Immunology & Cell Biology 2021; **99**: 146–156
www.wileyonlinelibrary.com/journal/icb

SPECIAL FEATURE REVIEW

OPEN

Detecting pathogenic variants in autoimmune diseases using high-throughput sequencing

Matt A Field[1,2]

1 Centre for Tropical Bioinformatics and Molecular Biology, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, QLD, Australia
2 John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia

Keywords
Autoimmune diseases, high-throughput sequencing, pathogenic variant, SNV, variant annotation, variant detection

Correspondence
Matt A Field, Centre for Tropical Bioinformatics and Molecular Biology, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, QLD 4878, Australia.
E-mail: matt.field@jcu.edu.au

Received 20 April 2020; Revised 22 June 2020; Accepted 2 July 2020

doi: 10.1111/imcb.12372

Immunology & Cell Biology 2021; **99**: 146–156

Abstract

Sequencing the first human genome in 2003 took 15 years and cost $2.7 billion. Advances in sequencing technologies have since decreased costs to the point where it is now feasible to resequence a whole human genome for $1000 in a single day. These advances have allowed the generation of huge volumes of high-quality human sequence data used to construct increasingly large catalogs of both population-level and disease-causing variation. The existence of such databases, coupled with a high-quality human reference genome, means we are able to interrogate and annotate all types of genetic variation and identify pathogenic variants for many diseases. Increasingly, sequencing-based approaches are being used to elucidate the underlying genetic cause of autoimmune diseases, a group of roughly 80 polygenic diseases characterized by abnormal immune responses where healthy tissue is attacked. Although sequence data generation has become routine and affordable, significant challenges remain with no gold-standard methodology to identify pathogenic variants currently available. This review examines the latest methodologies used to identify pathogenic variants in autoimmune diseases and considers available sequencing options and subsequent bioinformatic methodologies and strategies. The development of reliable and robust sequencing and analytic workflows to detect pathogenic variants is critical to realize the potential of precision medicine programs where patient variant information is used to inform clinical practice.

INTRODUCTION

Autoimmune diseases are a group of roughly 80 polygenic diseases characterized by aberrant immune responses where healthy tissues, organs and cells are attacked. This is caused by the failure of immune systems to respond appropriately to self-antigens and results in damage to tissues and organs. Autoimmune diseases are a heterogenous group of diseases with regard to pathogenicity, heritability and prevalence, and currently few effective therapies exist.[1] Some of the most common autoimmune diseases are rheumatoid arthritis, type 1 diabetes (T1D), inflammatory bowel syndrome, systemic lupus erythematosus (SLE) and Sjögren's syndrome.

Autoimmune diseases represent a global health burden with an estimated occurrence rate of 4.5%, and disproportionately affect females at a rate of 6.4% compared with 2.7% for males.[1] Prevalence rates of autoimmune diseases are rising, with a recent report from the British Society for Immunology estimating disease incidence growth at a rate of 3%–9% annually.[2] Prevalence of autoimmune diseases varies according to a wide variety of environmental and genetic factors; however, the influence of such factors varies considerably across the family of autoimmune diseases. Gender is a significant factor in some systemic conditions such as SLE and Sjögren's syndrome with 90% of cases occurring in females, whereas T1D and Guillain–Barré syndrome exhibit no gender bias.[3] Geography also plays a role with an estimated 1 in 12 people being affected by an autoimmune disease in the Western Hemisphere,[4] a higher estimate than the rest of the world. A smaller study found that individuals in Finland have six times higher rates of T1D compared with individuals from the adjacent Karelian republic of Russia despite sharing the same genetic background.[5] Ethnicity also plays a major role in many autoimmune diseases with significant differences observed with regard to incidence rates and disease severity.[6] For example, African Americans are five to nine times more likely to develop SLE than European Americans and typically develop more severe SLE which exhibits a greater number of manifestations and is more damaging.[7] However, other autoimmune diseases are more prevalent in Northern Europeans, as they are more susceptible to T1D than ethnic Chinese.[8] Little is known regarding the underlying mechanisms for the observed disparities between ethnicities; however, differences in human leukocyte antigen (HLA) regions are thought to contribute.[6]

Although environmental factors are known to contribute to autoimmune diseases, genetic factors are increasingly recognized to play a key role.[9] Many types of autoimmune diseases such as inflammatory bowel syndrome[10] and SLE[11] show familial clustering, and subsequent twin studies also exhibit high concordance rates among monozygotic twins.[12] Heritability estimates vary across autoimmune diseases, with a recent study of pediatric age autoimmune cohorts estimating 86% heritability for T1D at the high end compared with 43% for Crohn's disease at the low end.[4] The high estimated heritability and early successes in identifying pathogenic variants from patient sequence data[13] have led to increasingly large genetic studies being undertaken. These studies continue to link new genes to monogenic autoimmune disorders, with the latest Inborn Errors of Immunity report documenting 430 known defects, a gain of 64 additional gene defects in the last 2 years alone.[14]

Early work to elucidate the underlying contribution of genetic variation to autoimmune diseases focused on increasingly large genome-wide association studies (GWASs). GWASs successfully identified numerous risk loci, and a review identified 819 unique loci across 136 separate GWASs.[6] Although successful, GWAS was only able to account for a small portion of the estimated heritability in autoimmune diseases, meaning most heritability remained unexplained. A possible explanation for the missing heritability arises from a limitation of GWAS, as it only examines common single-nucleotide variants (SNVs). The advent of cheap sequencing allows other variation types to be interrogated, with results showing significant contribution to autoimmune diseases from rare SNVs,[15,16] indels,[17] somatic mosaicism[18] and structural and copy number variation.[19] In addition, immune system–specific applications such as the sequencing of HLA regions,[20,21] T-cell receptors (TCR)[18,22] and B-cell receptors (BCR)[23] have helped to better understand their unique role in autoimmune diseases.

