Predicting the occurrence of variants in *RAG1* and *RAG2*

Dylan Lawless[a,b,*], Hana Lango Allen[c,d], James Thaventhiran[e], NIHR BioResource–Rare Diseases Consortium[c], Flavia Hodel[b], Rashida Anwar[a], Jacques Fellay[b,f], Jolan E. Walter[g,h], Sinisa Savic[i,j,*]

[a] *Leeds Institute of Biomedical and Clinical Sciences, University of Leeds, Wellcome Trust Brenner Building, St James's University Hospital, Beckett Street, Leeds, UK.*
[b] *Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland.*
[c] *NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, CB20QQ, UK.*
[d] *Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB20XY, UK.*
[e] *MRC Toxicology Unit, School of Biological Sciences, University of Cambridge, UK.*
[f] *Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland.*
[g] *University of South Florida and Johns Hopkins All Children's Hospital, Saint Petersburg, Florida, USA.*
[h] *Division of Allergy Immunology, Massachusetts General Hospital for Children, Boston, Massachusetts, USA.*
[i] *Department of Clinical Immunology and Allergy, St James's University Hospital, Beckett Street, Leeds, UK.*
[j] *National Institute for Health Research Leeds Musculoskeletal Biomedical Research Centre and Leeds Institute of Rheumatic and Musculoskeletal Medicine, Wellcome Trust Brenner Building, St James's University Hospital, Beckett Street, Leeds, UK.*

Abstract

While widespread genome sequencing ushers in a new era of preventive medicine, the tools for predictive genomics are still lacking. Time and resource limitations mean that human diseases remain uncharacterised because of an inability to predict clinically relevant genetic variants. A strategy of targeting highly conserved protein regions is used commonly in functional studies. However, this benefit is lost for rare diseases where the attributable genes are mostly conserved. An immunological disorder exemplifying this challenge occurs through damaging mutations in *RAG1* and *RAG2* which presents at an early age with a distinct phenotype of life-threatening immunodeficiency or autoimmunity. Many tools exist for variant pathogenicity prediction but these cannot account for the probability of variant occurrence. Here, we present a method that predicts the likelihood of mutation for every amino acid residue in the RAG1 and RAG2 proteins. Population genetics data from approximately 146,000 individuals was used for rare variant analysis. Forty-four known pathogenic variants reported in patients and recombination activity measurements from 110 RAG1/2 mutants were used to validate calculated scores. Probabilities were compared with 98 currently known human cases of disease. A genome sequence dataset of 558 patients who have primary immunodeficiency but that are negative for RAG deficiency were also used as validation controls. We compared the difference between mutation likelihood and pathogenicity prediction. Our method builds a map of most probable mutations allowing

[*] Addresses for correspondence

Email addresses: `Dylan.Lawless@epfl.ch` (Dylan Lawless), `S.Savic@leeds.ac.uk` (Sinisa Savic)

pre-emptive functional analysis. This method may be applied to other diseases with hopes of improving preparedness for clinical diagnosis.

Funding

This work is funded by the University of Leeds 110 Anniversary Research Scholarship and by the National Institute for Health Research (NIHR, grant number RG65966). This work was also supported by the National Institutes of Health (sub-R01AI100887-05 to J.E.W.) and Robert A. Good Endowment, University of South Florida (to J.E.W.).

Acknowledgements

We gratefully acknowledge the participation of all NIHR BioResource volunteers, and thank the NIHR BioResource centres and staff for their contribution. We thank the National Institute for Health Research and NHS Blood and Transplant. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Ethics statement

The study was performed in accordance with the Declaration of Helsinki. The NIHR BioResource projects were approved by Research Ethics Committees in the UK and appropriate national ethics authorities in non-UK enrolment centres.

Key words

Recombination activating genes 1 and 2 (RAG1, RAG2), pathogenic variant, genomics, predictive.

Abbreviations

BCR (B-cell receptor), CADD (combined annotation dependent depletion), CID-G/A (combined immunodeficiency with granuloma and/or autoimmunity), GWAS (genome-wide association studies), HGMD (human gene mutation database), M_r (mutation rate), MRF (mutation rate residue frequency), PID (primary immunodeficiency), pLI (probability of being loss-of-function intolerant), *RAG1* (recombination activating gene 1), R_f (residue frequency), RNH (RNase H), RSS (recombination signal sequence), SCID (severe combined immunodeficiency), TCR (T-cell receptor).

Introduction

Costs associated with genomic investigations continue to reduce [1] while the richness of data generated increases. Globally, the adoption of wide scale genome sequencing implies that all new-born infants may receive screening for pathogenic genetic variants in an asymptomatic stage, pre-emptively [2]. The one dimensionality of individual genomes is now being expanded by the possibility of massive parallel sequencing for somatic variant analysis and by single-cell or lineage-specific genotyping; culminating in a genotype spectrum. In whole blood, virtually every nucleotide position may be mutated across 10^5 cells [3]. Mapping one's genotype across multiple cell types and at several periods during a person's life may soon be feasible [4]. Such genotype snapshots might allow for prediction and tracking of somatic, epigenetic, and transcriptomic profiling.

The predictive value of genomic screening highly depends on the computation tools used for data analysis and its correlation with functional assays or prior clinical experience. Interpretation of that data is especially challenging for rare human genetic disorders; candidate disease-causing variants that are predicted as pathogenic often require complex functional investigations to confirm their significance. There is a need for predictive genomic modelling with aims to provide a reliable guidance for therapeutic intervention for patients harbouring genetic defects for life-threatening disease before the illness becomes clinically significant.

The study of predictive genomics is exemplified by consideration of gene essentiality, accomplished by observing intolerance to loss-of-function variants. Several gene essentiality scoring methods are available for both the coding and non-coding genome [5]. Approximately 3,000 human genes cannot tolerate the loss of one allele [5]. The greatest hurdle in monogenic disease is the interpretation of variants of unknown significance while functional validation is a major time and cost investment for laboratories investigating rare disease.

Severe, life-threatening immune diseases are caused by genetic variations in almost 300 genes [6, 7] however, only a small percentage of disease causing variants have been characterised using functional studies. Several robust tools are in common usage for predicting variant pathogenicity. Compared to methods for pathogenicity prediction, a void remains for predicting mutation probability, essential for efficient pre-emptive validation. Our investigation aims to apply predictive genomics as a tool to identify genetic variants that are most likely to be seen in patient cohorts.

We present the first application of our novel approach of predictive genomics using Recombination activating gene 1 (RAG1) and RAG2 deficiency as a model for a rare primary immunodeficiency (PID) caused by autosomal recessive variants. *RAG1* and *RAG2* encode lymphoid-specific proteins that are essential for V(D)J recombination. This genetic recombination mechanism is essential for a robust immune response by diversification the T and B cell repertoire in the thymus and bone marrow, respectively [8, 9]. Deficiency of RAG1 [10] and RAG2 [11] in mice causes inhibition of B and T cell

development. Schwarz et al. [12] formed the first publication reporting that RAG mutations in humans causes severe combined immunodeficiency (SCID), and deficiency in peripheral B and T cells. Patient studies identified a form of immune dysregulation known as Omenn syndrome [13, 14]. The patient phenotype includes multi-organ infiltration with oligoclonal, activated T cells. The first reported cases of Omenn syndrome identified infants with hypomophic RAG variants which retained partial recombination activity [15]. RAG deficiency can be measured by in vitro quantification of recombination activity [16–18]. Hypomorphic *RAG1* and *RAG2* mutations, responsible for residual V(D)J recombination activity (on average 5-30%), result in a distinct phenotype of combined immunodeficiency with granuloma and/or autoimmunity (CID-G/A) [2, 19, 20].

Human RAG deficiency has traditionally been identified at very early ages due to the rapid drop of maternally-acquired antibody in the first six months of life. A loss of adequate lymphocyte development quickly results in compromised immune responses. More recently, we have found that RAG deficiency is also found for some adults living with PID [16].

RAG1 and *RAG2* are highly conserved genes but disease is only reported with autosomal recessive inheritance. Only 44% of amino acids in RAG1 and RAG2 are reported as mutated on GnomAD and functional validation of candidate variants is difficult [21]. Pre-emptive selection of residues for functional validation is a major challenge; a selection based on low allele frequency alone is infeasible since the majority of each gene is highly conserved. A shortened time between genetic analysis and diagnosis means that treatments may be delivered earlier. RAG deficiency may present with diverse phenotypes and treatment strategies vary. With such tools, early intervention may be prompted. Some patients could benefit from hematopoietic stem cell transplant [22] when necessary while others may be provided mechanism-based treatment [23]. Here, we provide a new method for predictive scoring that was validated against groups of functional assay values, human disease cases, and population genetics data. We present the list of variants most likely seen as future determinants of RAG deficiency, meriting functional investigation.

Methods

Population genetics and data sources

GnomAD (version r2.0.2) [21] was queried for the canonical transcripts of *RAG1* and *RAG2* from population genetics data of approximately 146,000 individuals; ENST00000299440 (*RAG1*) 1586 variants, GRCh37 11:36532259-36614706 and ENST00000311485 (*RAG2*) 831 variants, GRCh37 11:36597124 - 36619829. Data was filtered to contain the variant effect identifiers: frameshift, inframe deletion, inframe insertion, missense, stop lost, or stop gained. Reference transcripts were sourced from Ensembl in the FASTA format amino acid sequence for transcript RAG1-201 ENST00000299440.5

[HGNC:9831] and transcript RAG2-201 ENST00000311485.7 [HGNC:9832]. These sequences were converted to their three-letter code format using *One to Three* from the Sequence Manipulation Suite (SMS2) [24]. Combined Annotation Dependent Depletion (CADD) scores were sourced from https://cadd.gs.washington.edu/download (Nov 2018) and are reported by Kircher et al. [25]. The dataset used was "All possible SNVs" from whole genome data, from which we extracted the data for coding regions of *RAG1* and *RAG2*. We used the Human Gene Mutation Database (HGMD) from the Institute of Medical Genetics in Cardiff as a pre-defined source of known RAG deficiency cases http://www.hgmd.cf.ac.uk/ac/index.php (Feb 2019, free access version to NM_000448.2.) [26]. Data was formatted into CSV and imported into R for combined analysis with PHRED-scaled CADD scores and the main dataframe. The crystal structure render of DNA bound RAG complex was produced with data from RCSB Protein Data Bank (3jbw.pdb) [27]. Structures were visualised using the software VMD from the Theoretical and Computational Biophysics Group [28], imaged with Tachyon rendering [29], and colour mapped using our scoring method.

Data processing

The population genetics input dataset used GnomAD variant allele frequencies and reference sequences processed as CSV files, cleaned and sorted to contain only amino acid codes, residue numbers, alternate residues, alternate allele frequencies, and a score of 0 or 1 to indicate presence or absence of variants where 1 represented none reported. An annotation column was also provided to label where multiple alternate variants existed. Statistics and calculation steps are listed in order in **Supplemental Tables E3-E8**.

The percentage of conserved residues was calculated (55.99% of amino acids contained no reported variants in RAG1, 55.98% in RAG2 (**Table E4**)) . Basic protein statistics were generated using canonical reference transcript sequences of *RAG1* and *RAG2* with the SMS2 tool *Protein Stats* [24]. The resulting pattern percentage value was converted to a frequency based on the number of residues per protein to generate the residue frequency (R_f). The R_f values were found for both proteins as shown in **Table E5** and summarised in **Table E6**.

The count of variants per residue were found for both proteins and the mutation rates (M_r) per residue were calculated as shown in **Table E7**. M_r was found by counting the number of mutations per residue in a window, sized to contain each protein individually. For genome-wide application the window size may be increased or decreased. In this case the window consisted of only the coding regions. The M_r values were then converted to the frequencies based on the number of residues per protein. Separate, and overlapping windows could also be used based on genome phase data and regions of linkage disequilibrium to account for non-random association of alleles at different loci; this might be particularly important for disorders with multiple genetic determinants.

