Published in final edited form as:

Best Pract Res Clin Rheumatol. 2014 April ; 28(2): 175–189. doi:10.1016/j.berh.2014.05.001.

Genetics, genomics and their relevance to pathology and therapy

Michael J. Ombrello, M.D.,

National Institute of Arthritis, Musculoskeletal, and Skin Diseases, NIH, DHHS 9000 Rockville Pike, Building 10, Room 10C101A, MSC 1560 Bethesda, Maryland, 20892-1560, U.S.A.

Keith A. Sikora, M.D., and

National Institute of Arthritis, Musculoskeletal, and Skin Diseases, NIH, DHHS 9000 Rockville Pike, Building 10, Room 10-CRC/1-5256, MSC 1102 Bethesda, Maryland, 20892-1102, U.S.A.

Daniel L. Kastner, M.D., Ph.D.

National Human Genome Research Institute, NIH, DHHS 9000 Rockville Pike, Building 50, Room 5222, MSC 8002 Bethesda, Maryland, 20892-8002, U.S.A.

Michael J. Ombrello: michael.ombrello@nih.gov; Keith A. Sikora: keith.sikora@nih.gov; Daniel L. Kastner: daniel.kastner@nih.gov

Abstract

Genetic and genomic investigations are a starting point for the study of human disease, seeking to discover causative variants relevant to disease pathophysiology. Over the past 5 years, massively parallel, high-throughput, next-generation sequencing techniques have revolutionized genetics and genomics, identifying the causes of many mendelian diseases. The application of whole genome sequencing and whole exome sequencing to large populations has produced several publicly-available sequence datasets that have revealed the scope of human genetic variation, and have contributed to important methodological advances in the study of both common and rare genetic variants in genetically-complex diseases. The importance of noncoding genetic variation has been highlighted by the Encyclopedia of DNA Elements (ENCODE) project and NIH Roadmap Epigenomics Program, and integrated analyses of these datasets, together with disease-specific datasets, will provide an important and powerful tool for determining the mechanisms through which disease-associated, noncoding variation influence disease risk.

Keywords

pediatric rheumatology; somatic mosaicism; genomewide association study; targeted deep resequencing; miRNA; lncRNA

Correspondence to: Michael J. Ombrello, michael.ombrello@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest statement

In pediatric rheumatology, as in most of medicine, the causes of nearly every disease that we treat are either unknown or incompletely understood. Therefore, studies seeking to determine the pathogenic mechanisms of pediatric rheumatic diseases are of critical importance if we hope to identify and test new therapeutic agents. Genetic and genomic investigations are an integral component of the biomedical research endeavor, identifying variants that cause or influence the risk of developing a disease. In a field that is driven by technological advances, the emergence of massively parallel, high-throughput sequencing techniques, also known as **next-generation sequencing (NGS)**, has truly revolutionized clinical genomics with the largest immediate impact on **monogenic diseases**, which are caused by mutations of a single gene [1]. The wide adoption of NGS has produced large, publicly-available sequence datasets that have facilitated expanded use of single nucleotide polymorphism (SNP) imputation and ultimately improved the sensitivity of genomewide association studies (GWAS) in the investigation of **genetically-complex diseases** [2, 3]. Finally, efforts like the Encyclopedia of DNA Elements (ENCODE) project [4] and NIH Roadmap Epigenomics Program [5] have brought to the forefront the importance of non-coding variation. Moreover, the incorporation of these datasets into public repositories and genome browsers is now enabling the integrated interrogation of disease-specific genomic data and ENCODE data to determine the mechanisms through which disease-associated, noncoding variation influences disease risk. This is particularly important, considering that over 90% of GWAS-identified disease- and trait-associated variation falls within noncoding regions of the genome [6].

In this review, we will seek to demonstrate how advances in genetic and genomic technology may be applied to the investigation of pediatric rheumatic diseases. Given the paucity of examples in which these techniques have been applied to pediatric rheumatic diseases, to date, we will discuss these innovative approaches in the context of a range of rheumatic and immunologic diseases.

Investigating mendelian and monogenic pediatric rheumatic diseases

Methodologies

Prior to the advent of NGS technologies, investigations of diseases of **mendelian** inheritance typically employed linkage analyses of large families with multiple affected individuals to map the phenotypes of interest to specific genomic regions [7, 8]. By typing family members for a panel of markers evenly-spaced across the entire genome, one may identify the linked genomic regions where marker haplotypes segregate with disease status. From the linkage interval(s), candidate genes can be selected and sequenced by Sanger's method [9, 10] until the disease-causing mutation is identified. In Sanger sequencing, the genomic regions of interest are amplified by the polymerase chain reaction (PCR), and each strand of the double-stranded PCR product is further subjected to dye-terminator sequencing reactions. After separation and detection by a capillary sequencer, a chromatogram demonstrating the nucleotide sequence for each DNA fragment is examined. The evaluation of linkage intervals by conventional sequencing can be an expensive and time consuming process, particularly when linkage regions are large or contain many large, complex genes.

Compared to sequencing by Sanger's method, NGS approaches have many important improvements and benefits. NGS approaches can simultaneously interrogate tens to hundreds of billions of base pair positions in a single reaction, generating datasets whose breadth and depth of sequence coverage dwarf that of conventional sequencing methods. Additionally, NGS approaches are often less expensive and less time consuming than conventional sequencing. For all of these reasons, NGS has quickly replaced traditional sequencing in many applications. There are a variety of NGS platforms and techniques from which investigators may choose, but they each share a general work flow that includes the preparation of a library of DNA templates, the sequencing and imaging of the template library, and the post-sequencing data analysis. More specific information about NGS sequencing platforms and their methodologies may be found in these excellent reviews of the topic [11, 12].

DNA libraries for NGS sequencing may be generated through several approaches that are specific to the experimental design of the study. In whole-genome sequencing (WGS), one utilizes a library of DNA templates created from whole genomic DNA. In contrast, for **targeted deep resequencing** one extracts from the whole DNA only the regions of interest, producing a sequencing library from only those targeted regions, a process called target enrichment. Target enrichment may be accomplished through high-fidelity PCR amplification of the genomic regions of interest, in which case the sequencing libraries are created from the pooled PCR products. Alternatively, one may use hybrid capture-based methods for target enrichment, wherein a cocktail of custom-designed, synthetic oligonucleotide probes with sequences complementary to the genomic regions of interest is used to "capture" those desired regions. After additional processing, only the genomic regions of interest remain as the input material for the sequencing library. The most widely adopted target enrichment strategy is whole exome sequencing (WES), in which one seeks to capture and sequence the entirety of the coding regions of the genome, the so-called "exome", which accounts for only about 1% of the entire genome [13]. Due to the substantially lower computational effort and financial cost of performing WES, compared to WGS, combined with the expectation that a majority of mendelian diseases are caused by protein coding variants, WES is an attractive investigative strategy.

WES studies produce a very large set of sequence reads of various lengths and configurations, depending on the experimental design. These sequence reads are assembled and aligned to a known reference sequence and variant alleles are identified and cataloged. The resulting list of variants may then be filtered on the basis of various characteristics, including metrics of variant quality (depth of coverage and quality scores), indicators of potential importance of a variant (allele frequency, degree of evolutionary conservation, predicted effect on protein structure) and proper segregation within the experimental samples. The approaches and tools for the analysis and filtering of NGS datasets are rapidly evolving [14].