Autoimmune diseases are also variable in response to treatment and increasingly these differences are being attributed to genetic variation.[24] Individual patient sequencing can help inform clinical practice, yet this requires increasingly sophisticated bioinformatics software and methodologies to reliably detect pathogenic variants. This review focuses on the sequencing options and bioinformatics methodologies currently used to discover pathogenic variants that drive autoimmune diseases. While sequence data generation is now routine, strategies to effectively reduce the search space for pathogenic variants are critical to the development of successful personalized medicine programs for autoimmune diseases.

SEQUENCING OPTIONS

There is a wide variety of affordable high-throughput sequencing technologies available to help identify variants contributing to autoimmune diseases (Table 1).

Sequencing options for Mendelian disorders are numerous and the most common approach is short-read DNA-based methods that sequence either custom gene panels, exomes or whole genomes. Gene panel sequencing yields high-depth coverage across preselected genes of interest by performing an initial capture step. However, such approaches are limited as they presume an existing knowledge of disease-implicated genes and limit novel discoveries. Exome sequencing can capture over 95% of all exons and splice site regions across all known genes and is an extremely popular option, costing roughly one-third of the cost of genome sequencing. The disadvantage of exome sequencing is the inability to identify most noncoding variants and the failure to reliably detect types of variation larger than SNVs and small indels. Whole-genome sequencing offers the most comprehensive and unbiased view across all variation types; however, the associated cost with this method is the highest among available methods.

All high-throughput sequencing options generate large volumes of raw data that make pathogenic variant detection identification challenging. Additional options to reduce the variant search space are often deployed alongside short-read DNA-based methods. These include sequencing patient RNA, adding unique molecular identifiers (UMIs) to individual molecules, employing single-cell technologies to detect somatic and immune system subset–specific variation and utilizing long-read technologies such as Oxford Nanopore or PacBio.

UMI sequencing is commonly used to sequence heterogenous cell populations containing mixtures of

Table 1. Sample of sequencing options available for variant detection in autoimmune diseases

Sequencing type	Detectable variation	Advantages	Limitations
GWAS	Loci on GWAS chip	Cheap/large studies possible	Only common SNVs
Whole genome	All variant types: coding and noncoding	All variant types detectable	Expensive relative to targeted approaches
Exome	SNV and small indel in coding regions	Capture most coding regions	No noncoding or large variation
Gene panel	SNV and small indel in panel genes	High-depth coverage for panel genes	Nothing novel is detectable
Molecular tagging	Somatic and cell subset–specific variants	Analyze individual input molecules	Additional library preparation / custom software
Single cell	Somatic and cell subset–specific variants	Analyze individual cells	Additional library preparation / custom software
HLA typing	HLA genotypes	High-resolution phased HLA genotypes	Immune system specific
BCR	B-cell clonotypes	Construct and observe changes in BCR	Immune system specific
TCR	T-cell clonotypes	Construct and observe changes in TCR	Immune system specific
Transcriptome	Aberrant splicing/gene fusions/coding SNV	Observe effect of variants on genes	Miss rare transcripts/added expense
Long reads	All variant types: coding and noncoding	Resolve large variants/full gene transcripts	Higher error rate and higher per-base cost

BCR, B-cell receptor; GWAS, genome-wide association study; HLA, human leukocyte antigen; SNV, single-nucleotide variant; TCR, T-cell receptor.

both wild-type and disease-causing cells. The technology works by affixing a UMI to each individual input DNA molecule, often prior to PCR amplification. After sequencing, the software deconvolutes the UMIs and reads sharing UMIs are pooled together for analysis, with each group representing an individual input DNA molecule.[25] UMI sequencing is often combined with single-cell omics technologies, which are able to analyze large numbers of individual cells simultaneously. Single-cell technologies are changing our understanding of immunology by allowing us to examine many aspects of the immune system subsets in great detail, including their inherent variation.[26] For example, a recent study identified lymphoma driver mutations in a specific cell lineage that was producing pathogenic autoantibodies.[18]

RNA sequencing is increasingly employed in tandem with DNA sequencing to identify potential noncoding pathogenic variants. The functional information provided by RNA sequencing identifies dysregulated genes, many of which result from genetic changes, which allows the closer examination of the small number of candidate genes of interest. A recent study increased diagnostic rates by 35% relative to genome sequencing alone by identifying pathogenic variants responsible for exon skipping, exon expansion and intronic splice gains.[27] With RNA sequencing, it is critical to only sequence disease-specific tissue and to only compare samples across identical tissue types. This is not possible for all diseases, which limits its widespread applicability.

Long-read sequencing is increasingly recognized as a valuable tool to resolve larger complex genetic variants such as repeat expansions, copy number and structural variation and also for sequencing the TCR/BCR and HLA regions. While pathogenic variation detection using long-read sequencing has been most successful in cancer and neurological disorders detection thus far,[28] researchers are now applying this approach in autoimmune diseases. Long-read sequencing may also be used in pseudogene discrimination, with a recent study developing a robust diagnostic application that is able to unambiguously sequence three autoimmune diseases genes (*IKBKG*, *IRAK4* and *MYD88*) while bypassing the *IKBKGP1* pseudogene.[29] Currently, the major issues that prevent the uptake of long-read sequencing are the increased per-base cost and the higher error rates relative to short-read sequencing. However, both cost and error rates are continually improving.

