Characterising *RAG1* and *RAG2* with predictive genomics

Dylan Lawless[a,*], Hana Allen Lango[b,c], James Thaventhiran[d], Jolan E. Walter[e,f], Rashida Anwar[a], Sinisa Savic[g,h,*]

[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]*NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, CB20QQ, UK.*
[c]*Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB20XY, UK.*
[d]*Department of Medicine, University of Cambridge, Cambridge, UK.*
[e]*University of South Florida and Johns Hopkins All Children's Hospital, Saint Petersburg, Florida, USA.*
[f]*Division of Allergy Immunology, Massachusetts General Hospital for Children, Boston, Massachusetts, USA.*
[g]*Department of Clinical Immunology and Allergy, St James's University Hospital, Beckett Street, Leeds, UK.*
[h]*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. The structural or functional impact of a coding variant is mirrored by allele frequencies amongst the general population. Studies in protein function frequently target sites that are evolutionarily preserved. However, rare diseases are often attributable to variants in genes that are highly conserved. An immunological disorder exemplifying this challenge occurs through damaging mutations in *RAG1* and *RAG2*. RAG deficiency 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. We present variants in RAG1 and RAG2 proteins which are most likely to be seen clinically as disease-causing. Our method of mutation rate residue frequency builds a map of most probable mutations allowing pre-emptive functional analysis. We compare the accuracy of our predicted probabilities to functional measurements and provide the method for application to any monogenic disorder.

[*]Addresses for correspondence
Email addresses: D.Lawless@leeds.ac.uk (Dylan Lawless), S.Savic@leeds.ac.uk (Sinisa Savic)

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). Dr. Jolan Walter has received federal funding. The authors declare no conflict of interest.

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

RAG1, RAG2, 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), 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), $rf\text{-}igf$ (residue frequency,

inverse gene frequency), RNH (RNase H), RSS (recombination signal sequence), SCID (severe combined immunodeficiency), TCR (T-cell receptor), tf-idf (term frequency, inverse document frequency).

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 mutation 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. Most genomic investigations currently are not predictive for clinical outcome. 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 rhobust 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 (in 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 find 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 clinically relevant 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 very variable 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.

Results

RAG1 and RAG2 conservation and mutation rate residue frequency

Variant probability prediction is dependent on population genetics data. While many in-house or public database are available, our study queried GnomAD [21] to identify conserved residues using a Boolean score C (0 or 1, although allele frequency can be substituted). The gene-specific mutation rate M_r of each residue was calculated from allele frequencies. The gene-specific residue frequency R_f was also calculated and together the values are used to calculate the most probable disease-causing variants which have not yet been identified in patients. We term the resulting score a mutation rate residue frequency (MRF); where $MRF = C \times M_r \times R_f$. For visualisation, a noise reduction method was also applied; a sliding window was used to find the average MRF per 1% interval of each gene. The resulting scores are displayed with a cut-off threshold to highlight the top scoring residues (using the 75^{th} percentile).

Figure 1 presents the most probable unidentified disease-causing variants in RAG1/2. Phenotypic, epigenetic, or other such weighting data may also be applied to this model. 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 clinical presentation. Raw MRF scores are presented in **Figure 1 (ii)**. A histogram illustrates the MRF without Boolean scoring applied and **Figure 1 (iii)** provides a clearer visualisation. Variant sites most likely to present in disease cases are identified by high MRF scoring. **Table S1** provides all

MRF scores for both proteins as well as raw data used for calculations and the list of validated residues of RAG1 and RAG2. Analysis-formatted data is also available in the Supplemental zip "RAG MRF map" along with the associated R source file to allow for alternative visualisations as shown in **Figure S1**.