The M_r and R_f multiply to give the raw mutation rate residue frequency (MRF) value (**Table E8**). This value is also shown in **Tables 1 and E1**. Our investigation used a Boolean score C to account for the presence or absence of a mutation in the general population; 0 for any variant existing in the population and 1 for conserved residues. $C \times M_r \times R_f$, in our case, produced the MRF score for conserved residues. **Figure 1 (ii)** illustrates the raw MRF as a histogram and the MRF, after applying C, as a heatmap.

An important consideration for future application is whether to use this Boolean score or instead use a discrete variable which accounts for the true allele frequency in the general population. In the clinical setting, the likelihood of *de novo* mutations and inherited mutations have different impacts when considering recessive and dominant diseases. A patient is more likely to inherit a variant that exists even at a very low frequency than to acquire a random *de novo* mutation. Therefore, a value representing an allele frequency may be used to replace C in many investigations, particularity when considering variants that exist at low rates. PRHED-scaled CADD score data consisted of nucleotide level values. For comparison with MRF, the median CADD scores were averaged per codon as demonstrated in Supplemental text. A summary of data processing and analysis is illustrated in **Figure E1**.

Raw data availability and analysis script

The supplemental files *"Raw_data_R_analysis_for_figures"* contains all raw data and analysis methods used to produce figures (except illustrations in Figures 1 and 6). *"data_analysis.R"* is an R script that contains the methods used to produce figures. Each of the input data CSV files are explained on first usage within the analysis script. Running *"data_analysis.R"* from within the same directory as the associated input data CSV files will replicate analysis.

Data visualisation

For our visualisation of MRF scores, small clusters of high MRF values were of more appealing than individual highly conserved residues. Therefore, we applied a 1% average filter where values were averaged over a sliding window of N number of residues (10 in the case of RAG1, 6 in the case of RAG2). For a clear distinction of MRF clusters, a cut-off threshold was applied at the 75th percentile (e.g. 0.0168 in RAG1) as shown in heatmaps in **Figure 1 (iii)** and **6**. The gene heatmaps for coding regions in *RAG1* and *RAG2* (**Fig. 1**) were populated with (i) Boolean C score from population genetics data, (ii) raw MRF scores, and (iii) MRF clusters with 1% average and cut-off threshold. GraphPad Prism was used for heatmaps. The data used for heatmaps is available in **Table E1** and in the supplemental R source to allow for alternative visualisations. An example of alternative output for non-R users is shown in **Figure E2**. Adobe Illustrator and Photoshop were used for protein domain illustrations in **Figure 1 (iv)**.

Validation of MRF against functional data

The recombination activity of RAG1 and RAG2 was previously measured on known or candidate pathogenic variants [16–18]. Briefly, the pathogenicity of variants in *RAG1* and *RAG2* was measured functionally *in vitro* by either expression of RAG1 and RAG2 in combination with a recombination substrate plasmid containing recombination signal sequence (RSS) sites which are targeted by RAG complex during normal V(D)J recombination, or Abelson virus-transformed Rag2-/- pro-B cells with an RSS-flanked inverted GFP cassette. Recombination events were assessed by quantitative real-time PCR using comparative CT or expression of GFP evaluated by flow cytometry, respectively. The inverse score of recombination activity (0-100%) was used to quantify pathogenicity of variants in our study. Comparison between known pathogenicity scores and MRF was done by scaling MRF scores from 0-100% (100% being highest probability of occurring as damaging). A data and analysis is summarised in **Figure E1**.

Results

RAG1 and RAG2 conservation and mutation rate residue frequency

Variant probability prediction is dependent on population genetics data. Our study queried GnomAD [21] to identify conserved residues using a Boolean score C of 0 (present in population) or 1 (conserved). The gene-specific mutation rate M_r of each residue was calculated from allele frequencies. The gene-specific residue frequency R_f represented the frequency of a residue occurring per gene, acquired by converting gene residue percentage (from the SMS2 tool *Protein stats*) to a frequency [24]. Together the values were used to calculate the most probable disease-causing variants which have not yet been identified in patients. We termed the resulting score a mutation rate residue frequency, where $MRF = C \times M_r \times R_f$. This score represents the likelihood that a clinically relevant mutation will occur.

Figure 1 presents the most probable unidentified disease-causing variants in *RAG1/2*. Variants with a low MRF may still be damaging but resources for functional validation are best spent on gene regions with high MRF. Clusters of conserved residues are shown in **Figure 1 (i)** and are generally considered important for protein structure or function. However, these clusters do not predict the likelihood of mutation. Raw MRF scores are presented in **Figure 1 (ii)**. Histograms illustrates the MRF without Boolean scoring applied and **Figure 1 (iii)** provides a clearer illustration of top MRF score clusters. For visualisation, a noise reduction method was applied; a sliding window was used to find the average MRF per 1% interval of each gene. The resulting scores displayed in **Figure 1 (iii)** contain a cut-off threshold to highlight the top scoring residues (using the 75th percentile). Variant sites most likely to present in disease cases are identified by high MRF scoring. This model may be expanded by the addition of phenotypic or epigenetic data (**Supplemental**).

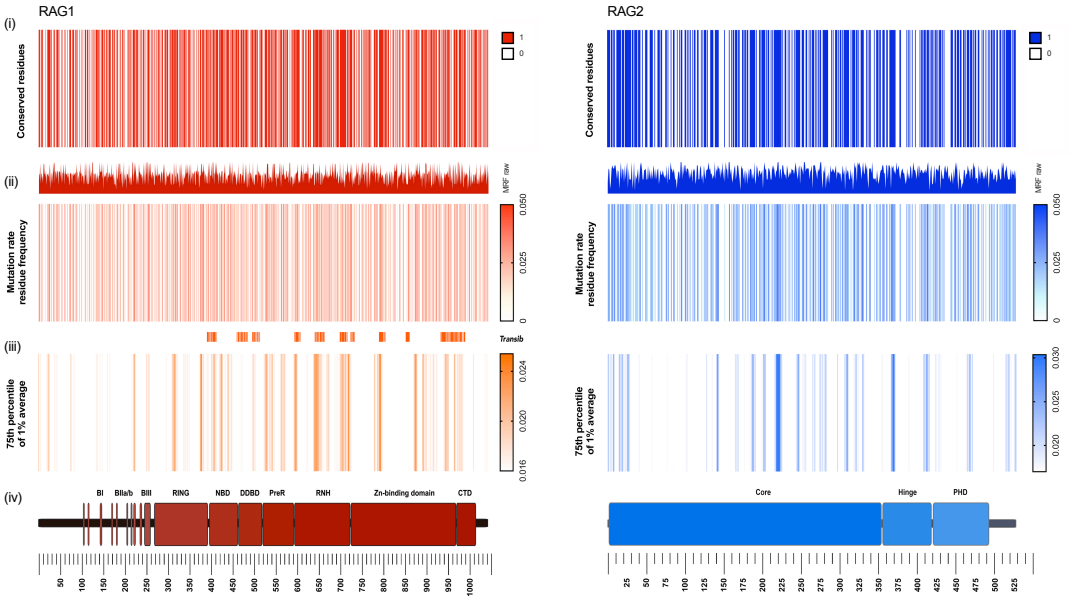

Figure 1: RAG1 (red, left) and RAG2 (blue, right) conservation and mutation rate residue frequency. (i) Gene conservation score; non-conserved 0 and conserved 1. Colour indicates no known mutations in humans. (ii) Histogram; raw MRF score. Heatmap; MRF prediction for conserved residues, graded 0 to 0.05 (scale of increasing mutation likelihood with human disease). (iii) Coloured bars indicate most likely clinically relevant variant clusters. MRF score averaged with 1% intervals for each gene and cut-off below 75th percentile, graded 0 to 0.03 (noise reduction method). (iv) Gene structure with functional domains. Full list of residues and scores available in **Table E1**.

Table E1 provides all MRF scores for both proteins. Raw data used for calculations and the list of validated residues of RAG1 and RAG2 are available in **Tables E3-E8**. **Summary table 1** shows the MRF mutation likelihood score for mutations that have also been reported as tested for recombination activity in functional assays. Analysis-ready files are also available in Supplemental data along with the associated R source file to allow for alternative visualisations as shown in **Figure E2**.

RAG1

MRF	Residue	Assayed	Recombition activity (%)	MRF	Residue	Assayed	Recombition activity (%)
0.030	56	I56T	3.5 ± 0.2	0.005	522	W522C	41.6 ± 1.9
0.030	86	K86VfsX33	2.7 ± 0.3	0.022	539	D539V	3.2 ± 0.2
0.014	99	G99S	113.2 ± 3.7	0.025	541	L541CfsX30	1.2 ± 0.9
0.030	100	M1006V	105.6 ± 6.8	0.043	559	R559S	1.0 ± 0.4
0.012	106	N106K	80.4 ± 16.4	0.043	561	R561H	2.0 ± 0.6
0.043	108	R108X	1.8 ± 0.3	0.041	601	S601P	0.0 ± 0.0
0.043	142	R142X	9.0 ± 4.0	0.026	612	H612R	121.6 ± 0.9
0.032	174	E174SfsX27	0.5 ± 0.2	0.043	624	R624H	0.0 ± 0.4
0.027	246	A246TfsX17	0.8 ± 0.1	0.041	626	S626X	0.0 ± 0.0
0.012	248	Q248X	1.2 ± 0.2	0.041	651	S651P	0.5 ± 0.5

0.026	249	H249R	112.2 ± 3.5	0.043	699	R699Q,W	45.9 ± 1.5, 19.3 ± 1.8
0.043	314	R314W	24.3 ± 5.2	0.032	722	E722K	0.0 ± 0.2
0.012	328	C328Y	16.0 ± 2.9	0.012	730	C730F	0.0 ± 0.0
0.030	383	K383RfsX7	0.1 ± 0.0	0.025	732	L732P	0.0 ± 0.0
0.013	386	F386CfsX4	0.2 ± 0.1	0.043	737	R737H	0.2 ± 0.0
0.030	391	K391E	6.5 ± 1.6	0.043	759	R759C	17.2 ± 3.3
0.043	394	R394Q	0.1 ± 0-0.1	0.043	764	R764P	0.0 ± 0.0
0.043	396	R396C	0.4-0.6 ± 0-0.1	0.008	768	Y768X	0.0 ± 0.0
0.041	401	S401P	0.0 ± 0.0	0.032	770	E770K	21.0 ± 0.4
0.020	403	T403P	0.0 ± 0.0	0.043	778	R778Q,W	8.6 ± 1.0, 4.6 ± 0.6
0.043	404	R404Q	1.2 ± 0.1	0.028	786	P786L	0.0 ± 0.1
0.043	410	R410Q	0.0 ± 0.0	0.030	820	K820R	117.9 ± 6.3
0.025	411	L411P	0.0 ± 0.0	0.025	836	L836V	75.0 ± 1.3
0.022	429	D429G	0.1 ± 0.0	0.043	841	R841Q,W	0.0 ± 0.0, 10.0 ± 0.5
0.028	433	V433M	0.2 ± 0.0	0.027	868	A868V	100.0 ± 5.0
0.019	435	M435V	23.6 ± 4.8	0.005	896	W896R	0.9 ± 0.1
0.027	444	A444V	1.4 ± 0.2	0.008	912	Y912C	6.9 ± 0.4
0.043	449	R449K	92.1 ± 3.6	0.005	959	W959X	0.0 ± 0.0
0.025	454	L454Q	5.4 ± 0.7	0.032	965	E965X	0.0 ± 0.0
0.019	458	M458SfsX34	0.0 ± 0.0	0.043	973	R973C	0.0 ± 0.2
0.027	472	A472V	0.4 ± 0.0	0.013	974	F974L	56.5 ± 0.8
0.043	474	R474C	125.4 ± 2.6	0.043	975	R975W,Q	57.9 ± 1.6, 53.5 ± 3.6
0.028	475	V475AfsX17	0.1 ± 0.0	0.012	981	Q981P	7.2 ± 0.1
0.025	506	L506F	1.0 ± 0.1	0.030	983	K983NfsX9	0.1 ± 0.0
0.043	507	R507W	15.9 ± 0.8	0.030	992	K992E	9.1 ± 1.2
0.014	516	G516A	40.2 ± 1.3				