Several sequencing applications are specific to the immune system. Deep sequencing of the HLA, TCR and BCR regions is now possible, with variation in these regions implicated in causing autoimmune diseases.[18,20,21] Although it is possible to identify HLA, TCR and BCR sequence using standard approaches, these applications yield better results with additional capture steps and bespoke software. For example, resolving HLA types is challenging using standard approaches because of low HLA sequence coverage and the incomplete representation of the HLA region in the human reference genome, resulting from its highly polymorphic nature. As such, many companies now offer deep sequencing of the HLA region by providing specific capture assays able to yield high-resolution phased HLA sequences. Similarly, deep-sequencing TCR and BCR clonotypes require additional steps including target enrichment, multiplex PCR or molecular tagging prior to sequencing. TCR/BCR sequencing is further confounded as a result of the fulllength V(D)J chain being 330 bp; this is longer than individual short reads and thus requires custom software to accurately reconstruct B- and T-cell clonotypes.

SAMPLE SELECTION STRATEGIES

In addition to sequencing options, careful patient sample selection has been shown to increase the success rates in pathogenic variant identification. While sample selection is not always an option, studies have shown that for singleton samples it is optimal to focus on early onset cases with extreme phenotypes and a clearly defined clinical phenotype.[30] These strategies have been used to identify a growing number of pathogenic variants in singleton samples which are deposited into clinical repositories that serve to link patient data from around the world. Searching across samples is a powerful approach, with a recent study describing a new immune dysregulation resulting from the linking of two singleton studies from unrelated cohorts in Australia and Japan.[31]

Compared with singleton sequencing, the most effective method for detecting pathogenic variants is sequencing multiple individuals within a family or pedigree. Sequencing a pedigree generates family-wide variant information, such as disease inheritance pattern, compound heterozygosity and genome phasing, later used for additional variant prioritization.[32] Another immediate benefit of this approach is the ability to catalog familial variation, often incorrectly assumed to be pathogenic because of its absence in databases of population-level variation. The greatest successes with pedigree sequencing come from sequencing trios, which consist of an affected child and unaffected parents. In such cases, it is likely the causal variant will be a *de novo* mutation in the affected child, which serves to greatly reduce the variant search space. This approach is also informative, with larger pedigrees exhibiting complex inheritance patterns, as variants shared between affected individuals are prioritized and variants shared with unaffected family members are deprioritized. Collectively, these sample selection strategies are able to greatly reduce the causal variant search space.

VARIANT DETECTION WORKFLOW

A typical variant detection analysis workflow consists of six major analysis steps: data quality control/adapter trimming, read alignment, alignment file preprocessing, variant detection, variant annotation and variant prioritization. The early workflow steps generally require more computation time and work with larger data sets, whereas the later workflow steps require more domain-specific analyses and offer a greater variety of software choices (Figure 1).

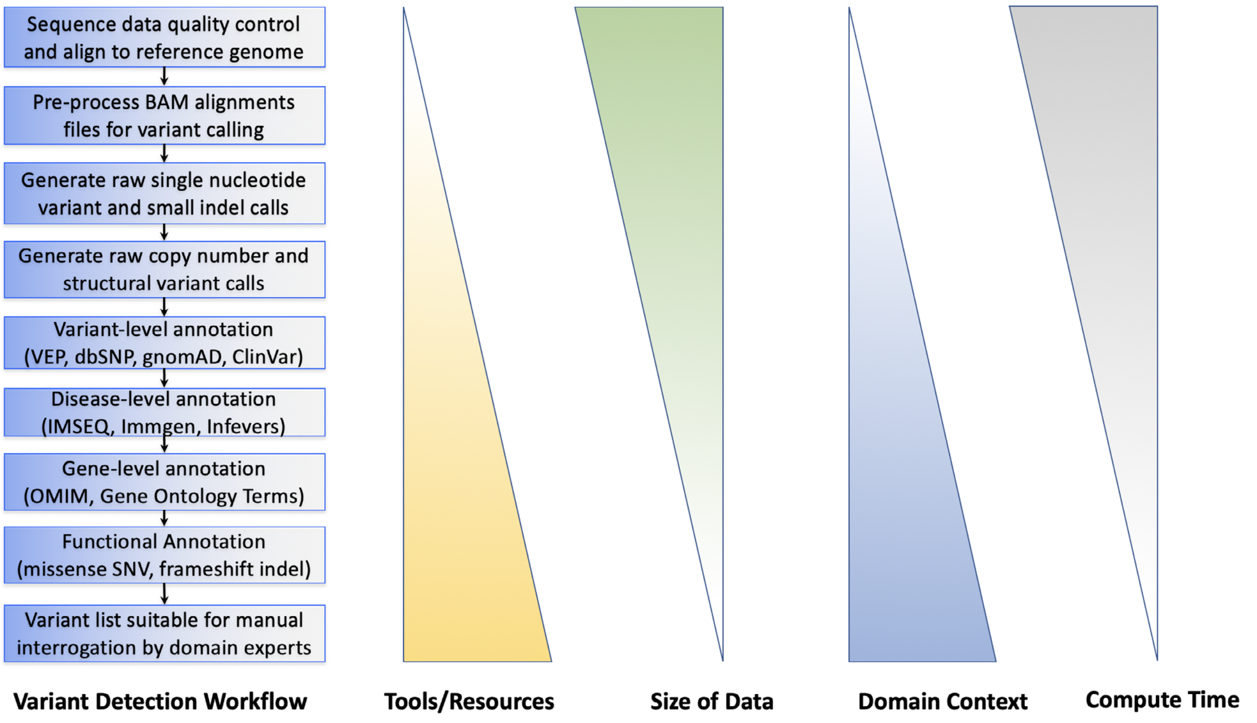
Sequence data quality control and align to reference genome
Pre-process BAM alignments files for variant calling
Generate raw single nucleotide variant and small indel calls
Generate raw copy number and structural variant calls
Variant-level annotation (VEP, dbSNP, gnomAD, ClinVar)
Disease-level annotation (IMSEQ, Immgen, Infevers)
Gene-level annotation (OMIM, Gene Ontology Terms)
Functional Annotation (missense SNV, frameshift indel)
Variant list suitable for manual interrogation by domain experts
Variant Detection Workflow
Tools/Resources
Size of Data
Domain Context
Compute Time

Figure 1. Variant detection workflow. SNV, single-nucleotide variant.

