ARTHRITIS & RHEUMATOLOGY
Vol. 67, No. 1, January 2015, pp 1–10
DOI 10.1002/art.38800

Arthritis & Rheumatology
An Official Journal of the American College of Rheumatology
www.arthritisrheum.org and wileyonlinelibrary.com

DISEASE MECHANISMS IN RHEUMATOLOGY—TOOLS AND PATHWAYS

Defining Functional Genetic Variants in Autoimmune Diseases

Shaofeng Wang, Graham B. Wiley, Jennifer A. Kelly, and Patrick M. Gaffney

Introduction

Autoimmune diseases develop through the exposure of genetically susceptible hosts to environmental triggers. Advances in high-throughput genotyping and sequencing technologies coupled with comprehensive databases of human genetic variation and the assembly of large cohorts of case and control subjects have led to substantial progress in defining the genetic risk factors that underlie autoimmune diseases (1). The workhorse statistical methodology has been the genome-wide association study (GWAS). Studies using the GWAS approach have convincingly and reproducibly identified ~700 genomic regions in 183 published studies of autoimmune diseases at the level of genome-wide statistical significance ($P < 10^{-8}$) (www.genome.gov/gwastudies). The majority of these genetic associations are near genes that map to critical immunoregulatory pathways, illuminating genetic effects that are shared across multiple autoimmune diseases and other genetic effects that are restricted to only a few (1).

GWAS studies detect most causal variants indirectly, by leveraging linkage disequilibrium (LD) throughout the human genome. Classically defined, LD is the nonrandom association of two or more loci, resulting in segments of the genome being inherited as haplotype "blocks." Knowledge of the allele at one variant predicts with high likelihood the alleles at the other variants on the same haplotype block (2). Though LD makes locus discovery by GWAS very efficient, because only one or two variants per haplotype block need to be genotyped in order to detect association, the high correlation of variants on associated haplotypes confounds the ability of genetic association methods to distinguish causal from noncausal variants. The *UBE2L3* locus associated with multiple autoimmune diseases, including systemic lupus erythematosus (SLE) (3), illustrates the problem vividly, where 34 SLE-associated variants are located within a 67-kb haplotype, but we assume that only a few are likely to be causal (3). With a situation such as this, which variants are we to choose for functional screening? No specific guidelines exist on how to answer this question.

In general, the approach to overcoming the LD problem is to first comprehensively understand the genetic architecture at a given locus. Doing this in multiple ethnic populations when possible, adds significant power to our ability to discern causal variants by allowing the comparison of haplotypes across populations. Comprehensive characterization of a locus may include the following activities: 1) locus enrichment—capture for analysis all available genetic variation present on the risk haplotype, 2) locus refinement—winnow down the associated variants within a locus to a prioritized list for functional testing, and 3) functional testing—identify allele-specific differences in biologic function that support the variant's role(s) in causality. In

Shaofeng Wang, MD, PhD, Graham B. Wiley, PhD, Jennifer A. Kelly, MPH, Patrick M. Gaffney, MD: Oklahoma Medical Research Foundation, Oklahoma City.

Address correspondence to Patrick M. Gaffney, MD, Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, 825 NE 13th Street, Oklahoma City, OK 73104. E-mail: gaffneyp@omrf.org.

Submitted for publication July 18, 2013; accepted in revised form July 22, 2014.

this review, we discuss each of these steps and describe in more detail the available molecular methods that can provide the functional evidence required to assign causality to variants associated with autoimmune diseases.

Locus enrichment approaches

The objective of locus enrichment is to include in the analyses as much genomic variation as possible so that the odds of capturing the effects of all potential causal variants can be assured. This approach primarily includes the application of genetic fine mapping, resequencing, and imputation.

Large-scale fine mapping is most efficient in collaborative experiments, as demonstrated in the Large Lupus Association Study 2 (LLAS2) and the ImmunoChip (IC) array. The LLAS2 fine mapped and replicated GWAS findings in SLE patients and tested for these associations in non-European SLE populations. Including non-European cohorts (subjects of African, East Asian, and native North American ancestry) facilitated the exploration of differences in LD structure across ethnic groups in what is known as transpopulation mapping (discussed below). In contrast to other consortium initiatives, the focus of LLAS2 was not to produce a single comprehensive article about the experimental results, but rather, to distribute locus-specific results to investigators with a particular interest in developing the data further. This effort led to the identification of multiple new SLE loci, including *IRF8*, *TREM39A*, and *IKZF3-ZBP2* (4).

The IC array was developed in 2010 by a consortium led by the Welcome Trust Case-Control Consortium, with the primary objective of providing a cost-effective platform for fine mapping loci identified in GWAS across multiple autoimmune diseases. Built on the Illumina Infinium platform, the IC contained ~196,500 single-nucleotide polymorphisms (SNPs) selected to provide both fine mapping of existing GWAS loci from patients with autoimmune and autoinflammatory diseases and to allow for deep replication of those traits. The IC content included 186 associated regions identified in 12 autoimmune diseases, consisting of SLE, type 1 diabetes mellitus, autoimmune thyroid disease, Crohn's disease, ulcerative colitis, psoriasis, rheumatoid arthritis, primary biliary cirrhosis, multiple sclerosis, celiac disease, IgA deficiency, and ankylosing spondylitis. The success of the IC has been driven in part by its low cost; at $39 per sample, it is far less expensive than other GWAS arrays. Through the use of IC, multiple putative causal variants have been identified in Parkinson's disease, celiac disease, psoriasis, and ankylosing spondylitis (5).