RAG2

MRF	Residue	Assayed	Recombition activity (%)	MRF	Residue	Assayed	Recombition activity (%)
0.013	1	M1T	65.3 ± 2.2	0.013	285	M285R	24.7 ± 0.8
0.006	16	Q16X	1.7 ± 0.4	0.004	307	W307X	0.2 ± 0.2
0.038	35	G35A,V	22.1 ± 3.1, 0.4 ± 0.3	0.017	386	F386L	109.1 ± 5
0.023	39	R39G	0.2 ± 0.1	0.025	407	E407X	2.9 ± 0.4
0.011	41	C41W	0.2 ± 0.4	0.004	416	W416L	1.4 ± 0.2
0.017	62	F62L	19.6 ± 3	0.025	437	E437K	0.9 ± 0.2
0.028	65	D65Y	6.8 ± 1.2	0.017	440	K440N	26.7 ± 2.4
0.023	73	R73H	12.4 ± 1.4	0.013	443	M443I	0.4 ± 0.2
0.034	77	T77N	42.6 ± 2.7	0.027	444	I444M	2.7 ± 0.3
0.038	95	G95R	0.3 ± 0.2	0.011	446	C446W	2.9 ± 0.1
0.013	110	M110L	74.6 ± 1.8	0.038	451	G451A	66.3 ± 4.8
0.017	127	K127X	0.1 ± 0	0.004	453	W453R	0.6 ± 0.1
0.038	157	G157V	0.4 ± 0.2	0.017	456	A456T	16 ± 2.9
0.030	160	S160L	5.8 ± 0.6	0.013	459	M459L	30.8 ± 0.6
0.023	180	P180H	31.1 ± 0.5	0.034	474	N474S	97.5 ± 5.9
0.019	195	Y195D	2 ± 0.3	0.011	478	C478Y	0.2 ± 0.1
0.034	215	T215I	67.2 ± 1	0.025	480	E480X	2.8 ± 0.6
0.023	229	R229Q,W	8.9 ± 1, 10.5 ± 0.5	0.017	481	H481P	23.8 ± 3.9
0.023	253	P253R	95.4 ± 2.3	0.013	502	M502V	99.6 ± 3.4
0.006	278	Q278X	0.1 ± 0.1				

Table 1: MRF likelihood scores for variants functionally assayed to date [16–18]. Increased MRF score indicate higher likelihood of occurrence. Recombination activity measured by functional assay is shown as a percentage of wildtype (% SEM). Residues with multiple mutations are shown with both alternative variants and values. $MRF_{max} = 0.043$ and $MRF_{min} = 0.004$. The full table of all protein positions can be found in Supplemental Table E1.

MRF scores select for confirmed variants in human disease

We have applied MRF scores to known damaging mutations from other extensive reports in cases of human disease [12, 15, 17, 19, 20, 30–53] [*originally compiled by* Notarangelo et al. [54]]. This dataset compares a total of 44 variants. We expected that functionally damaging variants (resulting in low recombination activity in vitro) that have the highest probability of occurrence would be identified with high MRF scores. MRF prediction correctly identified clinically relevant mutations in *RAG1* and *RAG2* (**Fig. 2 (i)**). Variants reported on GnomAD which are clinically found to cause disease had significantly higher MRF scores than variants which have not been reported to cause disease. We observed that rare and likely mutations provided high scores while rare but unlikely or common variants had low scores (**Fig. 2 (i)**).

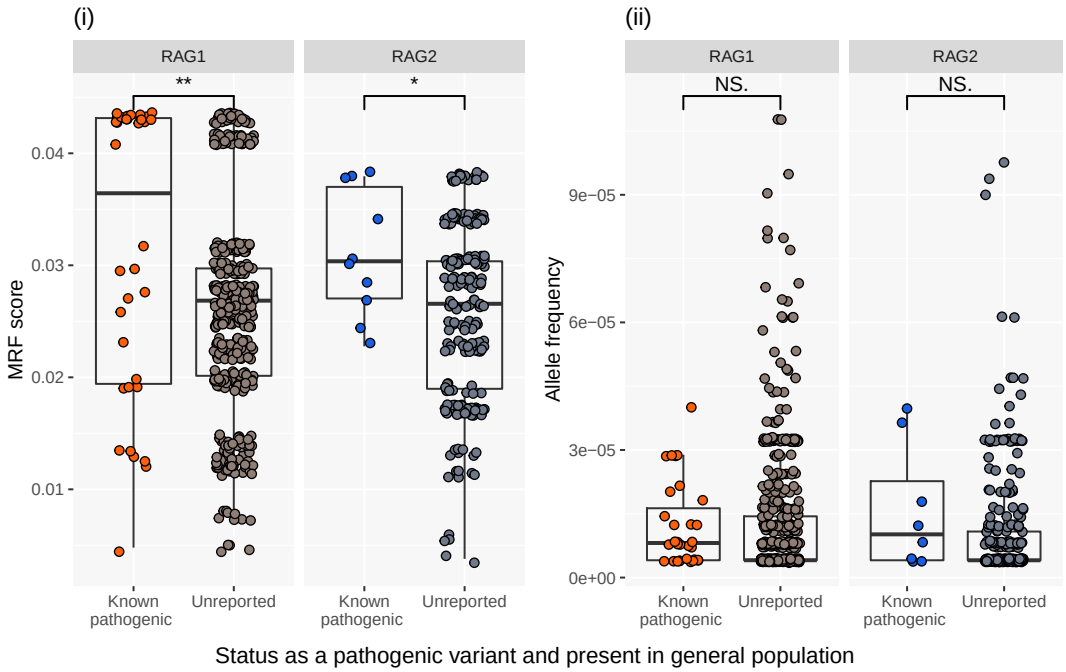

Figure 2: RAG1 and RAG2 MRF score predict the likelihood of mutations that are clinically relevant. (i) Known damaging variants (clinically diagnosed with genetic confirmation) reported on GnomAD have significantly higher MRF scores than unreported variants. (ii) GnomAD rare variant allele frequency <0.0001. No significant difference in allele frequency is found between known damaging and non-clinically reported variants. Unpaired t-test. RAG1 P-value 0.002** RAG2 P-value 0.0339*. MRF; mutation rate residue frequency, ns; non-significant.

Allele frequency is generally the single most important filtering method for rare disease in whole genome (and exome) sequencing experiments. Variants under pressure from purifying selection are more likely to cause disease than common variants. However, most RAG mutations are rare. Therefore,

allele frequencies of rare variants reported on GnomAD cannot differentially predict the likelihood of causing disease (**Fig. 2 (ii)**). As such, we found no significant difference between known damaging variants and those that have not yet been reported as disease-causing. The comparison between **Figure 2 (i) and (ii)** illustrates the reasoning for the design of our method.

Many non-clinically-reported rare variants may cause disease; the MRF score identifies the top clinically relevant candidates. Based on the frequency of protein-truncating variants in the general population, *RAG1* and *RAG2* are considered to be tolerant to the loss of one allele, as indicated by their low probability of being loss-of-function intolerant (pLI) scores of 0.00 and 0.01, respectively [21]. This is particularly important for recessive diseases such as RAG deficiency where most new missense variants will be of unknown significance until functionally validated.

Top candidate variants require validation

Functionally characterising protein activity is both costly and time consuming. RAG1 and RAG2 have now been investigated by multiple functional assays for at least 110 coding variants [16–18]. In each case, researchers selected variants in *RAG1* and *RAG2* that were potentially damaging or were identified from PID patients as the most probable genetic determinant of disease. Functional assays for RAG deficiency in those cases, and generally, measured a loss of recombination activity as a percentage of wild type function (0-100%).

Pre-emptively performing functional variant studies benefits those who will be identified with the same variants in the future, before the onset of disease complications. While more than 100 variants have been assayed in vitro, we calculated that only one-quarter of them are most probable candidates for clinical presentation. **Figure 3** illustrates that while functional work targeted "hand picked" variants that were ultimately confirmed as damaging, many of them may be unlikely to arise based on population genetics data. **Figure 3** presents, in increasing order, the number of potential variants based on likelihood of presentation and stacked by the number of variants per score category. Variants that have been measured for their loss of protein activity are coloured by severity. Potential variants that remain untested are coloured in grey. Only 21 of the top 66 most probable clinically relevant variants have been assayed in *RAG1*.

Supplemental Figure **E3** further illustrates the individual variants which have been tested functionally (the coloured *recombination activity* subset of Fig 3). We compared predicted MRF scores to assay measurements for 71 *RAG1* and 39 *RAG2* mutants. Most mutations tested showed severe loss of protein function (bottom panel of Supplemental Figure **E3**), while the likelihood each mutation occurring in humans varied significantly (top panels).

If MRF scoring was used in the same cases pre-emptively, the loss of investment would be minimal; only 8 variants out of 71 mutants tested had an above-average MRF score while being measured as

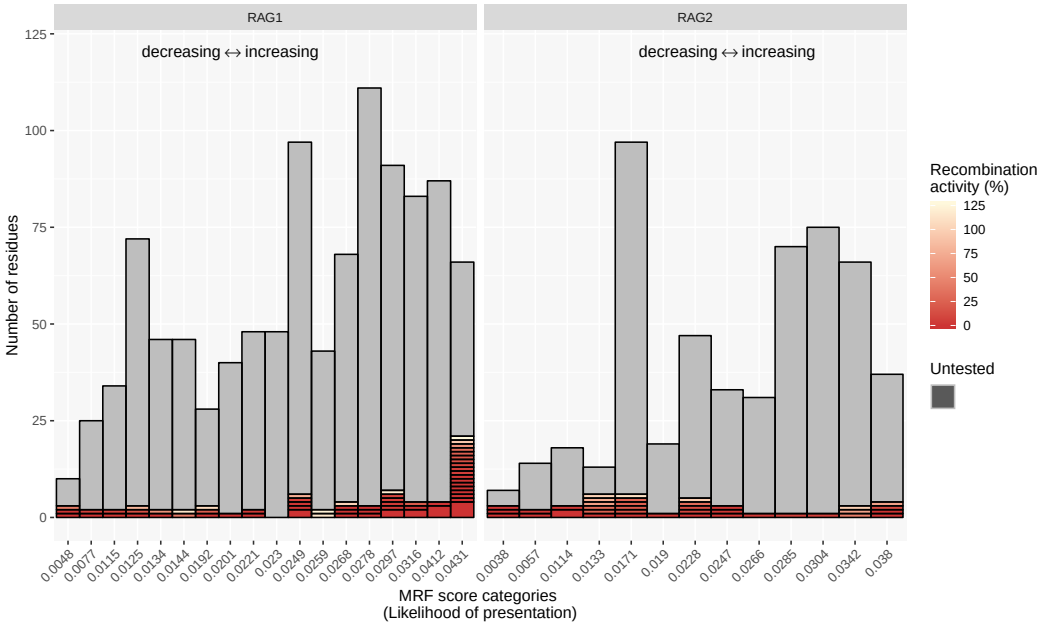

Figure 3: RAG1 and RAG2 MRF score categories and variants assayed to date. Protein residues are ranked and stacked into categories based on their MRF score. High scores (0.043 and 0.038 in RAG1 and RAG2, respectively) represent a greater mutation likelihood. Functional assays have measured recombination activity (as its inverse; % loss of activity) in a total of 110 mutants. The severity of protein loss of function is represented by a red gradient. Residues that have not been functionally tested are shown in grey. While many protein residues are critical to protein function, their mutation is less probable than many of the top MRF candidates. Data further expanded in Figure E3. MRF; mutation rate residue frequency.

functionally benign (a rate of 11.27%). RAG2 had only 3 out of 39 variants (7.69%) with an above-average MRF score while functionally benign. For the expended resources, approximately 30% more top candidates would have been tested in place of unlikely and functionally non-damaging mutations. However, the true measurement of accuracy is limited in that very few of the most likely clinically relevant variants predicted by MRF scoring have been tested to date.