The first NGS discovery

The first instance of a NGS study identifying a novel cause of a human disease was in 2010, when WES (and later WGS) were independently applied to the investigation of postaxial

acrofacial dysostosis (PAD), a rare, autosomal recessive malformation syndrome marked by features that include cleft lip and palate, posterior limb abnormalities and ocular anomalies [15, 16]. In their first study, the authors performed WES of 4 PAD patients from three families, generating a list of all variants within the protein-coding region of these 4 individuals [15]. By filtering the variant lists to include only genes harboring two novel, protein-altering mutations in all 4 individuals, the authors reduced their list to a single candidate gene, DHODH, encoding dihydroorotate dehydrogenase, which coincidentally is familiar to pediatric rheumatologists as the target of the drug, leflunomide. Using conventional sequencing of DHODH, the authors identified compound heterozygous mutations in 4 additional affected individuals. The second study of PAD, which used WGS to examine a family quartet that included 2 affected children and 2 unaffected parents, also identified mutations of DHODH as its cause [16]. It was immediately predicted that NGS would accelerate the discovery of genetic causes of many rare mendelian diseases [17], and within its first year of adoption, WES was responsible for the identification of the genetic causes of 12 mendelian diseases [18].

Next-generation sequencing discoveries and the pediatric rheumatologist

Deficiency of adenosine deaminase 2—Of the many NGS-driven mendelian discoveries that have been reported since that time, a subset has identified causes of immune-mediated diseases that are directly relevant to pediatric rheumatologists. Among these are autoinflammatory diseases, which are generally characterized by recurrent episodes of inflammation that occur in the absence of either an identifiable trigger or features typical of autoimmunity, as well as disorders of immune dysregulation with predisposition to autoimmunity, immune deficiency, and atopy. Perhaps the most striking of the newly-identified, immune-mediated diseases is the deficiency of adenosine deaminase 2 (DADA2), an autosomal recessive, autoinflammatory syndrome that is characterized by the presence of systemic inflammation with intermittent fevers, the early-onset of recurrent lacunar strokes, some with hemorrhagic transformation, and vasculopathy [19].

WES analysis of 3 unrelated, affected individuals and their unaffected parents identified compound heterozygous mutations of CECR1, which encodes adenosine deaminase 2 (ADA2), in each subject. Importantly, it was the combined analysis of multiple affected individuals that produced this discovery. The separate sequence analyses in patients 1 and 2 produced lists of 17 and 19 candidate genes, respectively, whereas the combined analysis of these two individuals identified CECR1 as the only candidate gene. Compound heterozygous mutations of CECR1 were identified subsequently in 3 additional affected individuals using conventional sequencing methods, further helping to make the case that variants of this gene are really causative. Further, all affected subjects had markedly reduced levels of ADA2 and its enzymatic activity in peripheral blood, as compared to healthy individuals. The authors demonstrated that upon morpholino-based silencing of a CECR1 paralog in zebrafish, the fish embryos developed intracranial hemorrhage and neutropenia, and moreover this phenomenon was reversible with the co-introduction of wild-type, human CECR1, but not with the mutated forms of CECR1. These experiments demonstrate the general value of animal models in establishing a functional link between a gene of interest and a disease. Specifically, this study demonstrates how the unique features of zebrafish,

including their very rapid development, their translucent embryos, and the ease with which they may be genetically manipulated, may make them an appealing alternative to other model organisms for the rapid investigation of genetic discoveries. This study went on to identify homozygous mutations of CECR1 in three patients with polyarteritis nodosa (PAN), which together with a concurrently-published study that also identified recessively-inherited mutations of CECR1 in 6 families with PAN, further expanded the scope of ADA2-mediated disease [20].

Chronic multifocal osteomyelitis with immune dysregulation—A very important benefit of NGS in the investigation of mendelian disease is its utility, even in cases of single affected individuals and family units too small for linkage analysis. This ability was demonstrated in a study that examined a child with immune dysregulation characterized by vitiligo, autoimmune hemolytic anemia, B lymphopenia and sterile chronic multifocal osteomyelitis which evolved into disseminated granulomatous disease [21]. Using WES of the affected child and his two unaffected parents, investigators identified 12 genes that contained homozygous, non-synonymous mutations in the affected child that were not present in either parent. Among these was RAG1, which encodes the recombinase-activating gene 1 and has been associated with immune dysregulation similar to that seen in the affected child. The presence of T and B lymphocytes in the affected child suggested that the missense mutation, R699W, produced reduced, but not absent, function, given that complete deficiency of RAG results T and cell deficiency,

Deficiency of HOIL-1 protein—Another example is the case of two siblings and a third unrelated child with a unique combination of autoinflammation and life-threatening immune deficiency, the cause of which was identified by a combination of SNP genotyping and WES. Early in life, all three affected children developed recurrent, episodic systemic inflammation, as well as antibody deficiency, recurrent pyogenic infections and the intracellular deposition of amylopectin-like material in the skeletal, smooth and cardiac myocytes. Interestingly, while bone marrow transplantation brought resolution to the infectious and inflammatory manifestations of one patient, that patient continued to experience progression of the amylopectinosis, ultimately succumbing to complications of heart failure. All three subjects were found to have recessively-inherited, loss-of-function mutations in RBCK1, which encodes the HOIL-1 protein, a component of the linear ubiquitin chain assembly complex (LUBAC) [22]. Through the addition of linear chains of ubiquitin, the LUBAC is critically involved in the targeting of proteins to specific cellular destinations, such as the targeting of nuclear factor kappa B (NF- κ B) essential modulator to the interleukin (IL)-1 receptor (IL-1R), tumor necrosis factor receptor (TNFR), and toll-like receptor (TLR) signalling complexes [23]. The authors went on to demonstrate that in fibroblasts and Epstein-Barr Virus-transformed B cells from these patients, the loss of HOIL-1 impaired assembly of the LUBAC assembly, leading to a disruption in NF- κ B activation.

Immune dysregulation mediated by mutations in PLCG2—NGS was also involved in the discoveries of two diseases caused by mutations of PLCG2, highlighting both the power and the limitations of NGS approaches. The cause of PLCG2-associated immune

deficiency, antibody deficiency and immune dysregulation (PLAID), an autosomal dominant syndrome characterized by cold urticaria and susceptibility to atopy, autoimmunity and infection, was not identified by an NGS approach, but rather was identified using SNP-based linkage analysis coupled with Sanger sequencing [8]. Through this approach, three distinct exon-containing deletions of *PLCG2* were identified as the cause of this syndrome. Importantly, WGS was undertaken in the first proband, however it failed to identify the causative mutation because the software to detect deletions hadn't been tested and tuned to detect hemizygous deletions of this size. While the reasons for them may vary, such failures are not uncommon, and in fact a recent report from a large, academic medical center found that among the first 250 consecutive patients referred for clinical WES to evaluate possible genetic conditions, WES failed to identify the causative variant in 75% of cases [1]. From a positive perspective, this reflects the successful identification of causal genetic variants in 1 out of 4 cases studies by WES, which is an extraordinary rate of success!

In contrast to PLAID, the cause of autoinflammatory PLAID (APLAID), which is characterized by antibody deficiency with recurrent infections, together with inflammatory manifestations of the skin, eyes, and gastrointestinal tract, was successfully identified using WES of a family trio that included an affected father and daughter, and an unaffected mother [24]. Using a variant filtering strategy that identified only high-quality sequence variants with >10X coverage that were non-synonymous, novel, evolutionarily conserved and predicted to be damaging, a list of 8 candidate variants was generated. Conventional sequencing of these 8 variants in the unaffected grandparents found that one of these mutations, the S707Y variant of *PLCG2*, was a de novo mutation in the affected father.