While there is currently no accepted end-to-end gold-standard methodology for identifying pathogenic variants, analysis steps leading to the generation of variant calls have become relatively standardized. By contrast, variant annotation and prioritization are specific to individual variant detection workflows and often contain analysis steps specific to the disease being studied. A sample of common open-source software packages for each workflow step is listed in Table 2.

The first analysis step is sequence data quality control and adapter trimming. Trimmomatic[33] is a popular tool used for this purpose, which identifies and removes adapter sequence, trims low-quality bases from the end of reads and removes reads with a high total fraction of low-quality bases using a sliding window approach. Following data quality control, reads are aligned to the gold-standard reference human genome GRCh38.p13 using a short-read aligner such as Burrows-Wheeler Aligner (BWA).[34] BWA first constructs an index of the reference genome and aligns individual reads to the index by anchoring small seed subsequences that allow base mismatches and gaps to account for sequencing errors. Local read alignments are expanded as far as possible around each matching seed and the highest scoring alignment is selected. Aligners output a compressed alignment file in binary alignment map (BAM) format which is then optimized for variant calling. This consists of marking potential duplicate reads, realigning reads around candidate indels to account for local misalignments and recalculating the base qualities to account for systematic errors made by the sequencing machine during the estimation of base call accuracies.

The processed BAM file is used as input to the variant calling algorithms where different types of variation are detected relative to the human reference genome. Variant detection algorithms aim to differentiate real genetic variation from experimental error by employing statistical methodologies specific to each type of variation, with the exception being the simultaneous detection of SNVs and small indels by algorithms such as Genome Analysis Toolkit (GATK).[35] Most variant detection algorithms assign a variant quality score and apply a hard cut-off when generating a list of true variants. However, an increasingly popular approach is to perform variant "group calling," where variants are detected simultaneously across a larger cohort. Group variant calling can be used to identify missed variants that would otherwise fall just below the quality cut-off scores because of issues such as low base coverage or skewed allele frequency ratios. Another increasingly common approach shown to improve variant calling quality is to run multiple algorithms and then take a consensus of variant calls.[39,40] This approach is particularly relevant in clinical variant detection workflows where false-negative variants are of the greatest concern.

Variant detection algorithms output raw variant lists in variant call format which are filtered to remove candidate false positives. Variants are generally filtered based on variant context characteristics such as low-quality alignment scores, low read depth, low-quality base scores or variant clustering. Yet, some algorithms such as Variant Quality Score Recalibration (VQSR) from GATK employ a machine learning approach that uses a variant truth set to differentiate true- and false-positive variants.

Generating raw variant lists for large sequence data sets (such as whole genomes) requires large computational resources and storage, typically 1000 CPU hours and 500-GB storage. In terms of variant numbers, a genome contains roughly 4 million SNVs, 400 000 small indels and 100 000 copy-number variations/structural variations. While generating these high-quality variant calls requires significant computational resources, the workflow is relatively standardized: the challenge is in reducing the variant search space to identify the variants most likely to be pathogenic. The approaches for annotating and prioritizing variants are less standardized

Table 2. Common software options for analysis steps in variant detection workflow

Analysis step	Example software	URL
Data quality control/trimming	Trimmomatic[33]	http://www.usadellab.org/cms/?page=trimmomatic
Read alignment	BWA[34]	http://bio-bwa.sourceforge.net/
BAM preprocessing	Picard	https://broadinstitute.github.io/picard/
Variant calling (SNV/indel)	GATK[35]	https://gatk.broadinstitute.org/hc/en-us
Variant calling (UMI tags)	DeepSNVMiner[25]	https://github.com/mattmattmattmatt/DeepSNVMiner
Variant calling (structural variation/copy-number variation)	Manta[36]	https://github.com/Illumina/manta
Variant calling (Pedigree)	VASP[32]	https://github.com/mattmattmattmatt/VASP
Variant annotation	Variant Effect Predictor[37]	https://ensembl.org/info/docs/tools/vep/index.html
Variant prioritization	PolyPhen-2[38]	http://genetics.bwh.harvard.edu/pph2/

SNV, single-nucleotide variant; UMI, unique molecular identifier.