False positives in Transib domains do not negatively impact prediction

Adaptive immunity is considered to have evolved through jawed vertebrates after integration of the RAG transposon into an ancestral antigen receptor gene [55, 56]. The *Transib* transposon is a 600 amino acid core region of RAG1 that targets RSS-like sequences in many invertebrates. A linked *RAG1/RAG2* was shown in the lower dueterosome (sea urchin), indicating an earlier common ancestor than the invertebrate [57], and more recently, a recombinatorially active RAG transposon (ProtoRAG) was found in the lower chordate amphioxus (or lancelet); the most basal extant chordate and a “living fossil of RAG” [58].

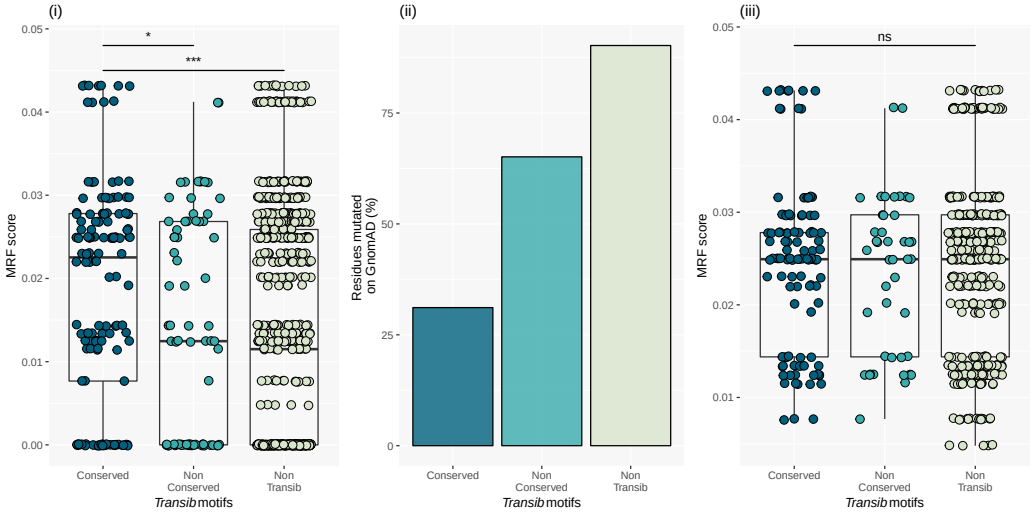

Figure 4: False positives in *Transib* domains do not worsen probability prediction. The *Transib* domains contain critical conserved protein residues. (i) False positives were simulated by scoring *Transib* domain MRF without omitting Boolean conservation weight $C = 0$. (ii) Allele frequencies on GnomAD had conservation levels inversely proportional to simulated false-positive MRF scoring. (iii) When testing for all Boolean component $C > 0$ after MRF calculation the effect of false positives remained non-significant, illustrating the non-negative impact of MRF for prediciting the mutation. Unpaired t-test, * P = 0.0195, *** P < 0.0001. MRF; mutation rate residue frequency, ns; non-significant.

A set of conserved motifs in core *RAG1* are shared with the *Transib* transposase, including the critical DDE residue catalytic triad (residues 603, 711, and 965) [59]. Ten *RAG1* core motifs are conserved amongst a set of diverse species including human [59]. This evolutionarily conserved region is considered as most important to protein function. Therefore, we chose this region to determine if MRF scoring would have a negative impact if mutations were falsely predicted as clinically important. To assess the influence of a false positive effect on prediction, the MRF scores for conserved residues in this group were compared to GnomAD allele frequencies. **Figure 4 (i)** plots the MRF (without omitting the Boolean component $C = 0$) for conserved *Transib* motif residues, non-conserved *Transib* motif residues, and non-*Transib* residues. **Figure 4 (ii)** shows the percentage of these which were reported as mutated on GnomAD. By accounting for unreported variants by applying $C > 0$, the resulting effect on incorrectly scoring MRF in the conserved *Transib* motifs remained neutral.

MRF predicts RAG deficiency amongst PID patients harbouring rare variants

We have previously measured the recombination activity of RAG1 and RAG2 disease-causing variants in several patients [16]. We have compiled our own and other functional assay data from Lee et al. [17]

and Tirosh et al. [18] to produce a panel of recombination activity measurements for coding variants in both *RAG1* and *RAG2*. RAG deficiency was measured as the level of recombination potential produced by the protein complex. Each method of investigation simulated the efficiency of wild-type or mutant proteins expressed by patients for their ability to produce a diverse repertoire of T-cell receptor (TCR) and B-cell receptor (BCR) and coding for immunoglobulins. In functional experiments, mutant proteins were assayed for their ability to perform recombination on a substrate which mimics the RSS of TCR and BCR in comparison to wild-type protein complex (as % SEM).

By gathering confirmed RAG deficiency cases, we compiled the MRF scores for 43 damaging *RAG1* variants in 77 PID cases and 14 damaging *RAG2* variants in 21 PID cases (MRF scores spanning over 22 categories). To test our method against a strong control group, we identified coding variants in patients with PID where RAG deficiency due to coding variants has been ruled out as the cause of disease. We obtained *RAG1/2* variants in 558 PID patients who had their genomes sequenced as part of the NIHR BioResource - Rare Diseases study [16]. Filtering initially identified 32 variants in 166 people. This set was trimmed to contain only rare variants; 29 variants over 26 MRF scoring categories from 72 cases of non-RAG-deficient PID. Linear regression on this control group produced negative or near-zero slopes for *RAG1* and *RAG2*, respectively. The same analysis for known-damaging mutations in disease cases had a significant prediction accuracy for *RAG1*. Analysis for *RAG2* was not significant. However, the sample size to date may be too small to significantly measure *RAG2* MRF scoring although a positive correlation was inferred in **Figure 5** [60]. R source and raw data can be found in supplemental material.

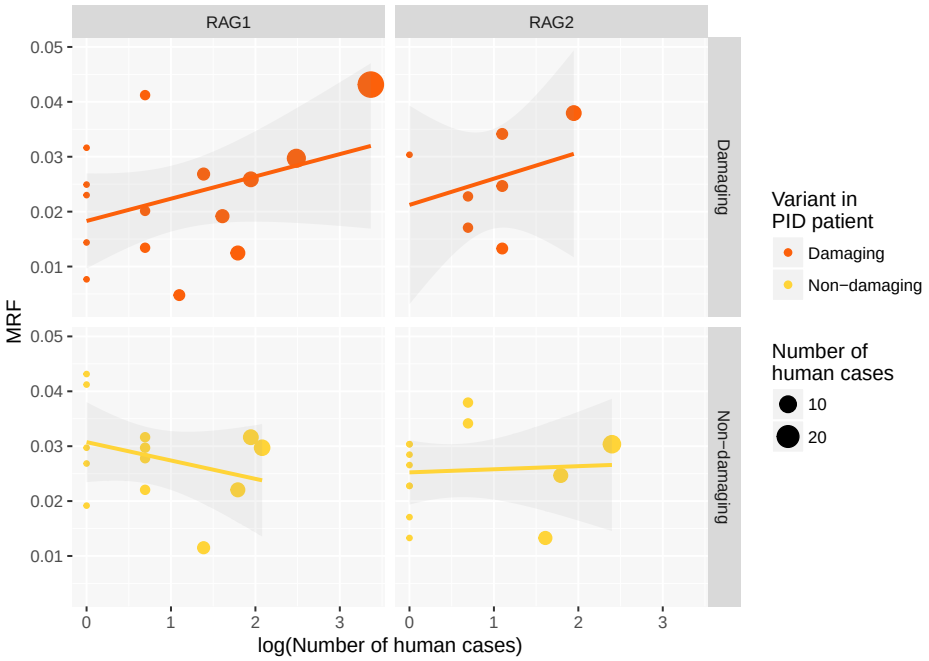

Figure 5: A linear regression model of RAG1/2 MRF scoring in cases of primary immune deficiency. MRF prediction correlates with clinical presentation. Damaging variants identified in confirmed RAG deficiency cases. Non-damaging variants sourced from cases of PID with rare variants but not responsible for disease. (Slopes of RAG1: Damaging: 0.0008* ($\pm$ 0.0004) P<0.05, intercept 5.82e-05 ***, Non-damaging: -0.0007 ($\pm$ 0.001). Slopes of RAG2; Damaging: 0.0023 ($\pm$ 0.0018), intercept 0.0312 *, Non-damaging 0.0001 ($\pm$ 0.0008). Source data and script in supplemental material).

MRF supplements pathogenicity prediction tools for translational research

CADD scoring [25] is an important bioinformatics tool that exemplifies pathogenicity prediction. While CADD is a valuable scoring method, its purpose is not to predict likelihood of variation. Similarly, MRF scoring is not a measure of pathogenicity. MRF scoring may be complemented by tools for scoring variant deleteriousness. We compare MRF to the PHRED-scaled CADD scores for all possible SNV positions in *RAG1* (**Fig. 6**) illustrating that pathogenicity prediction cannot account for mutation probability. Combining both methods allows researchers to identify highly probable mutations before querying predicted pathogenicity.

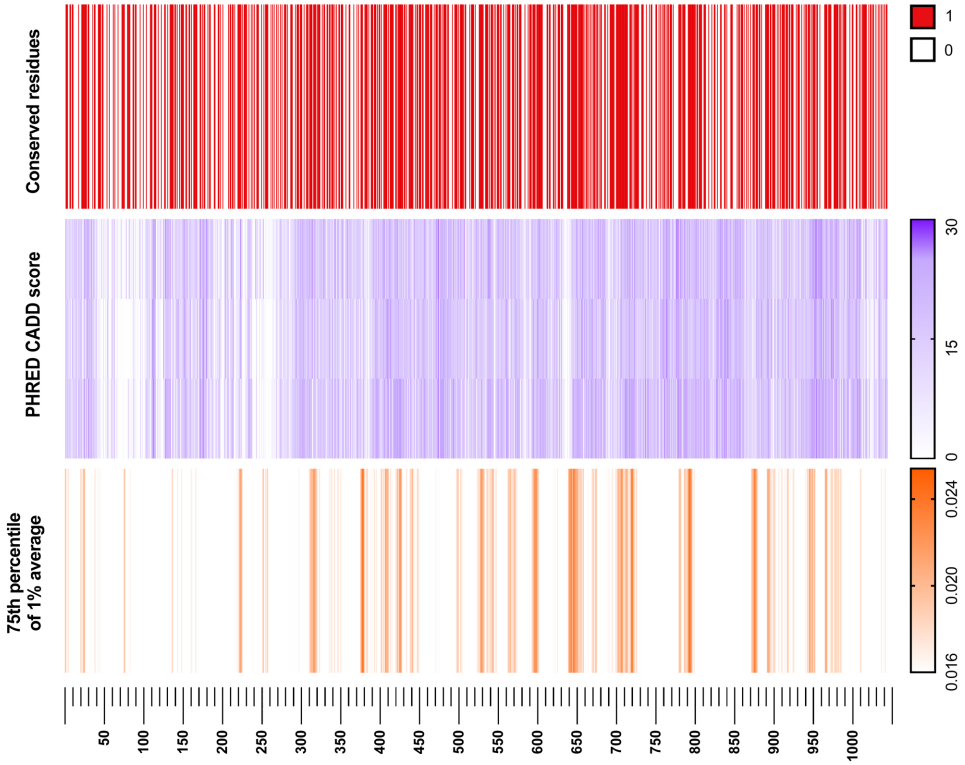

Figure 6: *RAG1* PHRED-scaled CADD score versus GnomAD conservation rate and MRF score. Allele frequency conservation rate (**top**) is vastly important for identifying critical structural and functional protein regions. The impact of mutation in one of these conserved regions is often estimated using CADD scoring (**middle**). CADD score heatmap is aligned by codon and separated into three layers for individual nucleotide positions. The MRF score (**bottom**)(visualised using the 75th percentile with 1% averaging) highlights protein regions which are most likely to present clinically and may require pre-emptive functional investigation.