Detecting somatic mutations that cause monogenic diseases—Somatic mutations, or mutations arising de novo in somatic cells, may be an important and under-recognized cause of disease. When a somatic mutation arises in an individual's gametes, that mutation may be transmitted, de novo, to their offspring, in whom the mutation would be present in the germ line and could be further transmitted. In contrast, when a somatic mutation occurs in somatic cells other than gametes, it produces somatic mosaicism with two genetically-distinct populations of cells. Because conventional sequencing cannot sensitively identify somatic mosaicism, there have been relatively few studies exploring the spectrum of nonmalignant diseases caused by somatic mutations. Given the enormous depth of coverage that NGS approaches can generate at each base position, it may now be possible to identify disease-causing somatic mutations, even if mutant cells represent a very low percentage of the larger cellular population. Although WES has been successfully used to identify causative somatic mutations in a number of diseases, including the cryopyrinopathies [25–28], the error rate of the whole genome amplification step of the most widely-adopted sequencing platform is relatively high at 1% [12, 29], which may confound the search for somatic mutations occurring at frequencies less than 10%. Studies of cancer genomics have demonstrated that this problem may be partially overcome by evaluating paired affected and unaffected tissues, and this approach has been widely applied, identifying an enormous range of cancer-associated somatic mutations [30]. By applying a similar strategy to the investigation of Proteus syndrome [31], a disease marked by segmental overgrowth and tissue hyperplasia that is thought to be the “Elephant Man”

disease, investigators identified the cause to be mosaic, activating mutations of AKT1. By performing WES of genomic DNA from 7 affected and 4 unaffected tissue biopsy specimens acquired from 6 Proteus syndrome patients, together with genomic DNA from 6 of their unaffected first degree relatives, investigators applied a filtering strategy designed to identify novel, protein-altering variants present in the affected, but not the unaffected tissues. After identifying the first mutation of AKT1 in a single patient, manual inspection and follow-up studies ultimately revealed that 26 of 29 Proteus syndrome patients had AKT1 mutations.

In addition to employing study designs that examine paired affected and unaffected tissues, there are several other methods that promise to improve the fidelity of NGS studies. By allowing for the identification and exclusion of variants created by PCR or sequencing errors from subsequent analyses, these methods greatly increase the sensitivity of NGS to detect ultra rare variants. In duplex sequencing, this is accomplished by incorporating oligonucleotide tags onto the 5' ends of each strand of duplex DNA templates [29]. Upon analysis of sequencing reads, if a variant is detected on tagged sequence reads from both strands, then one can conclude that the variant was present in the original sample. In contrast, mutations created by PCR or sequencing errors are introduced on a single DNA strand, and therefore variants identified on reads from only one DNA strand are excluded. In circle sequencing, DNA templates are circularized, making it possible to use a rolling circle polymerase to generate multiple copies of the circularized template DNA in tandem. By creating physically linked tandem copies, one can be confident that variants multiply detected in the tandem reads were each generated as independent copies of the circularized template, while variants not detected in tandem may be attributed to experimentally introduced errors [32].

Investigating pediatric rheumatic diseases with genetically-complex inheritance

In pediatric rheumatology, many of our most commonly encountered diseases, including the various forms of juvenile idiopathic arthritis (JIA), juvenile systemic lupus erythematosus (JSLE), the juvenile inflammatory myositides and the chronic regional pain syndromes (ie. juvenile fibromyalgia), are genetically-complex diseases. Instead of resulting from mutations of a single gene, as is the case in diseases of mendelian inheritance, genetically-complex diseases result from the combined effects of one or more genetic risk factors and one or more environmental risk factors. The risk of genetically-complex diseases may be influenced by common genetic variants, which typically have relatively small effect sizes on disease risk. Rare genetic variants also contribute to the risk of genetically-complex diseases, as has been demonstrated in type 1 diabetes (T1D) [33], rheumatoid arthritis (RA) [34] and Behçet's disease (BD) [35], among others. Rare variants often have higher penetrance and confer larger effects on an individual's risk of disease than do common variants. However because they are selected against, they remain rare and therefore make relatively small contributions to the disease risk in populations.

The gold-standard approach to investigate genetically-complex diseases is the association study, in which one compares the frequency of genetic variants between ancestrally similar

groups of disease-affected individuals and healthy control subjects. In such a study, if the frequency of a genetic variant differs significantly between the two groups, then the variant is said to be disease-associated and further studies are warranted to determine whether the disease-associated variant is the disease-causal variant, or if its association with the disease simply reflects its strong linkage disequilibrium (LD) with the actual causal variant. Although some candidate gene studies were informed by intervals defined by linkage mapping, many candidate gene studies relied upon best guesses and luck.

With the emergence of commercially-available methods to interrogate the entire genome for associations in a high-throughput manner, the GWAS quickly supplanted the candidate gene study and became the preferred method of identifying susceptibility loci in genetically-complex diseases. A GWAS is a hypothesis-neutral, unbiased study design that examines common genetic markers spanning the entire genome to identify disease susceptibility loci [36]. Using arrays of SNPs that are designed to query most of the common, independently-inherited LD blocks, GWAS can identify LD blocks harboring common, disease-associated variants. GWAS do not often identify the functionally important risk variant(s) within the LD block, though, and moreover, rare disease-associated variants may also exist within GWAS-implicated genes. For both of these reasons, GWAS are often coupled with candidate gene studies to more intensively interrogate the GWAS-implicated susceptibility loci to identify a functionally-relevant variant. A potential pitfall of these follow-up studies is the tendency to concentrate on the gene nearest to the variant, while ignoring the possibility of more long-range effects.

Methodological advances in the investigation of genetically-complex diseases

There have been a number of important advances in the performance of GWAS and analysis of genomewide datasets that warrant further discussion. First, the development and maturation of SNP imputation as a genetic statistical tool has made it a common element of a contemporary GWAS [37, 38]. SNP imputation is a statistical method that can, through repetitive comparisons of an experimental SNP dataset with a reference panel of dense SNP genotypes, accurately determine additional SNP genotypes in the experimental population, in silico. The maturation of SNP imputation was enabled both by the availability of large, sequencing-based reference populations, such as the 1000 Genomes project [2], as well as through methodological improvements to accommodate the substantially larger size of sequencing-based reference populations [3, 39, 40]. Ultimately, multi-reference imputation was pioneered with improved performance of imputation across experimental populations [2, 3]. The potential power and value of using imputation in GWAS is demonstrated by two recent studies of a single BD case-control collection. In the initial GWAS of this population, which examined 311,459 SNPs, significant associations were identified between BD and IL10, HLA-B and IL23R-IL12RB2 [41]. In the second study of the same case-control collection, SNP imputation was used to expand the dataset to include 779,465 SNPs, ultimately identifying independent associations between BD and CCR1, STAT4, KLRC4 and ERAP1 that were not detected in the first GWAS [42].

