, Benoist C. ImmVar project: insights and design considerations for future studies of “healthy” immune variation. *Semin Immunol*. 2015;27(1):51-57.

37. Ye CJ, Feng T, Kwon H-K, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*. 2014;345(6202):1254665.

38. Fairfax BP, Humburg P, Makino S, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*. 2014;343(6175):1246949.

39. van der Wijst M, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet*. 2018;50(4):493-497.

40. Thalayasingam N, Nair N, Skelton AJ, et al. CD4+ and B lymphocyte expression quantitative traits at rheumatoid arthritis risk loci in patients with untreated early arthritis. *Arthritis Rheumatol*. 2018;70(3):361-370.

41. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol*. 2015;16(4):245-257.

42. Tolhuis B, Palstra R, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol Cell*. 2002;10(6):1453-1465.

43. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev*. 2012;26(1):11-24.

44. Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*. 2014;157(1):13-25.

45. Dillon N, Trimborn T, Stroubloulis J, Fraser P, Grosveld F. The effect of distance on long-range chromatin interactions. *Mol Cell*. 1997;1(1):131-139.

46. Wijgerde M, Grosveld F, Fraser P. Transcription complex stability and chromatin dynamics in vivo. *Nature*. 1995;377(6546):209-213.

47. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. 2002;295(5558):1306-1311.

48. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293.
Ding J, Orozco G. Identification of rheumatoid arthritis causal genes using functional genomics. Scand J Immunol. 2019;89:e12753. https://doi.org/10.1111/sji.12753