To further develop this concept, we first annotated variants with MRF likelihood scores and

pathogenic prediction PHRED-scaled CADD scores (**Figure 7**), and secondly, performed a manual investigation of the clinical relevance of top candidates (**Table E2**). We used HGMD as an unbiased source of known RAG deficiency cases in both instances. CADD score was very successful at predicting the pathogenicity of a variant, (a high-density cluster of variants with CADD scores >25) as shown in **red** in **Figure 7 (i)**. At about the same rate, CADD score also predicted variants as pathogenic that are, to date, unreported (as **pink** in **Fig. 7 (i)**). Indeed, those unreported variants may very well be pathogenic. However, the likelihood of each mutation varies. As such, we developed the MRF score to account for that likelihood. As expected, the likelihood of mutations occurring that were unreported was low according to MRF (**Fig. 7 (ii)**, **pink**), while the mutations which did occur were highly enriched in at high MRF scores (**Fig. 7 (ii)**, **red** high-density cluster >0.043). Combining mutation prediction (MRF) with pathogenicity prediction (tools like CADD) increases the accuracy of pre-emptively targeting clinically relevant variants. **Figure 7 (iii)** shows that while the number of variants presented to date is relatively small, they already account for 36% of the top MRF score candidates.

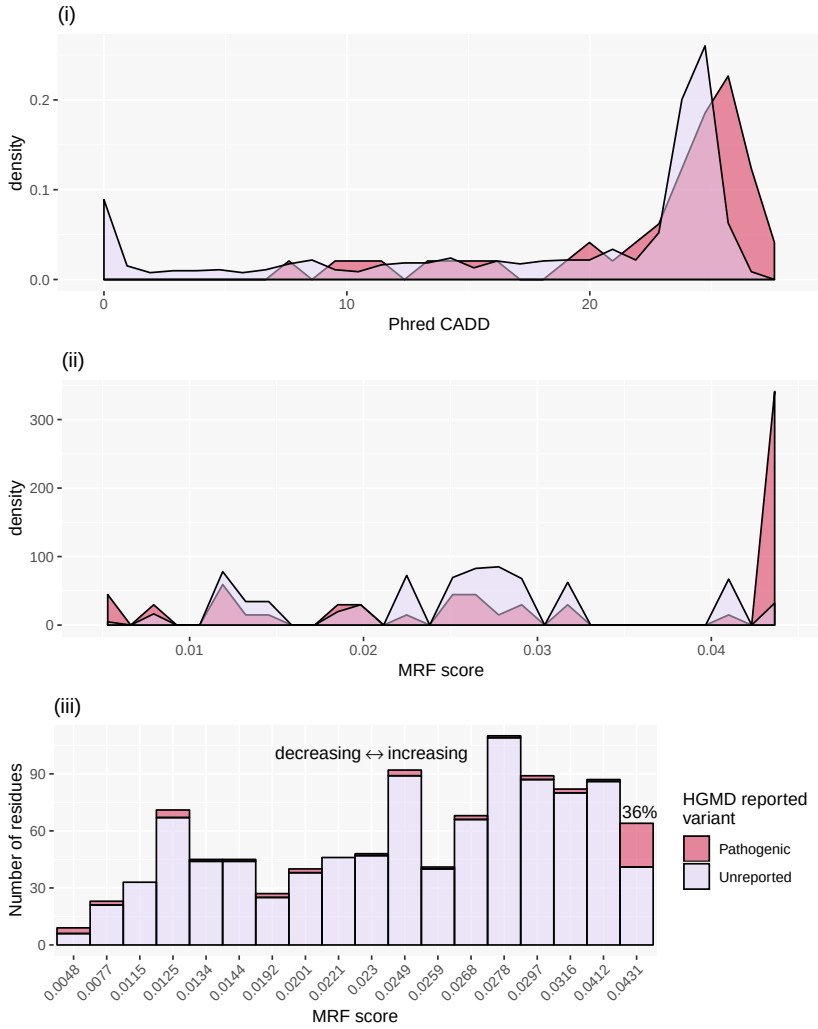

Figure 7: *RAG1* PHRED-scaled CADD score versus MRF score against HGMD data. (i) A high CADD score is a predictor of deleteriousness. Both reported (red) and non-reported residues (pink) have a high density of high CADD score. (ii) MRF scores only show a high-density cluster for high-likelihood variants, reflected by the high MRF score observed for known RAG deficiency variants. The number of pathogenic variants is outweighed by conserved residues; (i-ii) shows density of scores to normalise between groups. AUC overlap difference in CADD score of 21.43% and MRF score of 74.28% (above intersects >22.84 and >0.0409, in *(i-ii)* respectively). (iii) The number of residues per MRF category shows that disease reported on HGMD accounts for 36% of top MRF candidates. AUC; area under curve, CADD; Combined Annotation Dependent Depletion, HGMD; Human Gene Mutation Database.

Discussion

Determining disease-causing variants for functional analysis typically aims to target conserved gene regions. On GnomAD 56% of *RAG1* (approx. 246,000 alleles) is conserved with no reported variants. Functional validation of unknown variants in genes with this level purifying selection is generally infeasible. Furthermore, we saw that a vast number of candidates are "predicted pathogenic" by commonly used pathogenicity tools, which may indeed be damaging but unlikely to occur. To overcome the challenge of manual selection we quantified the likelihood of mutation for each candidate variant.

Targeting clearly defined regions with high MRF scores allows for functional validation studies tailored to the most clinically relevant protein regions. An example of high MRF score clustering occurred in the RAG1 catalytic RNase H (RNH) domain at p.Ser638-Leu658 which is also considered a conserved *Transib* motif.

While many hypothetical variants with low MRF scores may be uncovered as functionally damaging, our findings suggest that human genomic studies will benefit by first targeting variants with the highest probability of occurrence (gene regions with high MRF). **Table E1** lists the values for calculated MRFs for RAG1 and RAG2.

We have presented a basic application of MRF scoring for RAG deficiency. The method can be applied genome-wide. This can include phenotypically derived weights to target candidate genes or tissue-specific epigenetic features. In the state presented here, MRF scores are used for pre-clinical studies. A more advanced development may allow for use in single cases. During clinical investigations using personalised analysis of patient data, further scoring methods may be applied based on disease features. A patient phenotype can contribute a weight based on known genotype correlations separating primary immunodeficiencies or autoinflammatory diseases [6]. For example, a patient with autoinflammatory features may require a selection that favors genes associated with proinflammatory disease such as *MEFV* and *TNFAIP3*, whereas a patient with mainly immunodeficiency may have preferential scoring for genes such as *BTK* and *DOCK8*. In this way, a check-list of most likely candidates can be confirmed or excluded by whole genome or panel sequencing. However, validation of these expanded implementations requires a deeper consolidation of functional studies than is currently available.

Havrilla et al. [61] have recently developed a method with similar possible applications for human health mapping constrained coding regions. Their study employed a method that included weighting by sequencing depth. Similarly, genome-wide scoring may benefit from mutation significance cut-off, which is applied for tools such as CADD, PolyPhen-2, and SIFT [62]. We have not included an adjustment method as our analysis was gene-specific but implementation is advised when calculating genome-wide MRF scores.

The MRF score was developed to identify the top most probable variants that have the potential to cause disease. It is not a predictor of pathogenicity. However, MRF may contribute to disease prediction;

a clinician may ask for the likelihood of RAG deficiency (or any other Mendelian disease of interest) prior to examination (**Supplemental**).

Predicting the likelihood of discovering novel mutations has implications in genome-wide association studies (GWAS). Variants with low minor allele frequencies have a low discovery rate and low probability of disease association [63], an important consideration for rare diseases such as RAG deficiency. An analysis of the NHGRI-EBI catalogue data highlighted diseases whose average risk allele frequency was low [63]. Autoimmune diseases had risk allele frequencies considered low at approximately 0.4. Without a method to rank most probable novel disease-causing variants, it is unlikely that GWAS will identify very rare disease alleles (with frequencies <0.001). It is conceivable that a number of rare immune diseases are attributable to polygenic rare variants. However, evidence for low-frequency polygenic compounding mutations will not be available until large, accessible genetics databases are available, exemplified by the NIHR BioResource Rare Diseases study [16]. An interesting consideration when predicting probabilities of variant frequency, is that of protective mutations. Disease risk variants are quelled at low frequency by negative selection, while protective variants may drift at higher allele frequencies [64].

The cost-effectiveness of genomic diagnostic tests is already outperforming traditional, targeted sequencing [1]. Even with substantial increases in data sharing capabilities and adoption of clinical genomics, rare diseases due to variants of unknown significance and low allele frequencies will remain non-actionable until reliable predictive genomics practices are developed. Bioinformatics as a whole has made staggering advances in the field of genetics [65]. Challenges that remain unsolved, hindering the benefit of national or global genomics databases, include DNA data storage and random access retrieval [66], data privacy management [67], and predictive genomics analysis methods. Variant filtration in rare disease is based on reference allele frequency, yet the result is not clinically actionable in many cases. Development of predictive genomics tools may provide a critical role for single patient studies and timely diagnosis [23].

Conclusion

We provide a list of amino acid residues for RAG1 and RAG2 that have not been reported to date, but are most likely to present clinically as RAG deficiency. This method may be applied to other diseases with hopes of improving preparedness for clinical diagnosis.

Authorship Contributions

Dylan Lawless analysed data and conceived and wrote the manuscript, Hana Allen Lango analysed NIHR genomic data, James Thaventhiran provided genomic data, NIHR BioResource–Rare Diseases

Consortium provided genomic data, Rashida Anwar wrote the manuscript, Jolan E. Walter provided clinical genomics data and wrote the manuscript, Jacques Fellay wrote the manuscript, Sinisa Savic conceived and wrote the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Katherine Payne, Sean P Gavan, Stuart J Wright, and Alexander J Thompson. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nature Reviews Genetics*, 2018.

[2] Antonia Kwan, Roshini S Abraham, Robert Currier, Amy Brower, Karen Andruszewski, Jordan K Abbott, Mei Baker, Mark Ballow, Louis E Bartoshesky, Vincent R Bonagura, et al. Newborn screening for severe combined immunodeficiency in 11 screening programs in the united states. *Jama*, 312(7):729–738, 2014.

[3] L. Alexander Liggett, Anchal Sharma, Subhajyoti De, and James DeGregori. Conserved patterns of somatic mutations in human peripheral blood cells. *bioRxiv*, 2017. doi: 10.1101/208066.

[4] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):781, 2018.

[5] István Bartha, Julia di Iulio, J Craig Venter, and Amalio Telenti. Human gene essentiality. *Nature Reviews Genetics*, pages nrg–2017, 2017.

[6] Capucine Picard, H Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Yanick J Crow, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, et al. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. *Journal of clinical immunology*, 38(1):96–128, 2018.

[7] Mary Ellen Conley and Jean-Laurent Casanova. Discovery of single-gene inborn errors of immunity by next generation sequencing. *Current opinion in immunology*, 30:17–23, 2014.

[8] David G Schatz, Marjorie A Oettinger, and David Baltimore. The v (d) j recombination activating gene, rag-1. *Cell*, 59(6):1035–1048, 1989.

[9] Marjorie A Oettinger, David G Schatz, Carolyn Gorka, and David Baltimore. Rag-1 and rag-2, adjacent genes that synergistically activate v (d) j recombination. *Science*, 248(4962):1517–1523, 1990.

[10] Peter Mombaerts, John Iacomini, Randall S Johnson, Karl Herrup, Susumu Tonegawa, and Virginia E Papaioannou. Rag-1-deficient mice have no mature b and t lymphocytes. *Cell*, 68(5): 869–877, 1992.

[11] Yoichi Shinkai, Kong-Peng Lam, Eugene M Oltz, Valerie Stewart, Monica Mendelsohn, Jean Charron, Milton Datta, Faith Young, Alan M Stall, Frederick W Alt, et al. Rag-2-deficient mice lack mature lymphocytes owing to inability to initiate v (d) j rearrangement. *Cell*, 68(5):855–867, 1992.