Further locus enrichment with variants not directly genotyped is possible using the process of imputation, which requires a genotyped data set and a reference panel of variants configured as phased haplotypes. Public sequencing efforts, such as the HapMap Project (6) and the 1000 Genomes Project (7), have been instrumental in the development of reference panels of genetic variants and have greatly facilitated the utility of imputation, which is now a standard component of genetic analyses. Imputation programs leverage the LD between variants to populate the more sparsely genotyped data set with variants from the reference panel. For common variants, this can be done with high accuracy, thus increasing the chances of including causal variants in the analyses of association.

Several programs are available for imputing genotype data, including IMPUTE 2 (8), MaCH (9), and Beagle (10). The performance of each method is influenced by multiple factors, including ancestry (population substructure), genotyping platform (selection of SNPs), and the imputation reference panel. Overall, IMPUTE 2 has demonstrated superior imputation accuracy in most comparative studies (8,11). For quality of predictions, IMPUTE 2 and MaCH demonstrate lower error rates, from 5.16% and 5.46%, than Beagle, which is ~6.33% (12). Importantly, the imputation error rate increases as the minor allele frequency decreases due to reduced LD of rare variants with surrounding common variants, thus limiting the use of imputation for studies focusing on rare variants.

In addition, the HLA region, with its great functional diversity, is highly polymorphic and demonstrates functional importance in autoimmunity. Imputation methods have therefore been developed specifically for predicting the 4-digit HLA alleles based on the extended haplotype structure within the MHC region. These methods include HIBAG (13), HLA*IMP (14), and HLA*IMP:02 (14). The prediction accuracies of these methods for HLA alleles are comparable, with >90% for most loci. However, HIBAG has some advantages because it can be applied without the need to have access to large training data sets (unlike Beagle) or to upload data to an external website (unlike HLA*IMP and HLA*IMP:02) (14).

Next-generation sequencing (NGS) platforms have ushered in new possibilities for locus enrichment. Using NGS, entire risk haplotypes, loci, or even entire genomes can be sequenced, allowing for the identification of all possible variants. There are 3 main types of

NGS sequencing for genetic analysis: whole genome, whole exome, and targeted. Whole genome sequencing has the capability of providing researchers the identity of every variant contained within the genome; however, it may still be too expensive (~$1,500–2,000 per sample for sequencing only, not including data analysis) for most large-scale studies and may generate more data than required. Ongoing improvements in sequencing chemistry and hardware are rapidly dropping the price of whole genome sequencing to the point of making it economical for some large-scale studies, but the volume of data generated from whole genome sequencing may still be challenging for most laboratories to manage.

Whole exome sequencing isolates only the exonic sequences of the genome, allowing their specific capture and enrichment from noncoding sequences. The main limitation of exome sequencing is that it does not capture loci found in intergenic regions, such as long noncoding RNA or regulatory transcription factor binding sites, unless such sites are within LD blocks containing exon variants. Given that the majority (~90%) of disease-associated or trait-associated variants are located in noncoding DNA, this method is used primarily to assess the role of rare coding variants in autoimmune disease susceptibility (15).

Among these NGS methods, targeted resequencing has become the approach most frequently used to capture candidate causal variants within an established region of interest at manageable cost for large sample sizes. There are two primary methods for targeted resequencing: hybridization and amplicon. Hybridization uses oligomer "baits" tiled across regions of interest to enrich for complementary DNA sequences in the targeted region. Amplicon methods use primer walking to generate large amounts of polymerase chain reaction (PCR) products across the targeted locus in each sample. The resulting PCR products are then pooled and sequenced together. Hybridization approaches for targeted capture are generally preferred when targeting a large number of loci and have the capability to query up to 20 Mbp of sequence. Amplicon approaches are generally used for very specific sets of regions of <1 Mbp in total length but allow for large numbers of samples to be pooled together for high-throughput screening. Recent studies in SLE demonstrate several loci for which targeted resequencing successfully enriched functional variants and haplotypes, including *TNFAIP3* (16), *UBE2L3* (3), *IKBKE* (17), and *IFIH1* (17).

Locus refinement

Three primary activities are used to refine a locus following locus enrichment: conditional haplotype analysis, transpopulation mapping (TPM), and bioinformatic annotation. Conditional haplotype–based analyses are routinely used to evaluate for the presence of multiple independent association signals that could then indicate the presence of multiple causal variants within a given locus. To perform these analyses, stepwise logistic regression analysis is applied to data after adjusting for the most significantly associated variant. If variants in the locus continue to demonstrate evidence of association after adjustment, association can then be attributed to the presence of a second independent genetic effect. The process is usually repeated with the next most significant variant and so forth until all remaining association in the locus can be accounted for. Genetic analysis programs such as Plink (18) routinely implement this type of analysis.

The conditional haplotype approach was recently applied to an association between SLE and the HLA region. Using an alternative to stepwise logistic regression that allows parallel testing of different genetic models through Bayesian inference, Morris et al (19) identified 2 independent variants that were distinct from the classic HLA alleles, one in the class III region and the other in the class I region. The results suggest that there may be at least 5 independent genetic effects emanating from the HLA region that influence SLE predisposition.