In addition to expanding the density of SNP datasets, imputation-based approaches may also be used to infer classical major histocompatibility complex (MHC) alleles. Many GWAS,

particularly those of autoimmune and immune-related diseases, have identified disease associations among the MHC class I or class II gene clusters. Recently, a technique has evolved that allows one to statistically determine the MHC class I and class II types, as well as the subordinate amino acid identities, from SNP genotypes derived from the MHC region of nearly any commercially available GWAS array [43, 44]. By applying this approach to RA, which has a strong association with human leukocyte antigen (HLA)-DRB1 allele [45], the risk of RA was mapped to three amino acid positions of HLA-DRB1. This work further refined our understanding of the shared epitope of RA, while also identifying associations between BD and single residues of both HLA-B and HLA-DPB1, neither of which were previously known to influence RA risk [46]. In addition to RA, this method was recently used to investigate the role of MHC class-I molecules in BD, where it identified a group of 7 amino acid residues of HLA-B and HLA-A that independently influence disease susceptibility [47]. Moreover, the locations of the BD-associated residues implicate both peptide binding by both MHC class-I molecules and the regulation of cytotoxic cells through two distinct families of cytotoxicity receptors in the pathogenesis of BD. Given that numerous pediatric rheumatic diseases have associations with specific MHC molecules, this technique may be especially informative in the investigation of these diseases.

Particularly relevant to pediatric rheumatology, genomewide meta-analysis has been developed and adopted to bolster the power of GWAS in some of the most heavily studied diseases, such as RA [48, 49], where the latest genomewide meta-analysis includes approximately 10 million SNPs examined in 29,880 patients and 73,758 healthy controls of European and Asian ancestry [50]. Building on the 59 known RA susceptibility loci identified by previous studies, Okada *et al.* identified 42 novel, significant RA susceptibility loci [50]. Genomewide meta-analysis may also be utilized in the investigation of rare, genetically-complex diseases, such as systemic JIA, enthesitis-related JIA and psoriatic JIA, where genomewide meta-analysis may be required to even assemble a case collection of adequate size to perform the GWAS.

Low-frequency and rare genetic variants in genetically-complex diseases

Low-frequency variants (with frequencies below 0.05) and rare genetic variants (with frequencies below 0.01) are another important source of risk in genetically complex diseases. In fact, because rare variants were not included in commercially-available GWAS SNP arrays, it was not until the release of WGS and WES data from reference populations like the 1000 Genomes Project and the National Heart Lung and Blood Institute's Exome Sequencing Project that the landscape of rare genetic variation was truly appreciated [51]. Today, rare variants may be interrogated in three general ways: using NGS methods, including targeted deep resequencing, WES and WGS; using special genotyping arrays designed to examine rare variants, such as the ImmunoChip [52], a SNP array that densely genotypes SNPs within a subset of immunologically important genes; and using SNP imputation with 1000 Genomes reference data to infer the identities of rare variants. Importantly, the statistical methods for evaluating rare variants differ from those used to analyze common variants. Single point analyses, like those used to examine common variants, are under powered because the number of occurrences of rare alleles would require a prohibitively large study population. Instead, analyses that consider the distribution of rare

variants in an entire gene, referred to as gene-based tests or burden tests, are widely utilized to test for disease associations with rare variants [54]. These tests may be used to compare the distribution or “burden” of rare variants between two groups. For further information, please see the topical review by Panoutsopoulou *et al.* [54]

Targeted deep resequencing studies are the NGS version of the candidate gene study, allowing multiplex interrogation of all candidate genes or regions of interest. In one of the pioneering targeted resequencing studies, Nejentsev *et al.* examined 10 candidate loci in 480 type 1 diabetes (T1D) patients and 480 healthy control subjects, identifying rare variants of *IFIH1*, encoding melanoma differentiation-associated protein 5, which was located in a region previously implicated by GWAS of T1D. This study demonstrated that resequencing could clarify the source of risk in genomic regions initially identified by GWAS [33]. Using a similar approach, Kirino *et al.* examined 10 candidate genes implicated by GWAS and 11 genes of the innate immune system by deep resequencing in a collection of 2461 BD cases and 2458 healthy control subjects. Using three different tests of rare variant association, this study identified significant enrichment of rare variants of *TLR4*, *IL23R* and *NOD2* in BD, as well as an association between BD and the common M694V variant of *MEFV*, implicating innate immune mechanisms in the pathogenesis of BD [35]. Targeted deep resequencing studies of GWAS implicated risk loci have also identified rare variant associations in many other diseases, including RA and inflammatory bowel disease [34, 55].

Looking forward to genetically-complex pediatric rheumatic diseases

As these examples demonstrated, the coupling of GWAS and targeted, deep resequencing of known susceptibility loci is a strong and intuitive approach to harnessing the capabilities of NGS, while limiting the financial cost and computational time that may be incurred through WGS or WES of the GWAS and replication populations. It is also foreseeable that in the near future, the investigation of genetically-complex diseases will transition away from SNP genotyping arrays, instead employing WGS or WES to create population-based datasets. This choice would produce complete datasets appropriate for the thorough investigation of the full range of genetic variation from a single experimental procedure.

With respect to pediatric rheumatic diseases, the approaches discussed above hold great promise for advancing our understanding of diseases like JIA and JSLE, with the obvious challenge of sample recruitment. Although GWAS of common variation generally require sample sizes of at least 1000 affected subjects to detect effects of 10–20% increase in disease risk, studies in diseases like RA have continued to identify novel susceptibility loci, albeit most with smaller contributions to risk than those previously identified, as the size of study populations have grown [48–50]. Because most pediatric rheumatic diseases are rare diseases, assembling adequately sized patient collections to undertake association studies will always remain a challenge. However this challenge may be met by the formation of local, regional, national and international collaborations within the pediatric rheumatology community. The sharing and consolidation of patient sample collections through such networks can produce large, international investigations that will continue to provide new insights into the causes of pediatric rheumatic diseases [53, 56, 57].

Identifying the role of non-coding genetic variation in disease susceptibility

The majority of the trait-associated SNPs identified by GWAS fall within intronic or intergenic regions of DNA, while only 5–10% of them occur within the coding region [6, 58]. The historically poor understanding of the mechanisms through which noncoding variation contributes to the development of human disease have been a major obstacle to determining the functional relevance of the majority of GWAS-implicated loci. Fortunately, the development and refinement of strategies to aid in the selection of SNPs most likely to impart functional effects is a rapidly growing field.

In an effort to recognize functionally relevant regions of non-coding DNA, which could bring order to GWAS generated data, the ENCODE project employs various methods to identify genomic regions with regulatory functions. Mapping of DNase I hypersensitivity sites (DHS) is useful in identifying transcription factor-accessible regions of regulatory DNA [4]. Using 125 human cell and tissue types [59], Thurman *et al.* have provided us with a map of cell-specific regulatory regions of DNA represented as peaks (accessible chromatin) and valleys (inaccessible chromatin), the majority of them being downstream of transcriptional start sites, split between intronic and intergenic regions. Validating these findings, they found over 90% of peaks identified by chromatin immunoprecipitation-sequencing (ChIP-Seq) – another method for identifying regions of DNA bound by transcription factors [4] – were found to map within a reported DHS. Similarly, increased methylation of CpG sites, which is thought to be a marker of transcriptional silencing [60], was found to be inversely correlated with chromatin accessibility. Additionally, histone modification has been shown to influence the transcription of genes [61] and cell-specific mapping of these modification hotspots can be used to identify areas of the noncoding genome important for cell-specific gene regulation. Lastly, and perhaps most intriguing, non-coding region data sets are also offering us clues on the commonality of autoimmune disease. Despite their phenotypic differences, of all autoimmune disease GWAS variants that co-localize with DHS in immune cells, about 25% of them can be found within binding sites of 15 transcription factors that interact with the critical immune regulator, interferon regulatory factor 9 (IRF9) [6]