[12] Klaus Schwarz, George H Gauss, Leopold Ludwig, Ulrich Pannicke, Zhong Li, Doris Lindner, Wilhelm Friedrich, Reinhard A Seger, Thomas E Hansen-Hagge, Stephen Desiderio, et al. Rag mutations in human b cell-negative scid. *Science*, 274(5284):97–99, 1996.

[13] G de Saint-Basile, F Le Deist, JP De Villartay, N Cerf-Bensussan, O Journet, N Brousse, C Griscelli, and A Fischer. Restricted heterogeneity of t lymphocytes in combined immunodeficiency with hypereosinophilia (omenn's syndrome). *The Journal of clinical investigation*, 87(4):1352–1359, 1991.

[14] Frédéric Rieux-Laucat, Philippe Bahadoran, Nicole Brousse, Françoise Selz, Alain Fischer, Françoise Le Deist, and Jean Pierre De Villartay. Highly restricted human t cell repertoire in peripheral blood and tissue-infiltrating lymphocytes in omenn's syndrome. *The Journal of clinical investigation*, 102(2):312–321, 1998.

[15] Anna Villa, Sandro Santagata, Fabio Bozzi, Silvia Giliani, Annalisa Frattini, Luisa Imberti, Luisa Benerini Gatta, Hans D Ochs, Klaus Schwarz, Luigi D Notarangelo, et al. Partial v (d) j recombination activity leads to omenn syndrome. *Cell*, 93(5):885–896, 1998.

[16] Dylan Lawless, Christoph B Geier, Jocelyn R Farmer, Hana Lango Allen, Daniel Thwaites, Faranaz Atschekzei, Matthew Brown, David Buchbinder, Siobhan O Burns, Manish J Butte, et al. Prevalence and clinical challenges among adults with primary immunodeficiency and recombination-activating gene deficiency. *Journal of Allergy and Clinical Immunology*, 2018.

[17] Yu Nee Lee, Francesco Frugoni, Kerry Dobbs, Jolan E Walter, Silvia Giliani, Andrew R Gennery, Waleed Al-Herz, Elie Haddad, Francoise LeDeist, Jack H Bleesing, et al. A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. *Journal of Allergy and Clinical Immunology*, 133(4):1099–1108, 2014.

[18] Irit Tirosh, Yasuhiro Yamazaki, Francesco Frugoni, Francesca A Ververs, Eric J Allenspach, Yu Zhang, Siobhan Burns, Waleed Al-Herz, Lenora Noroski, Jolan E Walter, et al. Recombination activity of human rag2 mutations and correlation with the clinical phenotype. *Journal of Allergy and Clinical Immunology*, 2018.

[19] Jolan E Walter, Lindsey B Rosen, Krisztian Csomos, Jacob M Rosenberg, Divij Mathew, Marton Keszei, Boglarka Ujhazi, Karin Chen, Yu Nee Lee, Irit Tirosh, et al. Broad-spectrum antibodies against self-antigens and cytokines in rag deficiency. *The Journal of clinical investigation*, 125 (11):4135–4148, 2015.

[20] Catharina Schuetz, Kirsten Huck, Sonja Gudowius, Mosaad Megahed, Oliver Feyen, Bernd Hubner, Dominik T Schneider, Burkhard Manfras, Ulrich Pannicke, Rein Willemze, et al. An immunodeficiency disease with rag mutations and granulomas. *New England Journal of Medicine*, 358(19):2030–2038, 2008.

[21] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[22] Tami John, Jolan E Walter, Catherina Schuetz, Karin Chen, Roshini S Abraham, Carmem Bonfim, Thomas G Boyce, Avni Y Joshi, Elizabeth Kang, Beatriz Tavares Costa Carvalho, et al. Unrelated hematopoietic cell transplantation in a patient with combined immunodeficiency with granulomatous disease and autoimmunity secondary to rag deficiency. *Journal of clinical immunology*, 36(7): 725–732, 2016.

[23] Jean-Laurent Casanova, Mary Ellen Conley, Stephen J Seligman, Laurent Abel, and Luigi D Notarangelo. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *Journal of Experimental Medicine*, pages jem–20140520, 2014.

[24] Paul Stothard. The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *University of Alberta, Education and Research Archive*, 2000.

[25] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[26] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Shaw, Andrew D Phillips, and David N Cooper. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics*, 133(1):1–9, 2014.

[27] Heng Ru, Melissa G Chambers, Tian-Min Fu, Alexander B Tong, Maofu Liao, and Hao Wu. Molecular mechanism of v (d) j recombination from synaptic rag1-rag2 complex structures. *Cell*, 163(5):1138–1152, 2015.

[28] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[29] John Stone. An efficient library for parallel ray tracing and animation. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[30] Anna Villa, Cristina Sobacchi, Luigi D Notarangelo, Fabio Bozzi, Mario Abinun, Tore G Abrahamsen, Peter D Arkwright, Michal Baniyash, Edward G Brooks, Mary Ellen Conley, et al. V (d) j recombination defects in lymphocytes due torag mutations: severe immunodeficiency with a spectrum of clinical presentations. *Blood*, 97(1):81–88, 2001.

[31] Hassan Abolhassani, Ning Wang, Asghar Aghamohammadi, Nima Rezaei, Yu Nee Lee, Francesco Frugoni, Luigi D Notarangelo, Qiang Pan-Hammarström, and Lennart Hammarström. A hypomorphic recombination-activating gene 1 (rag1) mutation resulting in a phenotype resembling common variable immunodeficiency. *Journal of Allergy and Clinical Immunology*, 134(6):1375–1380, 2014.

[32] Necil Kutukculer, Nesrin Gulez, Neslihan Edeer Karaca, Guzide Aksu, and Afig Berdeli. Novel mutatıons and diverse clinical phenotypes in recombınase-activating gene 1 deficiency. *Italian journal of pediatrics*, 38(1):8, 2012.

[33] Cristina Sobacchi, Veronica Marrella, Francesca Rucci, Paolo Vezzoni, and Anna Villa. Rag-dependent primary immunodeficiencies. *Human mutation*, 27(12):1174–1184, 2006.

[34] Jeroen G Noordzij, Sandra de Bruin-Versteeg, Nicole S Verkaik, Jaak MJJ Vossen, Ronald de Groot, Ewa Bernatowska, Anton W Langerak, Dik C van Gent, and Jacques JM van Dongen. The immunophenotypic and immunogenotypic b-cell differentiation arrest in bone marrow of rag-deficient scid patients corresponds to residual recombination activities of mutated rag proteins. *Blood*, 100(6):2145–2152, 2002.

[35] Elena Crestani, Sharon Choo, Francesco Frugoni, Yu Nee Lee, Stephanie Richards, Joanne Smart, and Luigi D Notarangelo. Rag1 reversion mosaicism in a patient with omenn syndrome. *Journal of clinical immunology*, 34(5):551–554, 2014.

[36] Ilan Dalal, Uri Tabori, Bela Bielorai, Hana Golan, Eli Rosenthal, Ninette Amariglio, Gidi Rechavi, and Amos Toren. Evolution of a tb-scid into an omenn syndrome phenotype following parainfluenza 3 virus infection. *Clinical Immunology*, 115(1):70–73, 2005.

[37] Taco W Kuijpers, Hanna IJspeert, Ester MM van Leeuwen, Machiel H Jansen, Mette D Hazenberg, Kees C Weijer, Rene AW Van Lier, and Mirjam van der Burg. Idiopathic cd4+ t lymphopenia without autoimmunity or granulomatous disease in the slipstream of rag mutations. *Blood*, pages blood–2011, 2011.

[38] Tanja A Gruber, Ami J Shah, Michelle Hernandez, Gay M Crooks, Hisham Abdel-Azim, Sudhir Gupta, Sean McKnight, Drew White, Neena Kapoor, and Donald B Kohn. Clinical and genetic heterogeneity in omenn syndrome and severe combined immune deficiency. *Pediatric transplantation*, 13(2):244–250, 2009.

[39] Suk See De Ravin, Edward W Cowen, Kol A Zarember, Narda L Whiting-Theobald, Douglas B Kuhns, Netanya G Sandler, Daniel C Douek, Stefania Pittaluga, Pietro L Poliani, Yu Nee Lee, et al. Hypomorphic rag mutations can cause destructive midline granulomatous disease. *Blood*, 116(8): 1263–1271, 2010.

[40] David Buchbinder, Rebecca Baker, Yu Nee Lee, Juan Ravell, Yu Zhang, Joshua McElwee, Diane Nugent, Emily M Coonrod, Jacob D Durtschi, Nancy H Augustine, et al. Identification of patients with rag mutations previously diagnosed with common variable immunodeficiency disorders. *Journal of clinical immunology*, 35(2):119–124, 2015.

[41] Kerstin Felgentreff, Ruy Perez-Becker, Carsten Speckmann, Klaus Schwarz, Krzysztof Kalwak, Gasper Markelj, Tadej Avcin, Waseem Qasim, EG Davies, Tim Niehues, et al. Clinical and immunological manifestations of patients with atypical severe combined immunodeficiency. *Clinical immunology*, 141(1):73–82, 2011.

[42] Andreas Reiff, Alexander G Bassuk, Joseph A Church, Elizabeth Campbell, Xinyu Bing, and Polly J Ferguson. Exome sequencing reveals rag1 mutations in a child with autoimmunity and sterile chronic multifocal osteomyelitis evolving into disseminated granulomatous disease. *Journal of clinical immunology*, 33(8):1289–1292, 2013.

[43] Barbara Corneo, Despina Moshous, Tayfun Güngör, Nicolas Wulffraat, Pierre Philippet, Françoise Le Deist, Alain Fischer, and Jean-Pierre de Villartay. Identical mutations in rag1 or rag2 genes leading to defective v (d) j recombinase activity can cause either tb–severe combined immune deficiency or omenn syndrome. *Blood*, 97(9):2772–2776, 2001.

[44] Erika Asai, Taizo Wada, Yasuhisa Sakakibara, Akiko Toga, Tomoko Toma, Takashi Shimizu, Sheela Nampoothiri, Kohsuke Imai, Shigeaki Nonoyama, Tomohiro Morio, et al. Analysis of mutations and recombination activity in rag-deficient patients. *Clinical Immunology*, 138(2): 172–177, 2011.

[45] Tamaki Kato, Elena Crestani, Chikako Kamae, Kenichi Honma, Tomoko Yokosuka, Takeshi Ikegawa, Naonori Nishida, Hirokazu Kanegane, Taizo Wada, Akihiro Yachie, et al. Rag1 deficiency may present clinically as selective iga deficiency. *Journal of clinical immunology*, 35(3):280–288, 2015.

[46] Xiaomin Yu, Jorge R Almeida, Sam Darko, Mirjam van der Burg, Suk See DeRavin, Harry Malech, Andrew Gennery, Ivan Chinn, Mary Louise Markert, Daniel C Douek, et al. Human syndromes of immunodeficiency and dysregulation are characterized by distinct defects in t-cell receptor repertoire development. *Journal of Allergy and Clinical Immunology*, 133(4):1109–1115, 2014.

[47] Jean-Pierre De Villartay, Annick Lim, Hamoud Al-Mousa, Sophie Dupont, Julie Déchanet-Merville, Edith Coumau-Gatbois, Marie-Lise Gougeon, Arnaud Lemainque, Céline Eidenschenk, Emmanuelle Jouanguy, et al. A novel immunodeficiency associated with hypomorphic rag1 mutations and cmv infection. *The Journal of clinical investigation*, 115(11):3291–3299, 2005.

[48] Junyan Zhang, Linda Quintal, Adelle Atkinson, Brent Williams, Eyal Grunebaum, and Chaim M Roifman. Novel rag1 mutation in a case of severe combined immunodeficiency. *Pediatrics*, 116 (3):e445–e449, 2005.