The TPM approach to locus refinement requires that the locus of interest be genotyped in multiple ethnic groups so that differences in the LD structure across the various populations can be leveraged to potentially eliminate segments of the associated haplotype from further consideration. TPM has not been widely applied in autoimmune diseases since most studies have concentrated on populations of European ethnicity. The LLAS2 study in SLE (described above) was really the first to facilitate the broad application of TPM in the process of locus refinement. To perform TPM, association is identified in an index population, and specific haplotype blocks are identified. Variants are then genotyped in other ethnic populations, and associations with disease and haplotype block structure are determined. The populations are then evaluated for differences in LD structure that result in the loss of association with SNPs from the index population. The region demonstrating preserved evidence of association is the one most likely to harbor the causal allele. This scenario

assumes that the same causal variant is producing the signal in all associated populations. This approach has been successful in identifying causal variants in *ITGAM* in SLE (20) and *HLA–DQA1* in type 1 diabetes mellitus (21), as well as in the refinement of the *IL21* locus in SLE (22).

A corollary to this approach is to compare the locus-specific haplotype structures in populations that demonstrate association with populations that demonstrate no association. Under this scenario, it is assumed that the causal variant is present only in the associated populations and is present at a frequency similar to the risk haplotype. An example of this is seen in the identification of a causal regulatory variant near *TNFAIP3* (16).

Once genetic approaches have refined associated regions to their smallest possible segments, bioinformatic annotation is used to determine which variants reside in regions enriched for biologic importance. This activity is relatively straightforward, with the currently available data sets generated by large publicly funded consortia (e.g., the 1000 Genomes Project [23], the Encyclopedia of DNA Elements [ENCODE] Project [24], the UK10K Project [http://www.uk10k.org/]) and a variety of bioinformatics tools designed to predict the possible impact of each variant (Table 1).

To predict whether a coding variant may alter the function of a protein, several tools have been developed, including Sorting Intolerant From Tolerant (SIFT) (25), Polymorphism Phenotyping v2 (PolyPhen-2) (26), likelihood ratio test (LRT) (27), and MutationTaster (28). Each of these tools uses unique algorithmic methodology in an attempt to predict the deleterious nature of detected SNPs in protein-coding sequences. Due to their respective assumptions, they can and often do disagree as to the nature of variant damage potential. While there has been some comparison between tools, it is far from exhaustive (29). The Database for Nonsynonymous SNPs' Functional Predictions (dbNSFP) (30) has recently been developed to collate these disparate results and provide a resource wherein a particular variant may be queried across multiple prediction programs. These approaches have been useful in identifying functional variants in risk loci for SLE including *TNFAIP3* (16) and *ITGAM* (20).

Most complex disease-associated variants are located in DNA that does not encode for proteins, giving rise to the possibility that these variants may influence gene expression, RNA splicing, and/or transcription of noncoding RNAs through multiple complex mechanisms. To assess whether a variant may influence the expression of a nearby gene, a variety of in silico databases are available that allow a gene expression test to be performed. These databases include Gene Expression Variation (Genevar) (31), the messenger RNA (mRNA) by SNP Browser (32), the seeQTL (a searchable human eQTL browser and database) (33), and the Genotype-Tissue Expression (GTEx) eQTL (34) (Table 1) and provide researchers with a visualization of mapping results through either software or web browser interfaces.

Genevar allows users to switch between public services and local data on the same interface. Default services at the Sanger Institute have included 2 data sets: lymphoblastoid cell lines (LCLs) from 8 HapMap3 populations and 3 cell types derived from the umbilical cords of 75 Geneva GenCord subjects. The mRNA by SNP Browser includes data generated from LCLs derived from 400 children. SeeQTL collected 14 human eQTL data sets (HapMap LCLs, human cortical samples, and monocytes) and reanalyzed them using their own pipeline, combining quality control, population stratification control, association testing, and false discovery rate (FDR) control. Although the above 3 resources include databases generated from various sets of individuals, they are primarily based on LCLs. The GTEx eQTL database was recently established, providing central resources to enable the systematic study of genetic variation and the regulation of gene expression in multiple reference human tissues (34). Importantly, the gene expression profiles in the database are generated using multiple types of primary cells, such as liver, brain, and skin. The scope of this resource is to include any available tissues and to extend molecular phenotyping to other readouts, such as DNA methylation (34).

Following the completion of these approaches, the researcher is left with a prioritized list of variants for which genetic methods have determined a high likelihood that the causal variants are included. The following section describes the molecular approaches that can then be used to determine which SNPs produce a functional phenotype on protein expression or activity.