Armed with an increasing amount of cell-specific data that identifies regions of regulatory DNA, researchers can now organize and rank GWAS-implicated variants in order of their likelihood of having a functional impact. In a study of systemic lupus erythematosus [62], by cross-referencing the already established *TNFAIP3* risk loci with known DNA regulatory elements, one risk variant was located in an area defined by high DHS, a H3K27Ac histone mark (which is an indicator of transcriptional activity), and a convergent binding site for 7 transcription factors as determined by ChIP-Seq reads. Functional investigation of this variant revealed that it produced altered binding of NF- κ B, which led to attenuated protein expression in transformed B-cell lines. Likewise, researchers studying thrombocytopenia with absent radii (TAR) identified an association between TAR and a low-frequency 5'UTR variant and a novel intron 1 variant in *RBM8A*, encoding the RNA binding protein Y14 [63]. Based on histone modification data from the 7 human cell lines in the ENCODE project, as well as **formaldehyde-assisted isolation of regulatory elements (FAIRE) combined with sequencing (FAIRE-Seq)** data from megakaryocytes, both of the TAR-

associated variants were found to reside within regions of potentially active regulatory DNA. Further analysis indicated the 5' UTR variant introduced a binding site for the transcriptional repressor, EVI1, while the intronic variant disrupted the binding site for transcription factors MZF1 and RBPJ, leading to a significant decrease in *RMB8A* promoter activity in a human megakaryocytic cell line and a reduction of Y14 protein levels in patient platelet lysates.

Recognizing that alterations within regulatory DNA may have functional importance, some have sought to extend these findings through scrutiny of cell-specific regulatory DNA in an effort to further focus the search for a causal variant [6, 64]. That is, given that DNA regulatory regions differ based on cell-type, it is likely that variants may only have functional relevance in a specific cell type (or types), and that disease-associated SNPs enriched in cell-specific regulatory DNA may serve as an indicator of a cell type's importance in disease pathogenesis. Using previously identified variants from 4 distinct disease phenotypes, Trynka *et al.* were able to pinpoint the cell type whose gene regulation would be most affected. This was accomplished by assigning a cell type specificity score to each associated variant based on its co-localization with cell-specific H3K4me3 peaks, which are located in promoters and are associated with active transcription. Interestingly, they were able to identify the CD4⁺ regulatory T cell as the cell type whose gene regulation is most heavily influenced by known rheumatoid arthritis (RA) variants. Furthermore, they identified a SNP in LD with the published RA risk allele, rs13119723, that was situated between the *IL2* and *IL21* genes and only 116 base pairs from the summit of an H3K4me3 peak found predominantly in CD4⁺ regulatory T cells. Likewise, in a study of celiac disease [64–66], Trynka *et al.* describe the use of dense genotyping to identify a disease-associated SNP that is in strong LD with a potentially causal variant that was also located within a CD4⁺ regulatory T cell-specific H3K4me3 mark. Likewise, Maurano *et al.* [6] examined the locations of established Crohn's disease (CD) and multiple sclerosis (MS) risk variants within regulatory DNA regions of various cell types. For CD, they found a significant enrichment of CD-associated SNPs within the regulatory DNA of both T helper 17 (Th17) and T helper 1 (Th1) lymphocytes, while the MS-associated SNPs were enriched within the regulatory DNA of both cord blood-derived, CD3⁺ T cells and CD19⁺CD20⁺ B lymphocytes, thus raising questions about the contribution of T cell education and development, as well as B cell-mediated pathology in the pathogenesis of MS. Interestingly, they also noted that DHS within brain tissue were found to harbor few MS-associated SNPs, suggesting the effects of SNPs on neuronal gene expression in MS are of less importance than the ones located within cells of the immune system.

Noncoding RNA: regulators of regulation?

After evading detection literally decades since the first description of RNA [67], new species of non-coding RNA are currently being discovered at a rapid pace and non-coding RNA has become an area of active investigation in the fields of rheumatology and immunology. The emergence and growth of RNA sequencing (RNA-Seq) technologies have produced an unexpected expansion of what we have termed the “transcriptome”, shifting our understanding of RNA from a collection of mRNA species awaiting translation into their protein products to a complex collection of RNA molecules that also include important

regulatory elements whose mere existence is challenging the traditional definition of a gene [68, 69]. Adding even further complexity to the transcriptome, not only is its composition cell specific, but about 6% of both coding and non-coding transcripts share sequences with smaller RNA species, raising the possibility that these smaller RNAs may not have been directly transcribed, but produced from their larger counterparts [68]. In this section, we will review our current understanding of non-coding RNA, specifically **micro RNA** (miRNA) and long non-coding RNA (lncRNA), in relation to rheumatologic and immunological disease.

MicroRNA

miRNA are derived from larger transcripts via the cooperative actions of the RNase Drosha and DGCR8 complex in the nucleus, and the endoribonuclease Dicer together with co-factors in the cytosol [70, 71]. They impart their actions through complementary binding to mRNA, leading to inhibition of translation [72]. Three important miRNAs heavily involved in immune regulation are miR-146a, miR-155, and miR-223, whose elevations have been demonstrated in RA synovial tissue [73, 74]. miR-146a, an NF- κ B responsive element, negatively regulates TLR-mediated immune activation through inhibition of the translation of both tumor necrosis factor receptor-associated factor 6 (TRAF6) and interleukin-1 associated kinase 1 (IRAK1) message [75–77]. miR-155, an RNA species first identified in human B cell lymphomas [77], is believed to act as a promoter of inflammation. Highlighting its pro-inflammatory role, overexpression of miR-155 in CD14+ cells leads to a potentiated production of lipopolysaccharide (LPS)-induced proinflammatory cytokines and, in contrast, its knockdown in mice leads to a much-attenuated development of collagen-induced arthritis [78, 79].

Lastly, miR-223 was initially thought to be a myeloid-specific miRNA having anti-inflammatory functions, as its knockdown leads to increased peripheral neutrophilia, an amplified neutrophilic respiratory burst, and development of spontaneous inflammatory lung disease as well as exaggerated LPS-induced organ destruction [80]. In the macrophage, miR-223 limits the production of TLR-induced IL-1 β as well as IL-6 through targeting the activity of STAT3 [81]. Indeed, there exists a conserved miR-223 binding site in the 3' untranslated region of NLRP3, which could lead to decreased inflammasome activity [82, 83]. However, complicating this picture, miR-223 has been demonstrated to be increased in the peripheral T lymphocytes of RA patients [84]. Furthermore, blocking miR-223 in splenic lymphocytes leads to a decrease in LPS-induced interferon- γ (IFN- γ) [85] and transfection of miR-223 into colonic intraepithelial lymphocytes produces increased interleukin-17A production [86], thus raising the possibility that miR-223 could be proinflammatory in lymphocytes.

Long non-coding RNA

Long non-coding RNA (lncRNA) share many features with their translated brethren in that the majority are similarly transcribed by RNA polymerase II, undergo splicing, contain 5' terminal methylguanosine caps and polyadenylated tails [87] and even bind ribosomes, but nevertheless are not believed to be translated [88] due to absence of an open reading frame. The functions of lncRNA have thus far proven to be diverse, owing to their dual ability to

bind RNA (via an unprocessed lncRNA or through binding of smaller derivative species like miRNA, as noted above) and protein [69].