[49] Lauren A Henderson, Francesco Frugoni, Gregory Hopkins, Helen de Boer, Sung-Yun Pai, Yu Nee Lee, Jolan E Walter, Melissa M Hazen, and Luigi D Notarangelo. Expanding the spectrum of recombination-activating gene 1 deficiency: a family with early-onset autoimmunity. *Journal of Allergy and Clinical Immunology*, 132(4):969–971, 2013.

[50] Elizabeth Mannino Avila, Gulbu Uzel, Amy Hsu, Joshua D Milner, Maria L Turner, Stefania Pittaluga, Alexandra F Freeman, and Steven M Holland. Highly variable clinical phenotypes of hypomorphic rag1 mutations. *Pediatrics*, 126(5):e1248–e1252, 2010.

[51] AGL Riccetto, M Buzolin, JF Fernandes, F Traina, MLR Barjas-de Castro, MTN Silva, JB Oliveira, and MM Vilela. Compound heterozygous rag2 mutations mimicking hyper igm syndrome. *Journal of clinical immunology*, 34(1):7–9, 2014.

[52] Carlos A Gomez, Leon M Ptaszek, Anna Villa, Fabio Bozzi, Cristina Sobacchi, Edward G Brooks, Luigi D Notarangelo, Eugenia Spanopoulou, ZQ Pan, Paolo Vezzoni, et al. Mutations in conserved

regions of the predicted rag2 kelch repeats block initiation of v (d) j recombination and result in primary immunodeficiencies. *Molecular and cellular biology*, 20(15):5653–5664, 2000.

[53] Janet Chou, Rima Hanna-Wakim, Irit Tirosh, Jennifer Kane, David Fraulino, Yu Nee Lee, Soha Ghanem, Iman Mahfouz, André Mégarbané, Gérard Lefranc, et al. A novel homozygous mutation in recombination activating gene 2 in 2 relatives with different clinical phenotypes: Omenn syndrome and hyper-igm syndrome. *Journal of Allergy and Clinical Immunology*, 130(6):1414–1416, 2012.

[54] Luigi D Notarangelo, Min-Sung Kim, Jolan E Walter, and Yu Nee Lee. Human rag mutations: biochemistry and clinical implications. *Nature Reviews Immunology*, 16(4):234, 2016.

[55] Alka Agrawal, Quinn M Eastman, and David G Schatz. Transposition mediated by rag1 and rag2 and its implications for the evolution of the immune system. *Nature*, 394(6695):744, 1998.

[56] Kevin Hiom, Meni Melek, and Martin Gellert. Dna transposition by the rag1 and rag2 proteins: a possible source of oncogenic translocations. *Cell*, 94(4):463–470, 1998.

[57] Sebastian D Fugmann, Cynthia Messier, Laura A Novack, R Andrew Cameron, and Jonathan P Rast. An ancient evolutionary origin of the rag1/2 gene locus. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3728–3733, 2006.

[58] Shengfeng Huang, Xin Tao, Shaochun Yuan, Yuhang Zhang, Peiyi Li, Helen A Beilinson, Ya Zhang, Wenjuan Yu, Pierre Pontarotti, Hector Escriva, et al. Discovery of an active rag transposon illuminates the origins of v (d) j recombination. *Cell*, 166(1):102–114, 2016.

[59] Vladimir V Kapitonov and Jerzy Jurka. Rag1 core and v (d) j recombination signal sequences were derived from transib transposons. *PLoS biology*, 3(6):e181, 2005.

[60] Douglas G Altman and J Martin Bland. Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, 311(7003):485, 1995.

[61] James M Havrilla, Brent S Pedersen, Ryan M Layer, and Aaron R Quinlan. A map of constrained coding regions in the human genome. *bioRxiv*, 2017. doi: 10.1101/220814.

[62] Yuval Itan, Lei Shang, Bertrand Boisson, Michael J Ciancanelli, Janet G Markle, Ruben Martinez-Barricarte, Eric Scott, Ishaan Shah, Peter D Stenson, Joseph Gleeson, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature methods*, 13(2):109, 2016.

[63] Takashi Kido, Weronika Sikora-Wohlfeld, Minae Kawashima, Shinichi Kikuchi, Naoyuki Kamatani, Anil Patwardhan, Richard Chen, Marina Sirota, Keiichi Kodama, Dexter Hadley, et al. Are minor alleles more likely to be risk alleles? *BMC medical genomics*, 11(1):3, 2018.

[64] Yingleong Chan, Elaine T Lim, Niina Sandholm, Sophie R Wang, Amy Jayne McKnight, Stephan Ripke, Mark J Daly, Benjamin M Neale, Rany M Salem, Joel N Hirschhorn, et al. An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *The American Journal of Human Genetics*, 94(3):437–452, 2014.

[65] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321, 2015.

[66] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z. Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss. Scaling up dna data storage and random access retrieval. *bioRxiv*, 2017. doi: 10.1101/114553.

[67] Zhicong Huang, Erman Ayday, Huang Lin, Raeka S. Aiyar, Adam Molyneaux, Zhenyu Xu, Jacques Fellay, Lars M. Steinmetz, and Jean-Pierre Hubaux. A privacy-preserving solution for compressed storage and selective retrieval of genomic data. *Genome Research*, 26(12):10. 1687–1696, 2016.

[68] Attila Kumánovics, Yu Nee Lee, Devin W Close, Emily M Coonrod, Boglarka Ujhazi, Karin Chen, Daniel G MacArthur, Gergely Krivan, Luigi D Notarangelo, and Jolan E Walter. Estimated disease incidence of rag1/2 mutations: A case report and querying the exome aggregation consortium. *Journal of Allergy and Clinical Immunology*, 139(2):690–692, 2017.

Supplemental

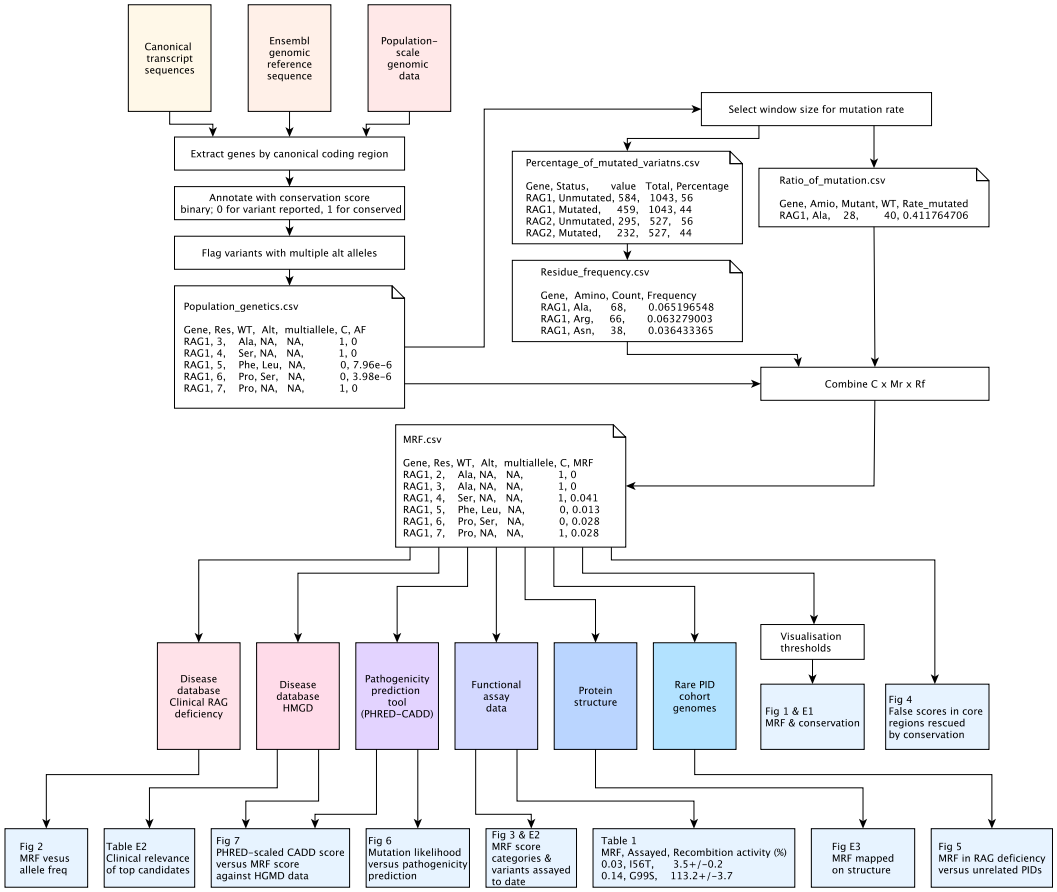

Figure E1: Data analysis summary map. Raw data and analysis scripts are provided in the supplemental. Analysis steps and data sources for each procedure described in *methods*. MRF; mutation rate residue frequency, PID; primary immunodeficiency.

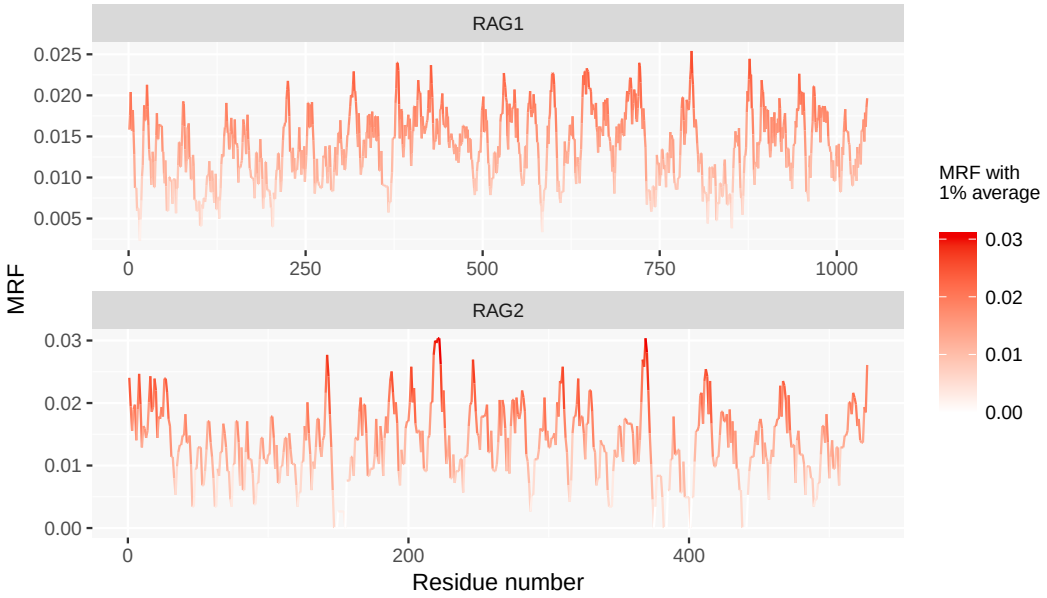

Figure E2: An alternative visualisation of MRF scores for RAG1 and RAG2 proteins. The data from Table E1 in column "Average over 1%" is displayed on both the y-axis and colour scale. An analysis-friendly long form CSV of the Table E1 data is also provided in the compressed supplemental R data "mrf.csv".