Functional testing

Characterization of transcription factor binding. For variants located in genetic regulatory elements, it is often important to determine if they alter the binding of nuclear protein complexes that may contain transcription factors and chromatin remodeling factors. Electrophoretic mobility shift assay (EMSA), EMSA-supershift (SS), and/or chromatin immunoprecipitation–

Table 1. Progress from tag SNP to functional mechanism in autoimmune diseases

Analytical method	Specific techniques and tools
Statistical genetic approach to identifying risk loci*	
Fine mapping of disease-associated regions	ImmunoChip
	Target resequencing
	Imputation (reference panel from 1KG or HM3)
Refine association signals	Conditional analysis
	Haplotype and linkage disequilibrium structure assessment
	Transpopulation mapping
Bioinformatics tools for evaluating functional potential†	
Functional annotation	ANNOVAR (http://www.openbioinformatics.org/annovar/)
	HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php)
	GenomeRunner (http://sourceforge.net/projects/genomerunner/)
	SCAN (http://www.scandb.org)
Prediction of protein function/structure change by coding variants	PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/)
	SIFT (http://sift.jcvi.org/)
	MutationTaster (http://www.mutationtaster.org/)
	dbNSFP (https://sites.google.com/site/jpopgen/dbNSFP)
Comprehensive bioinformatics database	ENCODE database (http://genome.ucsc.edu/ENCODE/)
	Genomatix (http://www.genomatix.de/)
Expression quantitative trait loci analysis	Genevar (http://www.sanger.ac.uk/resources/software/genevar/)
	uchicago eQTL browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/)
	SNPExpress (http://compute1.lsrc.duke.edu/softwares/SNPExpress/)
	SCAN (http://www.scandb.org)
	mRNA by SNP browser (http://www.sph.umich.edu/csg/liang/asthma/)
	SeeQTL (http://www.bios.unc.edu/research/genomic_software/seeQTL/)
	GTEx eQTL (http://www.gtexportal.org/home/)
Evaluate functional impact of risk variants‡	
Influence on gene expression	qPCR, Western blotting, microarray, RNA sequencing
Alternative splicing	RNA sequencing
Characterize regulatory mechanisms of functional variants§	
Alteration of transcription factor binding	EMSA, EMSA-supershift, DNA pulldown/mass spectroscopy, ChIP-qPCR
Alternative splicing	Double reports assay of altered splicing
Promoter activity	Luciferase promoter activity assay, DNA methylation (sodium bisulfite modification–based technique), histone modification (ChIP-qPCR)
Posttranscription modification	RNA EMSA
mRNA stability	Reporter gene assay of mRNA stability
Enhancer activity	Luciferase enhancer activity assay
Epigenetic modification	BiSeq, ChIP-Seq
Long-range DNA looping	3C, 4C-Seq, 5C, ChIA-PET, Hi-C
Noncoding RNA	Northern blotting, qPCR, RNA EMSA
Determine functional consequences of risk variants in vivo¶	
Cellular system	Zinc-finger nucleases
	TALENs
	CRISPR/Cas9
Animal model	Transgenic, knockout, knockin

* 1KG = 1000 Genomes Project; HM3 = HapMap3.
† ANNOVAR = Functional Annotation of Genetic Variants; SCAN = SNP and CNV Annotation Database; PolyPhen-2 = Polymorphism Phenotyping v2; SIFT = Sorting Intolerant From Tolerant; dbNSFP = Database for Nonsynonymous SNPs' Functional Predictions; ENCODE = Encyclopedia of DNA Elements; Genevar = Gene Expression Variation; eQTL = expression quantitative trait loci; SNP = single-nucleotide polymorphism; GTEx = Genotype-Tissue Expression.
‡ qPCR = quantitative polymerase chain reaction.
§ EMSA = electrophoretic mobility shift assay; ChIP = chromatin immunoprecipitation; BiSeq = bisulfite sequencing; 3C = chromatin conformation capture; 4C-Seq = chromatin conformation capture–on-chip with sequencing; 5C = chromatin conformation capture carbon copy; ChIA-PET = chromatin interaction analysis using paired-end tag sequencing.
¶ TALENs = transcription activator–like effector nucleases; CRISPR = clustered regularly interspaced short palindromic repeat.

quantitative PCR (ChIP-qPCR) are assays that effectively demonstrate the loss or gain of these interactions.

EMSA is a classic technique that involves separation of free DNA from DNA–protein complexes based on differences in their electrophoretic mobility in polyacrylamide gels (35) (Figure 1A). It is the core technology underlying a wide range of qualitative and quantitative analyses for the characterization of how DNA