The newly described tumor necrosis factor- α (TNF- α)- and heterogeneous nuclear ribonucleoprotein L (hnRNPL)-related immunoregulatory long intergenic noncoding RNA (lincRNA), THRIL, helps illustrate one of the functionalities of lncRNA [89]. Using a custom lncRNA microarray, Li et al. treated phorbol 12-myristate 13-acetate (PMA)-differentiated THP-1 cells with Pam3CSK4, a TLR1/2 agonist, and found a number of differentially expressed lncRNAs, which were then subsequently tested for their ability to influence TNF- α secretion via shRNA knockdown. THRIL, which is required for Pam3CSK4-induced TNF- α production, was found to bind specifically to hnRNPL in a complex that was determined to bind to the TNF- α promoter. Interestingly, they also found that TNF- α led to the downregulation of THRIL through a negative feedback mechanism, and in agreement with this, whole blood THRIL expression was found to be decreased in acute Kawasaki disease.

In addition to TNF- α , a lncRNA has also been shown to participate in the regulation of IFN- γ and the Th1-mediated immune response. Tmevpg1, or NeST, originally described to aid in the antiviral response [90], is located in proximity to Ifng and found to be upregulated in IFN- γ -producing Th1 polarized T lymphocytes, but not CD8+ T cells, and is dependent on STAT4 and T-bet [91]. Furthermore, Tmevpg1 was shown to interact with the protein WDR5, which catalyzes methylation of histone H3 at lysine 4 (H3K4me3), a chromatin mark associated with active transcription, causing an increase of H3K4me3 at the Ifng locus [92].

Translation of genetic discoveries to targeted therapeutics

The ultimate goal of performing genetic and genomic studies of any human disease is to identify and understand its pathogenesis. We seek to apply new insights about human disease pathogenesis to the identification of novel therapeutic strategies to treat human diseases with increased efficacy and safety. There can be an extremely long process between the identification of a genetic cause or risk factor and the development of a targeted therapy to modulate that gene or pathway, but there are also occasions where an agent originally developed for another disease may be repurposed for a particular disease based on a genetic discovery [93]. This was the case for the recombinant interleukin-1 receptor antagonist (IL1Ra), anakinra, and the treatment of the IL-1-mediated diseases, neonatal-onset multisystem inflammatory disorder (NOMID, also known as chronic infantile neurologic cutaneous and articular [CINCA] syndrome) and the deficiency of IL1Ra (DIRA). NOMID is caused by activating mutations of NLRP3, while DIRA is caused by recessive, loss of function mutations of IL1RN, both of which produce excessive IL-1-mediated signaling with profound consequences. In both cases, therapeutic administration of recombinant IL1RN was highly effective, eliminating both systemic and focal inflammatory manifestations [94, 95]. In the case of the recently identified autoinflammatory disease, DADA2, that is caused by loss of function mutations of ADA2, which has both enzymatic and growth factor-related function, replacement therapy is an intuitive therapeutic approach. Because ADA2 is present in the plasma of healthy individuals, efforts are underway to

investigate the use of fresh frozen plasma or cryoprecipitate as replacement therapy in the treatment of DADA2. Concerning the translation of genetic discoveries to therapeutic strategies in genetically-complex diseases, it has been argued that the product of GWAS studies has been underwhelming in their contributions to our basic understanding and the treatment of human disease [96]. To the contrary, a recent integrated genomic investigation of RA found substantial overlap between GWAS-implicated genes and RA therapeutic targets [50]. By integrating the analysis of the largest meta-analytic GWAS of RA with **functional annotations, expression quantitative trait locus** information and **pathway analysis**, the authors identified a significantly enriched overlap between 98 candidate genes (and the pathways to which they belong) and the 27 approved therapeutic targets for the treatment of RA [50]. Although this study identified the relationships between biologically relevant RA candidate genes and RA therapeutic targets retrospectively, as opposed to recognizing novel therapeutic targets prospectively, this study still affirmed the ability of GWAS to identify both biologically relevant pathways and promising therapeutic targets in genetically complex diseases.

Summary

Approaches for investigating the genetic bases of human diseases are advancing rapidly. NGS technologies have revolutionized genomics, with WES/WGS becoming an important tool for investigating monogenic diseases. Publicly available NGS datasets have become an important resource in genomic studies. In mendelian diseases, allele frequency data from reference populations are commonly employed for variant filtering. In genetically-complex diseases, these reference datasets have facilitated important improvements in SNP imputation and enabled large-scale genomewide meta-analyses. Although targeted deep resequencing is ideally suited to couple with GWAS, it seems likely that the comprehensive genomic datasets generated by WGS will eventually supplant SNP genotyping in the investigation of complex genetic traits. Somatic (or *de novo*) mutations, which are increasingly recognized as a potential cause of sporadic disease, are more readily detected by NGS approaches. Our understanding of the functional importance of noncoding genetic variation is also rapidly expanding. Annotated databases containing ENCODE and other data will improve our interpretation of GWAS findings and their relation to specific cell populations. The ability of NGS to efficiently and rapidly identify causal and risk variants has led to a bottleneck between genetic discovery and functional understanding. This has created the need for a prioritization scheme for testing the functional consequences of variants, and for new model systems and strategies – such as zebrafish models and genetic engineering with CRISPR/cas9 technology [97] – to more efficiently assess their function. The application of these approaches to pediatric rheumatic disease holds great promise for future discoveries.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health

Glossary

Expression quantitative trait loci (eQTL)	are genomic loci that regulate levels of gene expression
Formaldehyde-assisted isolation of regulatory elements-sequencing (FAIRE-Seq)	is a next-generation sequencing application that seeks to determine the sequences of genomic regions with regulatory activity
Functional annotations	are pieces of biological data, such as cell-type specific gene expression or transcription factor binding sites, that may be integrated into genomic analyses to identify known functional relevance for specific genetic variants or regions of interest
Genetically-complex diseases	are caused by a combination of genetic and environmental factors, as opposed to monogenic diseases, which are caused by mutation(s) of a single gene
Imputation	is a statistical process to infer missing or unknown data in an experimental dataset by comparison with a more complete dataset. In the case of SNP imputation, one may use densely-genotyped reference SNP datasets to determine, <i>in silico</i> , the genotypes of unobserved SNPs in less densely-genotyped experimental datasets
Mendelian diseases	are monogenic disease whose inheritance within families complies with Mendel's laws, including autosomal dominant, autosomal recessive and X-linked inheritance
MicroRNA	are short, non-coding RNA molecules that function in transcriptional and post-transcriptional regulation of gene expression
Monogenic diseases	are caused by mutation(s) of a single gene, including diseases of Mendelian inheritance and diseases caused by <i>de novo</i> or somatic mutations
Next-generation sequencing	refers to a group of high-throughput sequencing techniques that can simultaneously generate sequence data at hundreds of thousands to millions of positions
Pathway analysis	is an investigative approach to determine whether a set of disease-associated variants are enriched within known signalling pathways or cellular functions
Targeted deep resequencing	is a next-generation sequencing application that generates sequence data for a specific set of genomic target regions, as