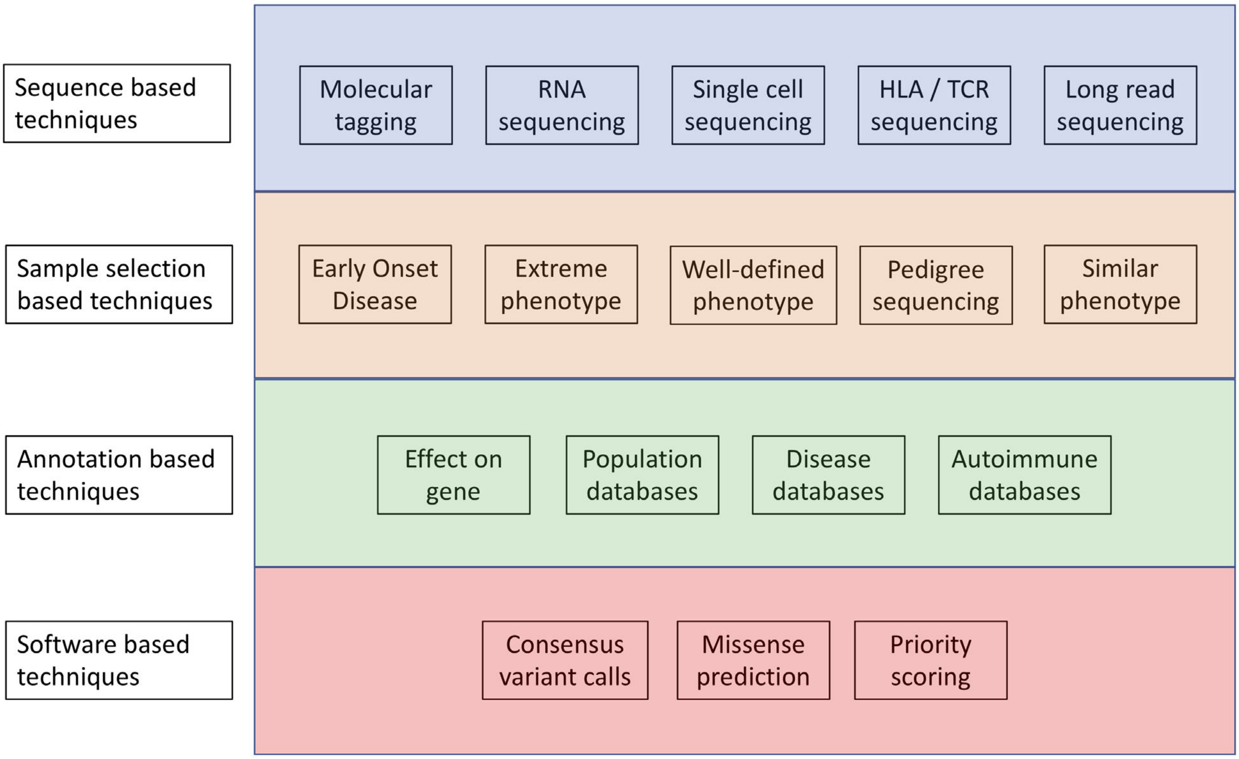

Figure 2. Strategies to reduce the variant search space for pathogenic variants. HLA, human leukocyte antigen; TCR, T-cell receptor.

than earlier steps and often combine publicly available software and data sets with custom code and annotations. Collectively, software-based strategies to prioritize pathogenic variants are combined with sequence-based and sample selection strategies that enrich for pathogenic variants to form the basis of a successful pathogenic variant detection workflow (Figure 2).

The first step in variant annotation is to determine the impact of the variant on genes or other important genomic features. This requires a reliable gene model such as Ensembl or RefSeq along with a variant annotation tool such as Ensembl's Variant Effect Predictor.[37] For SNVs, those classified as missense, nonsense or splice sites mutations are prioritized, with missense mutations further run through software such as PolyPhen-2,[38] which predicts the likely impact of the amino acid substitution on protein function. These predictive algorithms consider factors such as evolutionary sequence conservation, protein structure and overlap with protein features such as binding sites. While these algorithms generally have low false-negative rates, they suffer from high false positives, with a recent study finding that over 50% of predicted damaging missense mutations being functionally benign.[41] For other variant types, small indels are prioritized if they overlap genes and cause frameshift mutations, whereas larger variants are examined in terms of knocked out genes/exons or potential gene fusions. Variants are also compared with catalogs of both population-level variation (e.g. dbSNP[42]/gnomAD[43]) and disease-specific variants (e.g. ClinVar[44]). With population-level variant repositories, comparing with ethnically matched allele frequencies is critical to account for the often-large allele frequency differences observed between ethnic groups. The prioritization strategy differs depending on the nature of the data set with variants' overlapping entries in disease databases taken forward while variants found to occur at high allele frequency in the general population removed from further consideration.

The final step in the process is amalgamating all the information into prioritized ranked lists that contain the variants most likely to be pathogenic. Software such as GEMINI[45] attempts to generate prioritized variant lists; however, in general, this process is largely unsuitable for external software because of the development of custom in-house methodologies employed throughout the workflow. Ultimately any successful workflow will score true pathogenic variants highly which allows domain experts to identify pathogenic variants through manual interrogation of a small number of candidates.

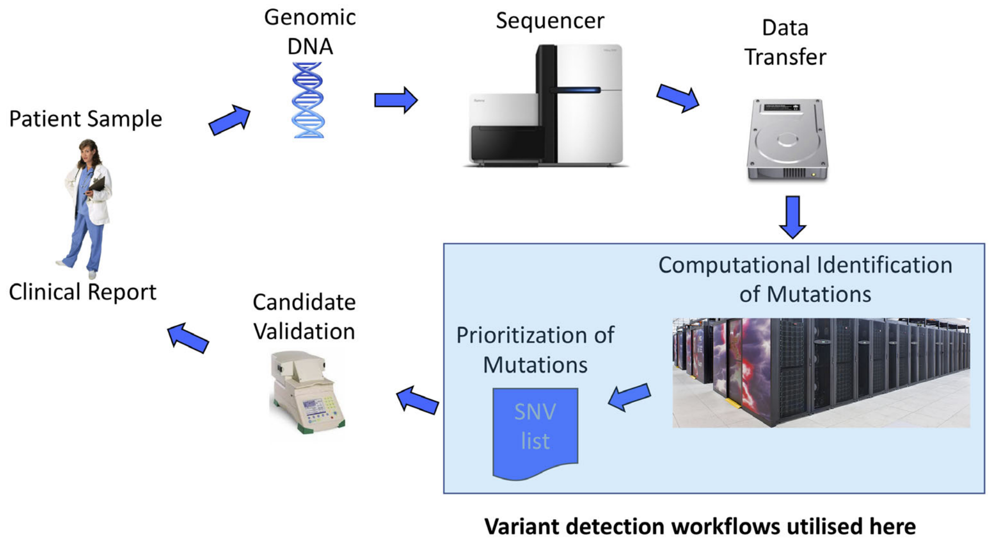
Patient Sample
Genomic DNA
Sequencer
Data Transfer
Computational Identification of Mutations
Prioritization of Mutations
SNV list
Candidate Validation
Clinical Report
Variant detection workflows utilised here

Figure 3. Personalized medicine workflow. SNV, single-nucleotide variant.