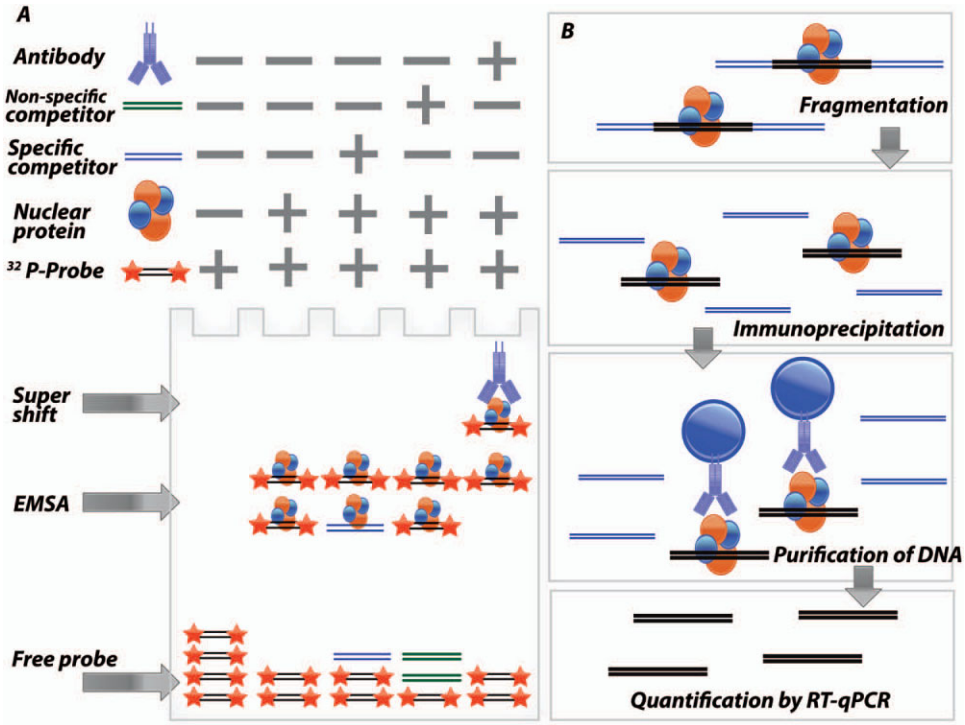

Figure 1. Characterization of the interaction between DNA and nuclear proteins. **A,** Electrophoretic mobility shift assay (EMSA) and supershift assay. Lanes contain (from left to right) free ^{32}P-probe, EMSA, EMSA in the presence of specific cold probe, EMSA in the presence of nonspecific probe, EMSA in the presence of antibodies against transcription factors. **B,** Chromatin immunoprecipitation–quantitative polymerase chain reaction (ChIP-qPCR) assay. Antibody against transcription factor is used to precipitate the crosslinked nuclear proteins and the target DNAs. Reverse transcription–qPCR (RT-qPCR) assay is used to determine the amount of precipitated DNA.

variants interact with DNA binding proteins (36). An antibody that recognizes the protein can be used in EMSA to identify a protein present in the protein–nucleic acid complex. This method is referred to as an EMSA-supershift assay. For characterization of candidate causal variants, differences in the binding affinity for protein complexes between risk and nonrisk alleles are supportive of a functional role. The reviews by Hellman and Fried (36) and by Carey and colleagues (37) comprehensively describe the technology and its strengths and weakness for correct interpretation of results.

EMSA and EMSA-SS assess the binding of transcription factors in vitro, independently of the native chromatin state. ChIP-qPCR assay is a widely used technique that complements EMSA by assessing the role of risk variants in influencing transcription factor binding in a native chromatin environment (Figure 1B). The method involves 2 basic steps: in vivo formaldehyde crosslinking of intact cells and selective immunoprecipitation of protein–DNA complexes with specific antibodies, followed by real-time qPCR measurement of the amount of DNA bound to the transcription factors.

The ChIP-qPCR procedure consists of several steps that can influence the final results; therefore, application of specific controls is important in the interpretation of the experimental data. To ensure the validity of the ChIP-qPCR signal, ChIP should be performed with 2 control antibodies: a negative control antibody (nonspecific isotype control) to evaluate background signal from nonspecific binding and a positive control antibody (e.g., anti–acetyl-histone H3) to be certain that the ChIP-qPCR assay was functioning properly. Once the qPCR data are obtained and normalized, they can be analyzed in different ways. The most commonly used methods are fold enrichment and percentage of input. With fold enrichment, the ChIP signals are divided by the antibody-free control signals, representing the ChIP signal as the fold increase in signal relative to background (38). In percentage of input, the qPCR signals derived from the ChIP samples are divided by the qPCR signals derived from the input sample taken early in the

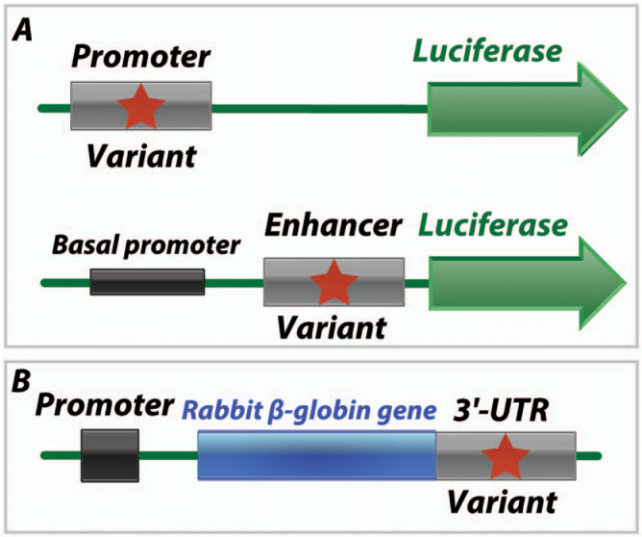

Figure 2. Characterization of functional variants using a luciferase reporter gene assay system. **A,** The promoter/enhancer activity assay system is based on a reporter gene encoding luciferase. Sequences of promoter or enhancer carrying candidate variants are cloned in front of the luciferase gene. **B,** Sequences of the 3′-untranslated region (3′-UTR) carrying candidate variants are cloned into a plasmid DNA after rabbit β-globin.

ChIP procedure (38). All normalization methods have their own advantages and drawbacks (38). Therefore, making a well-informed choice is important for correct interpretation of the data.