opposed to whole genome sequencing where one sequences the entire genome

References

1. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*. 2013; 369:1502–1511. [PubMed: 24088041]
2. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
3. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)*. 2011; 1:457–470. [PubMed: 22384356]
4. Bernstein BE, Birney E, Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
5. Chadwick LH. The NIH Roadmap Epigenomics Program data resource. *Epigenomics*. 2012; 4:317–324. [PubMed: 22690667]
6. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–1195. [PubMed: 22955828]
7. McDermott MF, Ogunkolade BW, McDermott EM, et al. Linkage of familial Hibernian fever to chromosome 12p13. *Am J Hum Genet*. 1998; 62:1446–1451. [PubMed: 9585614]
8. Ombrello MJ, Remmers EF, Sun G, et al. Cold urticaria, immunodeficiency, and autoimmunity related to PLCG2 deletions. *N Engl J Med*. 2012; 366:330–338. [PubMed: 22236196]
9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977; 74:5463–5467. [PubMed: 271968]
10. Remmers EF, Ombrello MJ, Kanno Y, et al. Principles and techniques in molecular biology. In: MC, Hochberg; AJ, Silman; JS, Smolen, et al., editors. *Rheumatology*. 5th edn. Philadelphia: Mosby/Elsevier; 2011. p. 105–118.
11. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010; 11:31–46. [PubMed: 19997069]
12. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008; 26:1135–1145. [PubMed: 18846087]
13. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009; 461:272–276. [PubMed: 19684571]
14. Goldstein DB, Allen A, Keebler J, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet*. 2013; 14:460–470. [PubMed: 23752795]
15. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*. 2010; 42:30–35. [PubMed: 19915526]
16. Roach JC, Glusman G, Smit AF, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*. 2010; 328:636–639. [PubMed: 20220176]
17. Biesecker LG. Exome sequencing makes medical genomics a reality. *Nat Genet*. 2010; 42:13–14. [PubMed: 20037612]
18. Koboldt DC, Steinberg KM, Larson DE, et al. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013; 155:27–38. [PubMed: 24074859]
19. Zhou Q, Yang D, Ombrello AK, et al. Early-Onset Stroke and Vasculopathy Associated with Mutations in ADA2. *N Engl J Med*. 2014
20. Elkan PN, Pierce SB, Segel R, et al. Mutant Adenosine Deaminase 2 in a Polyarteritis Nodosa Vasculopathy. *N Engl J Med*. 2014
21. Reiff A, Bassuk AG, Church JA, et al. Exome sequencing reveals RAG1 mutations in a child with autoimmunity and sterile chronic multifocal osteomyelitis evolving into disseminated granulomatous disease. *J Clin Immunol*. 2013; 33:1289–1292. [PubMed: 24122031]
22. Boisson B, Laplantine E, Prando C, et al. Immunodeficiency, autoinflammation and amylopectinosis in humans with inherited HOIL-1 and LUBAC deficiency. *Nat Immunol*. 2012; 13:1178–1186. [PubMed: 23104095]

23. Tokunaga F, Sakata S, Saeki Y, et al. Involvement of linear polyubiquitylation of NEMO in NF-kappaB activation. *Nat Cell Biol.* 2009; 11:123–132. [PubMed: 19136968]
24. Zhou Q, Lee GS, Brady J, et al. A hypermorphic missense mutation in PLCG2, encoding phospholipase Cgamma2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *Am J Hum Genet.* 2012; 91:713–720. [PubMed: 23000145]
25. Izawa K, Hijikata A, Tanaka N, et al. Detection of base substitution-type somatic mosaicism of the NLRP3 gene with >99.9% statistical confidence by massively parallel sequencing. *DNA Res.* 2012; 19:143–152. [PubMed: 22279087]
26. Jimenez-Trevino S, Gonzalez-Roca E, Ruiz-Ortiz E, et al. First report of vertical transmission of a somatic NLRP3 mutation in cryopyrin-associated periodic syndromes. *Ann Rheum Dis.* 2013; 72:1109–1110. [PubMed: 23486414]
27. Nakagawa K, Gonzalez-Roca E, Souto A, et al. Somatic NLRP3 mosaicism in Muckle-Wells syndrome. A genetic mechanism shared by different phenotypes of cryopyrin-associated periodic syndromes. *Ann Rheum Dis.* 2013
28. Omoyinmi E, Melo Gomes S, Standing A, et al. Brief Report: Whole-Exome Sequencing Revealing Somatic NLRP3 Mosaicism in a Patient With Chronic Infantile Neurologic, Cutaneous, Articular Syndrome. *Arthritis Rheumatol.* 2014; 66:197–202. [PubMed: 24431285]
29. Schmitt MW, Kennedy SR, Salk JJ, et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A.* 2012; 109:14508–14513. [PubMed: 22853953]
30. Watson IR, Takahashi K, Futreal PA, et al. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet.* 2013; 14:703–718. [PubMed: 24022702]
31. Lindhurst MJ, Sapp JC, Teer JK, et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med.* 2011; 365:611–619. [PubMed: 21793738]
32. Lou DI, Hussmann JA, McBee RM, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A.* 2013; 110:19872–19877. [PubMed: 24243955]
33. Nejentsev S, Walker N, Riches D, et al. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science.* 2009; 324:387–389. [PubMed: 19264985]
34. Diogo D, Kurreeman F, Stahl EA, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am J Hum Genet.* 2013; 92:15–27. [PubMed: 23261300]
35. Kirino Y, Zhou Q, Ishigatsubo Y, et al. Targeted resequencing implicates the familial Mediterranean fever gene MEFV and the toll-like receptor 4 gene TLR4 in Behcet disease. *Proc Natl Acad Sci U S A.* 2013; 110:8134–8139. [PubMed: 23633568]
36. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med.* 2010; 363:166–176. [PubMed: 20647212]
37. 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. [PubMed: 19543373]
38. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39:906–913. [PubMed: 17572673]
39. Howie B, Fuchsberger C, Stephens M, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet.* 2012; 44:955–959. [PubMed: 22820512]
40. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2012; 9:179–181. [PubMed: 22138821]
41. Remmers EF, Cosan F, Kirino Y, et al. Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R–IL12RB2 regions associated with Behcet's disease. *Nat Genet.* 2010; 42:698–702. [PubMed: 20622878]
42. Kirino Y, Bertsias G, Ishigatsubo Y, et al. Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B*51 and ERAP1. *Nat Genet.* 2013; 45:202–207. [PubMed: 23291587]
43. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One.* 2013; 8:e64683. [PubMed: 23762245]