Table 3. Resources to reduce variant search space in autoimmune diseases

Analysis type	Software/database	URL
HLA sequencing	HLAminer[46]	https://www.bcgsc.ca/resources/software/hlaminer
HLA sequencing	seq2HLA[47]	https://bitbucket.org/sebastian_boegel/seq2hla/src/default/
HLA sequencing	OptiType[48]	https://github.com/FRED-2/OptiType
HLA sequencing	PHLAT[49]	https://sites.google.com/site/phlatfortype/
TCR/BCR sequencing	MiXCR[50]	https://mixcr.readthedocs.io/en/master/
TCR/BCR sequencing	VDJPuzzle[51]	https://github.com/simone-rizzetto/VDJPuzzle
TCR/BCR sequencing	IMSEQ[52]	http://www.imtools.org/
BCR sequencing	IgDiscover[53]	https://github.com/NBISweden/IgDiscover/
Annotations	ImmGen[54]	http://www.immgen.org/
Annotations	InnateDB[55]	http://www.innatedb.com
Annotations	Immuno Polymorphism Database[56]	https://www.ebi.ac.uk/ipd/index.html
Annotations	Centre for Personalised Immunology	https://database.cpi.org.au/cpi28/
Annotations	Infevers[57]	https://infevers.umai-montpellier.fr/web/
Annotations	LOVD 2.0[58]	http://www.lovd.nl

BCR, B-cell receptor; HLA, human leukocyte antigen; TCR, T-cell receptor.

Such workflows are increasingly forming the basis of personalized medicine programs, such as the Centre for Personalised Immunology in Australia or the Relent Project in Europe. Precision medicine programs are broader in scope than variant detection workflows and begin with patient recruitment and culminate in the creation of a concise clinical variant report used to inform clinical decision making (Figure 3).

For complex diseases with heterogenous genetic causes and confounding environmental factors, such as autoimmune diseases, the default resultant variant lists are often large and unsuitable for manual interrogation and require additional custom analyses to further reduce the variant search space. Additional measures to further reduce the variant search space for autoimmune diseases include immune system–specific annotation, sequencing the HLA region and sequencing the TCR/BCR regions (Table 3).

The most common approach in autoimmune diseases is to annotate variants with immune system–specific data sets such as ImmGen,[54] InnateDB,[55] IMSEQ,[52] Immuno Polymorphism Database,[56] Infevers[57] and locus-specific LOVD databases.[58] These databases contain information covering a variety of aspects of the immune system which can be used to prioritize pathogenic autoimmune diseases variants. For example, ImmGen contains gene expression data for immune cells in mouse, whereas Infevers contains information on hereditary autoinflammatory disorder mutations. Although such resources are useful, an overreliance on any single data source is ill advised, as entries are often inconsistent because of evaluations made using incomplete functional evidence. To illustrate, a follow-up study of 239 annotated disease-causing variants listed in the Human Gene Mutation Database[59] was only able to recapitulate the results for 7.5% of the entries. This lack of reproducibility highlights the importance of working as much as possible with up-to-date resources that are expertly curated and rigorous in their inclusion criteria such as the Inborn Errors of Immunity report.[14]

Sequencing applications such as deep sequencing of the BCR, TCR and HLA regions are unique to studies of the immune system and require application-specific software. While software specific to these applications is maturing, recent benchmarking reviews of the available software for both HLA sequencing[60] and TCR sequencing[61,62] report high levels of variability in overall software performance. Recent reviews of BCR sequencing also discuss available software; however, individual algorithms were not benchmarked in these studies.[63,64] Both reviews stress the importance of using the extensive IMGT database[56] for clonotype assignment and the importance of constructing a complete catalog of all allelic variants using algorithms such as IgDiscover.[53] In the review of HLA typing software, six algorithms were run across a "gold-standard" data set, and found OptiType[48] to be the most accurate at 99%. However, the algorithm only detects Class I HLA genotypes, thus limiting its clinical utility. Among algorithms able to detect both Class I and II HLA genotypes, PHLAT[49] had the highest accuracy at 81%: it was noted that this is likely insufficient for clinical utility, and a consensus software approach was proposed as a possible hybrid solution.[40] In the review of TCR sequencing software, the first study generated an *in silico* data set and assessed clonotype detection, CDR3 identification, error correction and gene segment assignment accuracy.[61] This study found that not all algorithms were able to run all four subanalyses and that the performances varied greatly across individual algorithms, particularly for gene assignment and error correction. The second study performed a similar analysis[62] and concluded that no single tool performed optimally for all types of analyses but recommended MiXCR[50] if limited to a single analysis. All review studies note that superior results can be obtained using UMIs; however, none of the software assessed in the previous studies was able to incorporate this information.