The EMSA, EMSA-SS, and ChIP-qPCR assays have been used in the identification of functional variants in risk loci for SLE, including *TNFAIP3* (16).

Transcriptional activity assays. Given that variants may alter transcription factor binding and modulate gene expression, assessment of their function in promoter activity can offer insight into their functional significance. Reporter gene assays are an efficient method for the indirect measurement of relative rates of transcription and make possible the fine mapping of transcriptional control regions in enhancer, promoter, or silencer regulatory elements (39). Using standard DNA cloning methods, the regulatory region carrying functional variants can be cloned upstream of a reporter gene, such as firefly luciferase (40) (Figure 2A). The expression of the reporter is then evaluated as a readout of the regulatory potential of the inserted regulatory element. Using a reporter assay, risk variants and their ability to influence reporter gene expression can be quantified.

As promoters and enhancers could function in a cell type–dependent manner, the testing of transcriptional activity of given DNA elements in the context of cell types is needed. This effort will not only help in the identification of functional variants, it will also provide valuable insight into the connection between functional variants and the pathogenesis of autoimmune diseases in the context of different organs and tissues. The reporter assay may also be designed to test the function of variants in the 3′-untranslated region, which are predicted to influence mRNA stability (41) (Figure 2B).

Chromatin conformation capture (3C). Recent attention to the 3-dimensional conformation of chromosomes in the nucleus suggests that long-range looping events, which bring into close proximity distant regulatory elements with gene promoters, is an important mechanism in the regulation of gene transcription (42). Measuring this dynamic aspect of the genome is particularly important for characterizing the effects of risk variants that are located in regulatory elements some distance away from the genes they regulate. The 3C methodology allows the in vivo genomic organization to be explored at a scale of a few tens to a few hundred kbps (42). To perform a 3C experiment, as shown in Figure 3, chromatin is fixed in vivo with formaldehyde to crosslink interacting sites. The chromatin is then digested with a restriction enzyme and ligated at a low DNA concentration to ensure that ligation between crosslinked fragments is favored over ligation between random fragments. Each ligation product reflects an interaction between 2 genomic loci and can be detected by qPCR using specific primers (Figure 3). The abundance of each ligation product provides a measurement of the frequency with which the 2 loci interact (42). Although 3C is experimentally straightforward, one must carefully design 3C experiments and implement the conscientious use of controls to draw meaningful conclusions. The reviews by Hagege et al (42) and by Dekker (43) comprehensively describe the controls for the correct interpretation of the results of 3C analysis.

In recent years, many robust approaches have been developed to allow researchers to efficiently investigate the role of chromosome conformation in gene regulation on a fully genome-wide scale. The methods include chromatin conformation capture–on-chip (4C) (44), 4C with sequencing (4C-Seq) (45), chromatin conformation capture carbon copy (5C) (46), Hi-C, the latest method based on 3C, (47), and chromatin interaction analysis using paired-end tag sequencing (ChIA-PET) (48). In addition, publically available data, such as 5C maps, ChIA-PET with anti–RNA polymerase II, and anti–CCCTC-binding factor (anti-CTCF) antibody re-