44. Pereyra F, Jia X, McLaren PJ, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*. 2010; 330:1551–1557. [PubMed: 21051598]
45. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*. 1987; 30:1205–1213. [PubMed: 2446635]
46. Raychaudhuri S, Sandor C, Stahl EA, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*. 2012; 44:291–296. [PubMed: 22286218]
47. Ombrello MJ, Kirino Y, de Bakker PIW, et al. Behcet's disease-associated MHC class I residues implicate antigen binding and regulation of cell-mediated cytotoxicity. *Proc Natl Acad Sci U.S.A.* 2014; 111:8867–72. [PubMed: 24821759]
48. Stahl EA, Raychaudhuri S, Remmers EF, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*. 2010; 42:508–514. [PubMed: 20453842]
49. Okada Y, Terao C, Ikari K, et al. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat Genet*. 2012; 44:511–516. [PubMed: 22446963]
50. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014; 506:376–381. [PubMed: 24390342]
51. Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–58. [PubMed: 20811451]
52. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. *Arthritis Res Ther*. 2011; 13:101. [PubMed: 21345260]
53. Hinks A, Cobb J, Marion MC, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet*. 2013; 45:664–669. [PubMed: 23603761]
54. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. *Hum Mol Genet*. 2013; 22:R16–21. [PubMed: 23922232]
55. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011; 43:1066–1073. [PubMed: 21983784]
56. Thompson SD, Marion MC, Sudman M, et al. Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis Rheum*. 2012; 64:2781–2791. [PubMed: 22354554]
57. Thompson SD, Sudman M, Ramos PS, et al. The susceptibility loci juvenile idiopathic arthritis shares with other autoimmune diseases extend to PTPN2, COG6, and ANGPT1. *Arthritis Rheum*. 2010; 62:3265–3276. [PubMed: 20722033]
58. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009; 106:9362–9367. [PubMed: 19474294]
59. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012; 489:75–82. [PubMed: 22955617]
60. Siegfried Z, Eden S, Mendelsohn M, et al. DNA methylation represses transcription in vivo. *Nat Genet*. 1999; 22:203–206. [PubMed: 10369268]
61. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007; 39:311–318. [PubMed: 17277777]
62. Adrianto I, Wen F, Templeton A, et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. *Nat Genet*. 2011; 43:253–258. [PubMed: 21336280]
63. Albers CA, Paul DS, Schulze H, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet*. 2012; 44:435–439. S431–S432. [PubMed: 22366785]
64. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*. 2013; 45:124–130. [PubMed: 23263488]

65. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet.* 2010; 42:295–302. [PubMed: 20190752]
66. Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet.* 2011; 43:1193–1201. [PubMed: 22057235]
67. Eddy SR. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet.* 2001; 2:919–929. [PubMed: 11733745]
68. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature.* 2012; 489:101–108. [PubMed: 22955620]
69. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol.* 2013; 20:300–307. [PubMed: 23463315]
70. Bernstein E, Caudy AA, Hammond SM, et al. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature.* 2001; 409:363–366. [PubMed: 11201747]
71. Lee Y, Ahn C, Han J, et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature.* 2003; 425:415–419. [PubMed: 14508493]
72. Soifer HS, Rossi JJ, Saetrom P. MicroRNAs in disease and potential therapeutic applications. *Mol Ther.* 2007; 15:2070–2079. [PubMed: 17878899]
73. Stanczyk J, Pedrioli DM, Brentano F, et al. Altered expression of MicroRNA in synovial fibroblasts and synovial tissue in rheumatoid arthritis. *Arthritis Rheum.* 2008; 58:1001–1009. [PubMed: 18383392]
74. Murata K, Yoshitomi H, Tanida S, et al. Plasma and synovial fluid microRNAs as potential biomarkers of rheumatoid arthritis and osteoarthritis. *Arthritis Res Ther.* 2010; 12:R86. [PubMed: 20470394]
75. Taganov KD, Boldin MP, Chang KJ, et al. NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A.* 2006; 103:12481–12486. [PubMed: 16885212]
76. Nahid MA, Satoh M, Chan EK. Mechanistic role of microRNA-146a in endotoxin-induced differential cross-regulation of TLR signaling. *J Immunol.* 2011; 186:1723–1734. [PubMed: 21178010]
77. Eis PS, Tam W, Sun L, et al. Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc Natl Acad Sci U S A.* 2005; 102:3627–3632. [PubMed: 15738415]
78. Kurowska-Stolarska M, Alivernini S, Ballantine LE, et al. MicroRNA-155 as a proinflammatory regulator in clinical and experimental arthritis. *Proc Natl Acad Sci U S A.* 2011; 108:11193–11198. [PubMed: 21690378]
79. Bluml S, Bonelli M, Niederreiter B, et al. Essential role of microRNA-155 in the pathogenesis of autoimmune arthritis in mice. *Arthritis Rheum.* 2011; 63:1281–1288. [PubMed: 21321928]
80. Johnnidis JB, Harris MH, Wheeler RT, et al. Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature.* 2008; 451:1125–1129. [PubMed: 18278031]
81. Chen Q, Wang H, Liu Y, et al. Inducible microRNA-223 down-regulation promotes TLR-triggered IL-6 and IL-1beta production in macrophages by targeting STAT3. *PLoS One.* 2012; 7:e42971. [PubMed: 22937006]
82. Haneklaus M, Gerlic M, Kurowska-Stolarska M, et al. Cutting edge: miR-223 and EBV miR-BART15 regulate the NLRP3 inflammasome and IL-1beta production. *J Immunol.* 2012; 189:3795–3799. [PubMed: 22984081]
83. Bauernfeind F, Rieger A, Schildberg FA, et al. NLRP3 inflammasome activity is negatively controlled by miR-223. *J Immunol.* 2012; 189:4175–4181. [PubMed: 22984082]
84. Fulci V, Scappucci G, Sebastiani GD, et al. miR-223 is overexpressed in T-lymphocytes of patients affected by rheumatoid arthritis. *Hum Immunol.* 2010; 71:206–211. [PubMed: 19931339]
85. Dai R, Phillips RA, Zhang Y, et al. Suppression of LPS-induced Interferon-gamma and nitric oxide in splenic lymphocytes by select estrogen-regulated microRNAs: a novel mechanism of immune modulation. *Blood.* 2008; 112:4591–4597. [PubMed: 18791161]
86. Schaefer JS, Montufar-Solis D, Vigneswaran N, et al. Selective upregulation of microRNA expression in peripheral blood leukocytes in IL-10–/– mice precedes expression in the colon. *J Immunol.* 2011; 187:5834–5841. [PubMed: 22043014]

87. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009; 458:223–227. [PubMed: 19182780]
88. Guttman M, Russell P, Ingolia NT, et al. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*. 2013; 154:240–251. [PubMed: 23810193]
89. Li Z, Chao TC, Chang KY, et al. The long noncoding RNA THRIL regulates TNFalpha expression through its interaction with hnRNPL. *Proc Natl Acad Sci U S A*. 2014; 111:1002–1007. [PubMed: 24371310]
90. Vigneau S, Rohrlich PS, Brahic M, et al. Tmevpg1, a candidate gene for the control of Theiler's virus persistence, could be implicated in the regulation of gamma interferon. *J Virol*. 2003; 77:5632–5638. [PubMed: 12719555]
91. Collier SP, Collins PL, Williams CL, et al. Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *J Immunol*. 2012; 189:2084–2088. [PubMed: 22851706]
92. Gomez JA, Wapinski OL, Yang YW, et al. The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell*. 2013; 152:743–754. [PubMed: 23415224]
93. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*. 2013; 12:581–594. [PubMed: 23868113]
94. Goldbach-Mansky R, Dailey NJ, Canna SW, et al. Neonatal-onset multisystem inflammatory disease responsive to interleukin-1beta inhibition. *N Engl J Med*. 2006; 355:581–592. [PubMed: 16899778]
95. Aksentijevich I, Masters SL, Ferguson PJ, et al. An autoinflammatory disease with deficiency of the interleukin-1-receptor antagonist. *N Engl J Med*. 2009; 360:2426–2437. [PubMed: 19494218]
96. Goldstein DB. Common genetic variation and human traits. *N Engl J Med*. 2009; 360:1696–1698. [PubMed: 19369660]
97. Ran FA, Hsu PD, Wright J, et al. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*. :2281–2308.