Lastly, an illustrative example is described where a novel pathogenic variant resulted in the description of a new syndrome characterized by global immune dysregulation. The variant was added to the Inborn Errors of Immunity database in 2019. In this study, two unrelated patients from Australia and Japan exhibited similar phenotypes resulting in the destruction of lymphocytes that lead to excessive inflammation.[31] Both patients were exome sequenced and analyzed using an existing variant detection pipeline[32] which identified a candidate causal heterogeneous variant in the *IKBKB* gene (inhibitor of nuclear factor kappa-B kinase subunit beta). The variant was prioritized as it was novel, resulted in a missense mutation that was predicted to be damaging and occurred in an active site of *IKBKB*, a gene which was previously implicated in causing combined immune deficiency.[65] The variant was confirmed with Sanger sequencing to replicate the result using the current gold-standard sequencing method. Further evidence was provided for the Australian patient by sequencing the unaffected parents which, following confirmation of paternity, demonstrated that the mutation had arisen *de novo* in the patient. While the evidence was substantial, functional validation was required using CRISPR–Cas [clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated] technology to engineer the exact mutation into a mouse model which generated a similar immunodeficiency phenotype to the observed patients. This example highlights the value of using sequencing technologies to elucidate the underlying genetic cause of autoimmune diseases.

FUTURE APPROACHES

This review discusses the current landscape in high-throughput sequencing and bioinformatic workflows for pathogenic variant detection in autoimmune diseases. Sequence-based approaches continue to grow, with an increasing number of precision medicine initiatives around the world focused on autoimmune diseases. While such initiatives are currently limited to short-read DNA-based sequencing that use either gene panels, exomes or whole genomes, increasingly long-read sequencing, single-cell technologies, transcriptome sequencing, immune profiling and molecular tagging techniques are being incorporated. Looking beyond the current approaches, researchers are now recognizing the impact of the epigenome[66] and the microbiome[67] on autoimmune diseases and in the future will integrate these data types into workflows. Combining the disparate data types will require a new generation of complex bioinformatics software and statistical methodologies able to quickly and efficiently elucidate the cause of autoimmune diseases.

ACKNOWLEDGMENTS

I thank the National Computational Infrastructure (Australia) for continued access to significant computation resources and technical expertise. I thank the members of the Centre for Personalised Immunology for ongoing support and collaboration. The author is funded by NHMRC (Grant No. APP5121190).

CONFLICT OF INTEREST

The author has no conflict of interest.

AUTHOR CONTRIBUTION

Matt A Field: Conceptualization; Writing-original draft; Writing-review & editing.

REFERENCES

1. Hayter SM, Cook MC. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev* 2012; **11**: 754–765.
2. Immunology BSf. Report reveals the rising rates of autoimmune conditions 2018. https://www.immunology.org/news/report-reveals-the-rising-rates-autoimmune-conditions
3. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. *Front Neuroendocrinol* 2014; **35**: 347–369.
4. Li YR, Zhao SD, Li J, *et al.* Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nat Commun* 2015; **6**: 8442.
5. Kondrashova A, Reunanen A, Romanov A, *et al.* A six-fold gradient in the incidence of type 1 diabetes at the eastern border of Finland. *Ann Med* 2005; **37**: 67–72.
6. Ramos PS, Shedlock AM, Langefeld CD. Genetics of autoimmune diseases: insights from population genetics. *J Hum Genet* 2015; **60**: 657–664.
7. Lewis MJ, Jawad AS. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. *Rheumatology (Oxford)* 2017; **56**(suppl_1): i67–i77.
8. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Curr Diab Rep* 2013; **13**: 814–823.
9. Anaya JM, Gomez L, Castiblanco J. Is there a common genetic basis for autoimmune diseases? *Clin Dev Immunol* 2006; **13**: 185–195.
10. Nunes T, Fiorino G, Danese S, Sans M. Familial aggregation in inflammatory bowel disease: is it genes or environment? *World J Gastroenterol* 2011; **17**: 2715–2722.
11. Alarcon-Segovia D, Alarcon-Riquelme ME, Cardiel MH, *et al.* Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. *Arthritis Rheum* 2005; **52**: 1138–1147.
12. Bogdanos DP, Smyk DS, Rigopoulou EI, *et al.* Twin studies in autoimmune disease: genetics, gender and environment. *J Autoimmun* 2012; **38**: J156–J169.
13. Taupin D, Lam W, Rangiah D, *et al.* A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Hum Genome Var* 2015; **2**: 15013.
14. Tangye SG, Al-Herz W, Bousfiha A, *et al.* Human inborn errors of immunity: 2019 update on the classification from the International Union of Immunological Societies Expert Committee. *J Clin Immunol* 2020; **40**: 24–64.
15. Jiang SH, Athanasopoulos V, Ellyard JI, *et al.* Functional rare and low frequency variants in BLK and BANK1 contribute to human lupus. *Nat Commun* 2019; **10**: 2201.
16. Johar AS, Mastronardi C, Rojas-Villarraga A, *et al.* Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjögren's syndrome. *J Transl Med* 2015; **13**: 173.
17. Dideberg V, Kristjansdottir G, Milani L, *et al.* An insertion-deletion polymorphism in the interferon regulatory Factor 5 (IRF5) gene confers risk of inflammatory bowel diseases. *Hum Mol Genet* 2007; **16**: 3008–3016.
18. Singh M, Jackson KJL, Wang JJ, *et al.* Lymphoma driver mutations in the pathogenic evolution of an iconic human autoantibody. *Cell* 2020; **5**: 878–894.
19. Yim SH, Jung SH, Chung B, Chung YJ. Clinical implications of copy number variations in autoimmune disorders. *Korean J Intern Med* 2015; **30**: 294–304.
20. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. *Nat Rev Immunol* 2018; **18**: 325–339.
21. Bodis G, Toth V, Schwarting A. Role of human leukocyte antigens (HLA) in autoimmune diseases. *Rheumatol Ther* 2018; **5**: 5–20.
22. McGuire HM, Watkins TS, Field M, *et al.* TCR deep sequencing of transgenic RAG-1-deficient mice reveals endogenous TCR recombination: a cause for caution. *Immunol Cell Biol* 2018; **96**: 642–645.
23. Hampe CSB. Cell in autoimmune diseases. *Scientifica (Cairo)* 2012; **2012**: 215308.
24. Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat Med* 2015; **21**: 730–738.
25. Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ* 2016; **4**: e2074.
26. Gaublomme JT, Yosef N, Lee Y, *et al.* Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* 2015; **163**: 1400–1412.
27. Cummings BB, Marshall JL, Tukiainen T, *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* 2017; **9**: 386.
28. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet* 2019; **10**: 426.