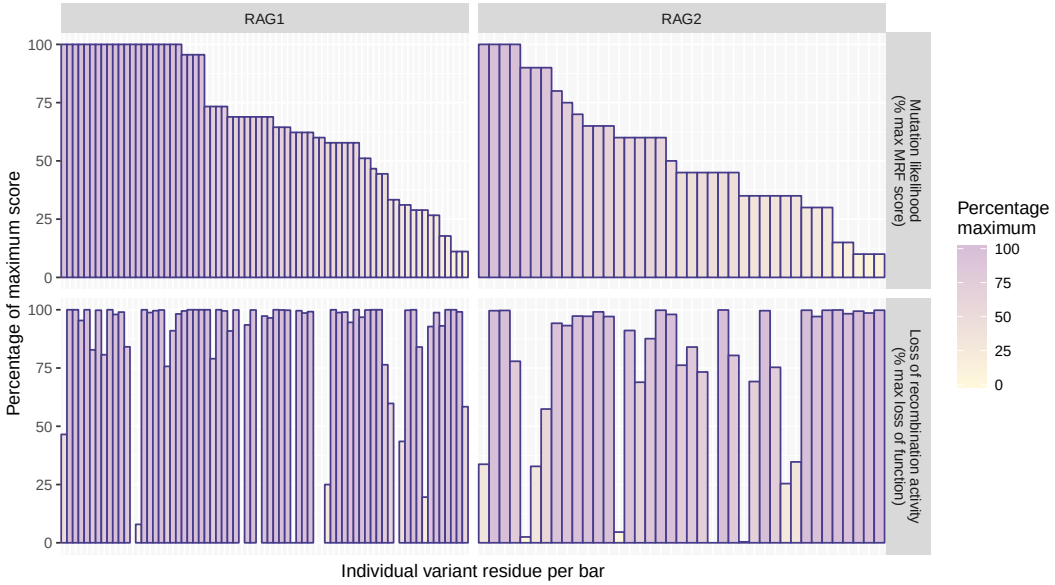

Figure E3: MRF likelihood score versus known functional activity. We compiled all variants that we know to have been assayed for protein function to date. The inverse of functional assay measurements were used, where 0% activity represents 100% loss of activity. MRF scores are presented as a percentage of the maximum score per gene (i.e., for RAG1 $MRF_{max} = 0.043$ (100%) and $MRF_{min} = 0.0048$ (0%)). **Top panels** show how likely each mutation is predicted to occur in humans. **Bottom panels** show the loss of protein activity as a percentage compared to wild-type (% SEM); most mutations tested produced severe loss of protein function, regardless of their mutation likelihood. Subset of *Recombination activity* data from Figure 3.

Main supplemental data table

Table E1: MRF data tables. The complete scores are listed for each protein. Both the wild type and alternative variants reported on GnomAD are shown [21]. Multiple alternative variants are reported for some residues which are annotated in a separate column. A column listing the 1% average of MRF scores using a sliding window is present; this is used in Figure 1 (iii) with a cutoff threshold at the 75th percentile to clearly visualise high scoring clusters. The Boolean conservation score is based on population genetics data, where 1 represents no known variant at a residue site. The percentile values for 0.9 and 0.75 are provided to allow for alternative visualisation methods using the R source supplemental *"mrf.csv"* file.

Clinical relevance of top candidates

The top scoring candidates in RAG1 were assessed for potential clinical relevance (**Table E2**). HGMD was chosen as a reliable, curated source of identifying pathogenic variants. 45% of RAG1 variants reported on HGMD (23 of 51) were predicted by our model as the most likely candidates seen clinically (the top scoring MRF group of had 66 residues total). The remaining variants in the top MRF group, which were not reported by HGMD (43 of 66), were assessed manually for their likelihood as potentially disease causing. 21 (49%) were highly conserved, not reported on GnomAD, and would be considered probable RAG deficiency on presentation as homozygous or compound heterozygous with a second damaging variant. The remainder had allele frequencies <0.0006, were only found as low frequency heterozygous in the general population, and justify functional validation. We expect that none of the top candidate mutations are benign.

Most likely mutation candidates	
Variant type	**Number variants**
Top MRF score candidates total	66
(i) Of which are reported on HGMD	23
(ii) Not reported on HGMD to date	43

Top MRF score candidates not reported on HGMD			
Frequency in Population	**Number variants**	**Unreported top candidate (%)**	**GnomAD allele frequency**
Not found	21 of 43	49%	0
Very rare	15 of 43	35%	<0.00002*
Very rare	7 of 43	16%	<0.00006**

Table E2: Clinical relevance of top candidates. 23 top MRF score variants were reported as pathogenic on HGMD to date. The remaining variants (the 43 not reported) were assessed by their frequency in population based on GnomAD (allele frequencies vary between individual variants but equate to approximately <6* and 9-77** heterozygous per 125,000 individuals). Therefore, no top candidates should be considered benign without functional validation. HGMD; Human Gene Mutation Database, MRF; mutation rate residue frequency.

Supplemental analysis tables

Table E3: Known damaging in human cases. Amino acids and residue numbers are listed along with their basic mutation rate residue frequency value for known RAG deficiency [54]. Supplemental CSV file.

Table E4: Percentage of variants per gene. Percentage of mutated versus non-mutated amino acids in RAG1 and RAG2 based on GnomAD population genetics data [21]. Supplemental CSV file.

Table E5: Residue frequency. Basic statistics for RAG1/2 were produced using SMS2 [24]. Results are shown for the canonical sequences, a 1043-residue sequence of RAG1-201 peptide ENSP00000299440 and for a 527-residue sequence of RAG2-201 peptide ENSP00000308620. Both percentage and frequency of residue usage are provided. Supplemental CSV file.

Table E6: Simplified residue frequency. The data from Table E5 is simplified for use in data analysis to only include residues, count, total, and frequency per protein expressed. Supplemental CSV file.

Table E7: Ratio of mutation per gene. The number of times each residue is mutated was found in population genetic data. The number of mutant versus wild-type is shown, from which the rate of each is derived. Supplemental CSV file.

Table E8: Basic MRF scores. The mutation rate and frequency is shown which were used to calculate the basic mutation rate residue frequency for the main analysis dataframe and Table E1. Supplemental CSV file.

Protein structure application

With the availability of a structured protein complex, modelling can be carried out prior to functional assays. Residues with the highest MRF for both RAG1 and RAG2 were mapped in **Figure E4**.

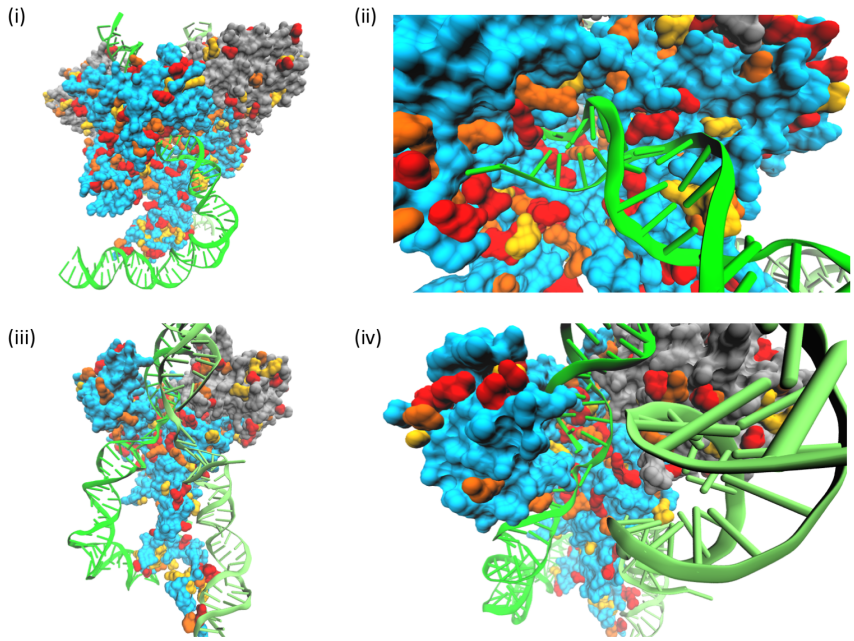

Figure E4: The RAG1 (blue) and RAG2 (grey) protein structure with top candidate MRF scores. (i) Protein dimers and (ii=iv) monomers illustrating the three highest category MRF scores for predicted clinically relevant variants. Increasing in score the top three MRF categories (illustrated in **Figure 3**) for each protein are highlighted; yellow, orange, red. DNA (green) is bound by the RAG protein complex at recombination signal sequences (PDB:3jbw).

Median CADD score per residue

The sourced PHRED-scaled CADD score data consisted of nucleotide level values. We were interested in CADD scores averaged per codon. For every nucleotide position there were three alternative variants to consider, e.g.

Chrom	Pos	Ref	Alt1	Alt2	Alt3	PHRED1	PHRED2	PHRED3
11	36594855	A	C	G	T	22.3	18.81	22.4

The PHRED-scaled scores are listed here; raw CADD scores are also included in the original database. To produce a working input we used the median score per codon, that is three scores per nucleotide and three nucleotides per codon. This produced median PHRED-scaled score per codon / residue, e.g.:

Chrom	Pos	PHRED1	PHRED2	PHRED3
11	36594855	22.3	18.81	22.4
11	36594856	25.3	23.6	24.6
11	36594857	24.8	24.3	24.5

Median PHRED = 24.3

Supplemental file *'RAG1.cadd.amino.csv'* within the analysis data *'Raw_data_R_analysis_for_figures'* contains the median values over a three-nucleotide window, starting at nucleotide 1 to produce input data with the correct reading frame. The "PHRED-scaled" values are used as a normalised and externally comparable unit of analysis, rather than raw CADD scores. The area under the curve was calculated for density plots to quantify the difference between pathogenic and unreported variants with high scores, above the intersects >0.0409 and >22.84 for MRF and CADD, respectively, using score value (x) versus density (y) (Fig. 7 (i-ii)) with $\int_a^b f(x)\,\mathrm{d}x \approx\ (b-a)\left[\frac{f(a)+f(b)}{2}\right]$.

Genome-wide and disease-specific application

Weighting data can also be applied to the MRF score model to amplify the selectivity. The mutation rate can be applied genome wide with a process common in the study of information retrieval; term frequency, inverse document frequency ($tf - idf$). In this case the "term" and "document" are replaced by amino acid residue r and gene g , respectively such that,

$$rf - igf_{r,g} = rf_{r,g} \times igf_r \tag{1}$$

We may view each gene as a vector with one component corresponding to each residue mutation in the gene, together with a weight for each component that is given by (1). Therefore, we can find the overlap

score measure with the $rf - igf$ weight of each term in g, for a query q;

$$\text{Score}(q, g) = \sum_{r \in q} \text{rf-igf}_{r,g}.$$

In respect to MRF scoring, this information retrieval method might be applied as follows; the $rf - igf$ weight of a term is the product of its rf weight and its igf weight ($W_{r,g} = rf_{r,g} \times \log \frac{N}{gf_r}$) or ($W_{r,g} = (1 + \log rf_{r,g}) \times \log \frac{N}{gf_r}$). That is, firstly, the number of times a residue mutates in a gene ($rf = rf_{r,g}$) and secondly, the rarity of the mutation genome-wide in N number of genes ($igf = N/gf_r$). Finally, ranking the score of genes for a mutation query q by;

$$\text{Score}(q, g) = \sum_{r \in q \cap g} \text{rf-igf}_{r,g}$$

The score of the query (Score(q, g)) equals the mutations (terms) that appear in both the query and the gene ($r \in q \cap g$). Working out the $rf - igf$ weight for each of those variants ($rf.igf_{r,g}$) and then summing them ($\sum$) to give the score for the specific gene with respect to the query.

Bayesian probability

MRF score may provide a limiting component required for applying Bayesian probability to disease prediction. A clinician may ask for the likelihood of RAG deficiency (or any Mendelian disease of interest) for a patient given a set of gene variants $P(H|E)$ using Bayes' theorem,

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where $P(H)$ is the probability of a patient having RAG deficiency, $P(E|H)$ is the probability of RAG deficiency due to a set of variants that have been pre-emptively assayed, and $P(E)$ is the probability of having a set of gene variants.

$P(H)$ is known since the rate of RAG deficiency is estimated at an incidence of 1:181,000 [68], SCID at a rate of 1:330,000 [2], and we also recently show the rate of RAG deficiency in adults with PID [16]. Being a recessive disease, $P(E)$ must account for biallelic variants and is the most difficult value to determine. This may be found from population genetics data for (i) the rate of two separate, compound heterozygous variants, (ii) the rate of a homozygous variant or potential consanguinity, or (iii) the rate of de novo variation [21]. $P(E|H)$ would be identified where all variants are functionally validated. This requires a major investment, however the MRF score provides a good approximation.