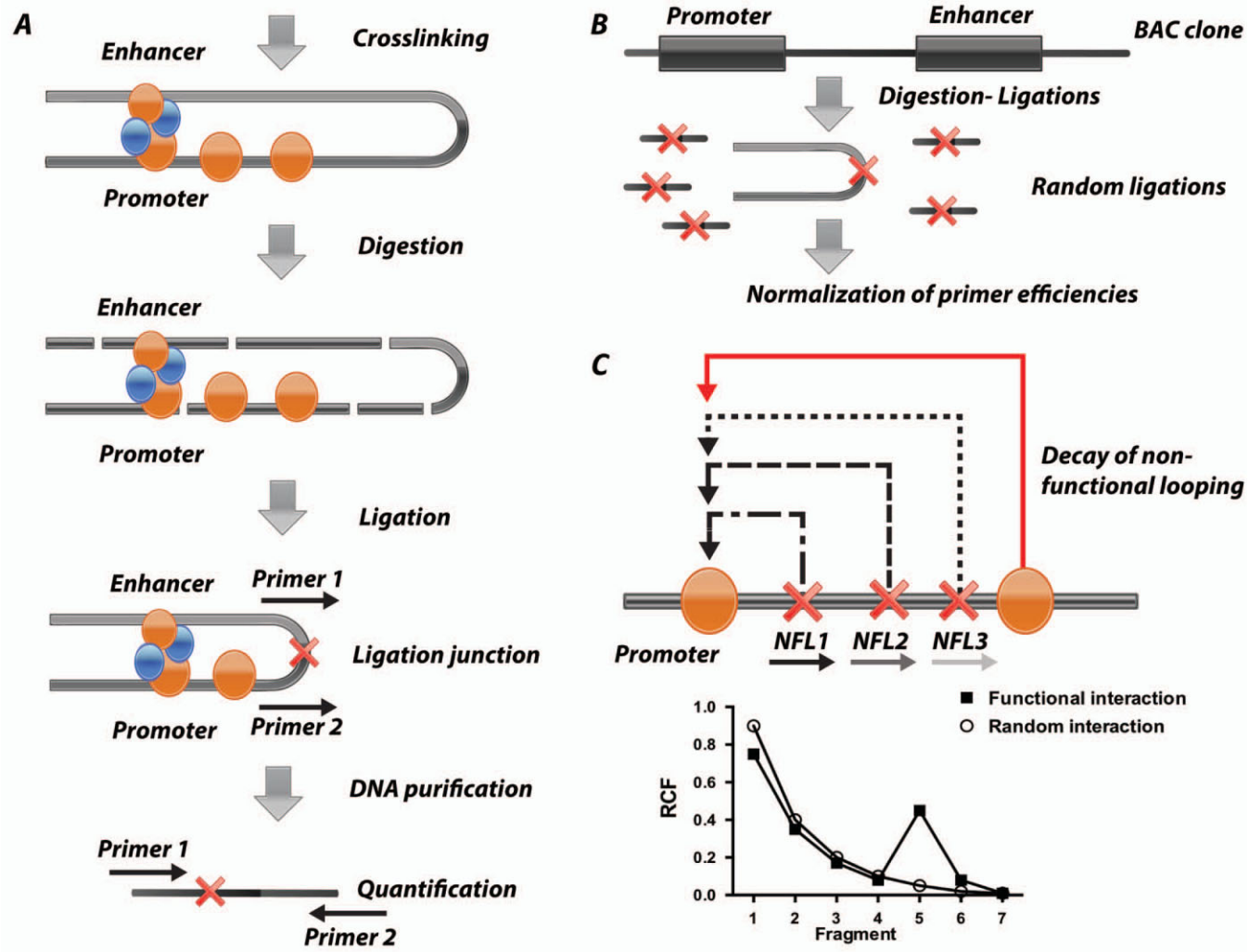

Figure 3. Detection of interacting genetic regulatory elements by chromatin conformation capture–quantitative polymerase chain reaction (3C-qPCR) assay. **A,** Interacting promoter and enhancer are crosslinked and digested with an appropriate restriction enzyme. Intramolecular ligation of crosslinked fragments is performed to generate chimeric DNA, the amount of which is then determined by reverse transcription–qPCR assay. **B,** Bacterial artificial chromosome (BAC) clone carrying the DNA region of interest is digested with the same enzyme used in **C,** followed by a random ligation to generate a positive control template containing all possible ligation products. **C,** The interaction frequencies for anchor/primer sets are shown by the **arrows**. **Arrows** marked NFL1–3 indicate nonfunctional looping. The strength of the interaction frequency is proportional to the weight of the **arrow**. Note that the interaction frequency is inversely related to the genome distance. Interaction resulting from functional looping is shown by the **orange arrow**. The line graph at the bottom shows the relative crosslinking frequency (RCF).

sults have been generated for GM12878, K562, and HeLa-S3 cells through the work of the ENCODE Project. These data are integrated with other data from the ENCODE Consortium, providing insight into the complex 3-dimensional interactions of the genome (http://genome.ucsc.edu/). As these data continue to accrue, they will serve to clarify a major impediment to the functional characterization of putative causal variants by defining which gene-enhancer interactions are most important in regulating gene expression in the context of specific cell lineages and activation states.

Functional characterization of causal variants using engineered cell lines. In recent years, new technologies have emerged that allow researchers to further investigate the role of causal variants through the introduction of specific variants into engineered cell lines. These technologies include zinc-finger nucleases (ZFNs), transcription activator–like effector nucleases

(TALENs), and clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 genome-editing tools. ZFNs and TALENs are artificial restriction enzymes generated through the fusion of a DNA-binding domain to a DNA-cleavage domain, while the CRISPR/Cas9 system uses short RNA–guided site-specific DNA nuclease technology. These engineered nucleases are capable of introducing double-strand DNA breaks (DSBs) in nearly any gene in a wide range of cell types. DSBs trigger the nonhomologous end–joining (NHEJ) DNA repair machinery, which can result in small insertions or deletions into the targeted site.

Recent studies have shown that ZFN-induced mutations are much more evenly distributed between deletions and insertions, while TALEN-induced deletions occur at a frequency 89% higher than that of insertions (1.6%) (49). Unlike these site-specific nucleases, target recognition by the Cas9 protein requires a "seed" sequence within the CRISPR RNA (crRNA) and a conserved dinucleotide-containing protospacer adjacent motif (PAM) sequence upstream of the crRNA-binding region (50) that may constrain some applications (50). Timely reviews by Kim et al (49) and Jinek (50) compare and contrast the various technologies for specific applications.

Introducing a donor DNA sequence with homology to the region being cut will stimulate incorporation of the donor sequence into the targeted site and the introduction of precise basepair changes into the genome. These approaches have opened up new possibilities for the functional study of genetic variants associated with autoimmune diseases by allowing researchers to study isolated causal variants on a genetic background free of the contribution of correlated variants carried on the same risk haplotype. These studies could be performed in human cell lines to complement and expand data derived from model organisms or in induced pluripotent stem cells to allow exploration of variant effects in tissue-specific contexts.

Conclusion

An important challenge for complex disease genetics is the identification and functional characterization of causal variants responsible for the association signals detected by GWAS. Accomplishing this task will require a multidisciplinary approach that includes genetic fine mapping, genomic sequencing, bioinformatics, and functional assays. Technological advances in genomic analyses have significantly enhanced the pace at which genetic data can be generated and annotated, highlighting the next bottleneck—functional assays. While EMSA, ChIP, luciferase assays, and 3C are widely applied and are mainstays in molecular genetic analysis, the development of high-throughput screening methods that can sensitively and accurately screen functional candidate variants would represent a significant technological breakthrough in this area.