29. Frans G, Meert W, der Werff V, *et al*. Conventional and single-molecule targeted sequencing method for specific variant detection in IKBKG while bypassing the IKBKGP1 pseudogene. *J Mol Diagn* 2018; **20**: 195–202.
30. Johar AS, Anaya JM, Andrews D, *et al*. Candidate gene discovery in autoimmunity by using extreme phenotypes, next generation sequencing and whole exome capture. *Autoimmun Rev* 2015; **14**: 204–209.
31. Cardinez C, Miraghazadeh B, Tanita K, *et al*. Gain-of-function IKBKB mutation causes human combined immune deficiency. *J Exp Med* 2018; **215**: 2715–2724.
32. Field MA, Cho V, Cook MC, *et al*. Reducing the search space for causal genetic variants with VASP: variant analysis of sequenced pedigrees. *Bioinformatics* 2015; **31**: 2377–2379.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
35. McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
36. Chen X, Schulz-Trieglaff O, Shaw R, *et al*. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016; **32**: 1220–1222.
37. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; **26**: 2069–2070.
38. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
39. Waardenberg AJ, Field MA. consensusDE: an R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. *PeerJ* 2019; **7**: e8206.
40. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies. *PLoS One* 2015; **10**: e0143199.
41. Miosge LA, Field MA, Sontani Y, *et al*. Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci USA* 2015; **112**: E5189–E5198.
42. Sherry ST, Ward MH, Kholodov M, *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.
43. Karczewski KL, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**(7809): 434–443.
44. Landrum MJ, Lee JM, Riley GR, *et al*. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**(D1): D980–D985.
45. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol* 2013; **9**: e1003153.
46. Warren RL, Choe G, Freeman DJ, *et al*. Derivation of HLA types from shotgun sequence datasets. *Genome Med* 2012; **4**: 95.
47. Boegel S, Lower M, Schafer M, *et al*. HLA typing from RNA-Seq sequence reads. *Genome Med* 2012; **4**: 102.
48. Szolek A. HLA typing from short-read sequencing data with OptiType. *Methods Mol Biol* 2018; **1802**: 215–223.
49. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genom* 2014; **15**: 325.
50. Bolotin DA, Poslavsky S, Mitrophanov I, *et al*. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 2015; **12**: 380–381.
51. Rizzetto S, Koppstein DNP, Samir J, *et al*. B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* 2018; **34**: 2846–2847.
52. Kuchenbecker L, Nienen M, Hecht J, *et al*. IMSEQ–a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* 2015; **31**: 2963–2971.
53. Corcoran MM, Phad GE, Vazquez Bernat N, *et al*. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* 2016; **7**: 13642.
54. Shay T, Kang J. Immunological Genome Project and systems immunology. *Trends Immunol* 2013; **34**: 602–609.
55. Breuer K, Foroushani AK, Laird MR, *et al*. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res* 2013; **41**(D1): D1228–D1233.
56. Robinson J, Halliwell JA, Hayhurst JD, *et al*. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 2015; **43**(D1): D423–D431.
57. Sarrauste de Menthiere C, Terriere S, Pugnere D, Ruiz M, Demaille J, Touitou I. INFEVERS: the Registry for FMF and hereditary inflammatory disorders mutations. *Nucleic Acids Res* 2003; **31**: 282–285.
58. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD vol 2.0: the next generation in gene variant databases. *Hum Mutat* 2011; **32**: 557–563.
59. Stenson PD, Mort M, Ball EV, *et al*. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017; **136**: 665–677.
60. Bauer DC, Zadoorian A, Wilson LOW, Melbourne Genomics Health A, Thorne NP. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief Bioinform* 2018; **19**: 179–187.
61. Zhang Y, Yang X, Zhang Y, *et al*. Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. *Brief Bioinform* 2019.
62. Afzal S, Gil-Farina I, Gabriel R, *et al*. Systematic comparative study of computational methods for T-cell receptor sequencing data analysis. *Brief Bioinform* 2019; **20**: 222–234.

63. Kim D, Park D. Deep sequencing of B cell receptor repertoire. *BMB Rep* 2019; **52**: 540–547.
64. Chaudhary N, Wesemann DR. Analyzing immunoglobulin repertoires. *Front Immunol* 2018; **9**: 462.
65. Pannicke U, Baumann B, Fuchs S, *et al*. Deficiency of innate and acquired immunity caused by an IKBKB mutation. *N Engl J Med* 2013; **369**: 2504–2514.
66. Ye J, Gillespie K, Rodriguez S. Unravelling the Roles of Susceptibility Loci for Autoimmune Diseases in the Post-GWAS Era. *Genes* 2018; **9**(8): 377.
67. De Luca F, Shoenfeld Y. The microbiome in autoimmune diseases. *Clin Exp Immunol* 2019; **195**: 74–85.