We envision that as the field of genome engineering evolves, technologies that facilitate rapid and reliable genome editing, coupled with current and future high-throughput screening approaches could provide a path forward. Such advances would facilitate a more rapid approach to dissecting the confounding effects of linkage disequilibrium and speed translation of GWAS results into biologic understanding that can be applied to improve the lives of patients with autoimmune diseases.

ACKNOWLEDGMENTS

We thank He Li, Indra Adrianto, and John Ice for critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Lessard CJ, Ice JA, Adrianto I, Wiley GB, Kelly JA, Gaffney PM, et al. The genomics of autoimmune disease in the era of genome-wide association studies and beyond. Autoimmun Rev 2012;11:267–75.
2. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225–9.
3. Wang S, Adrianto I, Wiley GB, Lessard CJ, Kelly JA, Adler AJ, et al. A functional haplotype of UBE2L3 confers risk for systemic lupus erythematosus. Genes Immun 2012;13:380–7.
4. Lessard CJ, Adrianto I, Ice JA, Wiley GB, Kelly JA, Glenn SB, et al. Identification of IRF8, TMEM39A, and IKZF3-ZPBP2 as susceptibility loci for systemic lupus erythematosus in a large-scale multiracial replication study. Am J Hum Genet 2012;90:648–60.
5. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 2013;14:661–73.
6. International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
7. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
8. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.
9. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–34.
10. Browning SR. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 2006;78:903–13.

11. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511.
12. Hao K, Chudin E, McElwee J, Schadt EE. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet 2009;10:27.
13. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J 2014;14:192–200.
14. Dilthey A, Leslie S, Moutsianas L, Shen J, Cox C, Nelson MR, et al. Multi-population classical HLA type imputation. PLoS Comput Biol 2013;9:e1002877.
15. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337:1190–5.
16. Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat Genet 2011;43:253–8.
17. Wang C, Ahlford A, Laxman N, Nordmark G, Eloranta ML, Gunnarsson I, et al. Contribution of IKBKE and IFIH1 gene variants to SLE susceptibility. Genes Immun 2013;14:217–22.
18. Weeks JP. Plink: an R package for linking mixed-format tests using IRT-based methods. J Stat Softw 2010;35:1–33.
19. Morris DL, Taylor KE, Fernando MM, Nititham J, Alarcon-Riquelme ME, Barcellos LF, et al. Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. Am J Hum Genet 2012;91:778–93.
20. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, Gilkeson GS, et al. A nonsynonymous functional variant in integrin-α(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nat Genet 2008;40:152–4.
21. Todd JA, Mijovic C, Fletcher J, Jenkins D, Bradwell AR, Barnett AH. Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. Nature 1989;338:587–9.
22. Hughes T, Kim-Howard X, Kelly JA, Kaufman KM, Langefeld CD, Ziegler J, et al. Fine-mapping and transethnic genotyping establish IL2/IL21 genetic association with lupus and localize this genetic effect to IL21. Arthritis Rheum 2011;63:1689–97.
23. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
24. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, et al. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 2011;39:D871–5.
25. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations [letter]. Nat Methods 2010;7:248–9.
27. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553–61.
28. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations [letter]. Nat Methods 2010;7:575–6.
29. Castellana S, Mazza T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools [review]. Brief Bioinform 2013;14:448–59.
30. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
31. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics 2010;26:2474–6.
32. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, et al. A genome-wide association study of global gene expression. Nat Genet 2007;39:1202–7.
33. Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, et al. seeQTL: a searchable database for human eQTLs. Bioinformatics 2012;28:451–2.
34. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
35. Garner MM, Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 1981;9:3047–60.
36. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc 2007;2:1849–61.
37. Carey MF, Peterson CL, Smale ST. Electrophoretic mobility-shift assays. Cold Spring Harb Protoc 2013;2013:636–9.
38. Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant Methods 2007;3:11.
39. Himes SR, Shannon MF. Assays for transcriptional activity based on the luciferase reporter gene. Methods Mol Biol 2000;130:165–74.
40. Kricka LJ. Chemiluminescent and bioluminescent techniques. Clin Chem 1991;37:1472–81.
41. Tanguay RL, Gallie DR. Translational efficiency is regulated by the length of the 3′ untranslated region. Mol Cell Biol 1996;16:146–56.
42. Hagege H, Klous P, Braem C, Splinter E, Dekker J, Cathala G, et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc 2007;2:1722–33.
43. Dekker J. The three ‘C’s of chromosome conformation capture: controls, controls, controls. Nat Methods 2006;3:17–21.
44. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 2006;38:1348–54.
45. Van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods 2012;9:969–72.
46. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 2006;16:1299–309.
47. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 2012;58:268–76.
48. Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, et al. ChIA-PET analysis of transcriptional chromatin interactions. Methods 2012;58:289–99.
49. Kim Y, Kweon J, Kim JS. TALENs and ZFNs are associated with different mutation signatures [letter]. Nat Methods 2013;10:185.
50. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 2012;337:816–21.