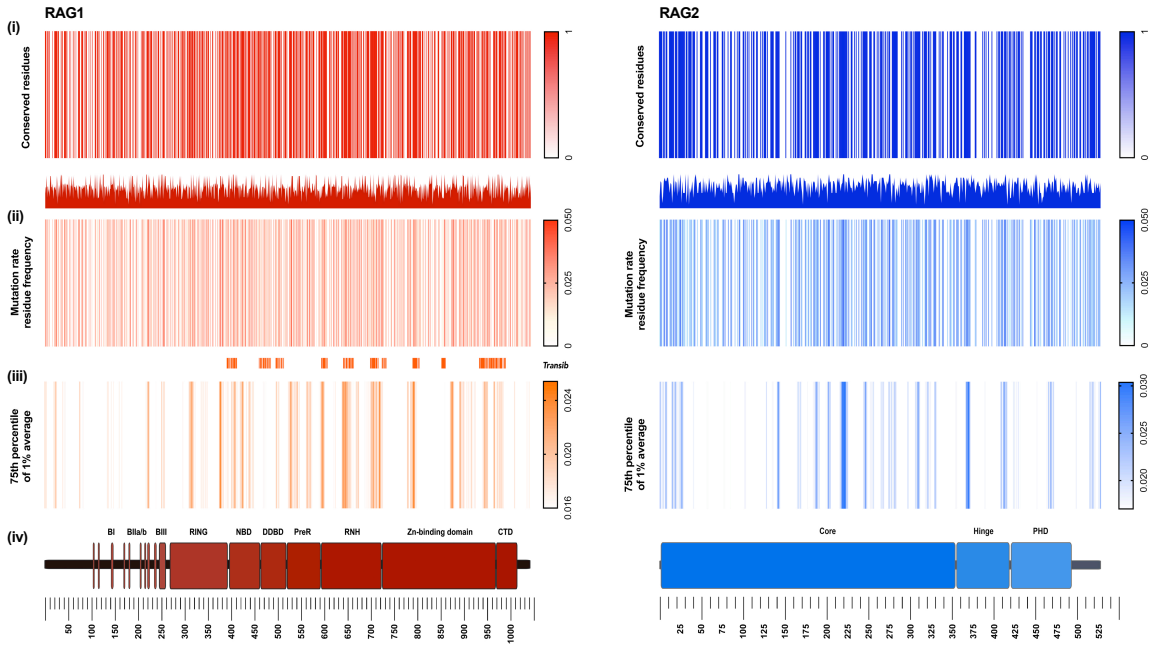

Figure 1: RAG1 and RAG2 conservation and mutation rate residue frequency. (i) Gene conservation score, non-conserved 0 and conserved 1. (ii) Histogram; raw MRF score. Heatmap; MRF prediction for conserved residues, graded 0 to 0.05 (scale of increasing likelihood for producing disease). (iii) MRF score averaged with 1% intervals for each respective gene and cut-off below 75th percentile, graded 0 to 0.03 (Noise reduction method). (iv) Gene structure with functional domains.

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, 24–47] [*originally compiled by* Notarangelo et al. [48]]. 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 damaging mutations in RAG1 and RAG2 (**Fig. 2 (i)**). Variants reported on GnomAD which are clinically found to cause disease have significantly higher MRF scores than variants which have not been reported to cause disease (**Fig. 2 (i)**).

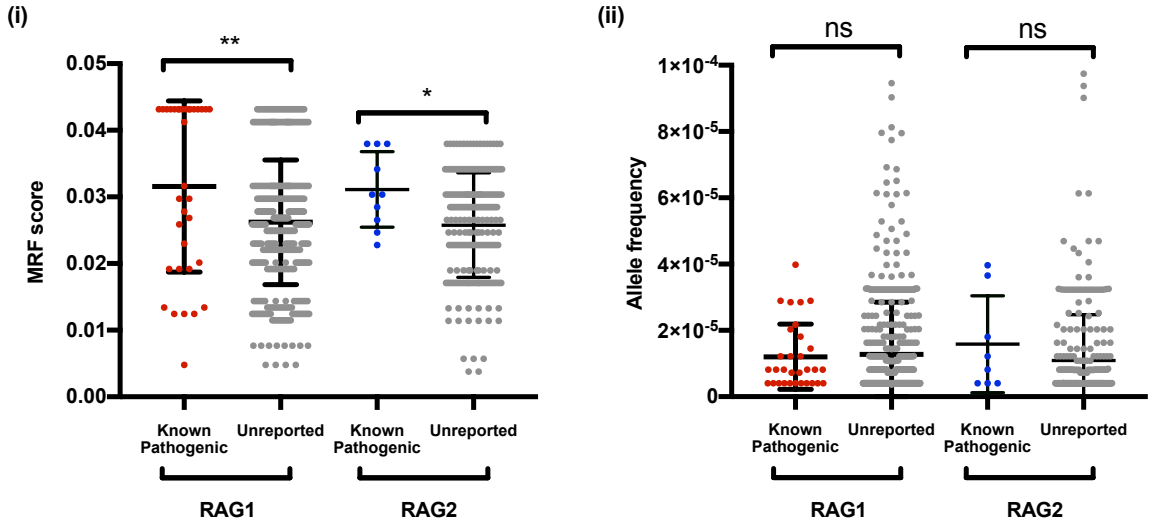

Figure 2: RAG1 and RAG2 mutation rate residue frequency accurately predicts disease. (i) Known damaging variants (clinically diagnosed with genetic confirmation) reported on GnomAD have significantly higher MRF scores than non-pathogenic variants. (ii) GnomAD rare variant allele frequency <0.0001. No 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. 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, indicated by their low probability of being loss-of-function intolerant (pLI) scores of 0.00 and 0.01, respectively [21]. Therefore, allele frequencies of rare variants reported on GnomAD cannot differentially predict the likelihood of causing disease (**Fig. 2 (ii)**). This is particularly important for recessive diseases such as RAG deficiency. As such we find no significant difference between known damaging variants and those which have not been reported yet as disease-causing, illustrating the reasoning for our method design (**Fig. 2 (ii)**). Many non-clinically-reported rare variants may cause disease; the MRF score identifies the top clinically-relevant candidates.

MRF scores predict functional validation

The functional validation of MRF predictions is presented in **Figure 3**. 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 is measured as the level of recombination potential produced by the protein complex. Each method of investigation simulates 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) coding for immunoglobulins.

In functional experiments, mutant proteins were assayed for their ability to perform recombination on a substrate which mimics the recombination signal sequences (RSS) of TCR and BCR in comparison to wild type protein complex (as % SEM with SD). Our investigation uses the inverse of these measurements, 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%)). We compared predicted MRFs to assay measurements for 71 and 39 mutant *RAG1* and *RAG2* plasmids, respectively.

The accuracy for correctly identifying all disease-causing variants reported to date is shown per-variant in **Figure 3 (i-ii)**. Best performance was seen for RAG1. We found >80% accuracy for 21 known variants tested, >50% accuracy for 48 tested and <50% accuracy for

only 23 tested (**Fig. 3 (iii-iv)**). 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 MRF score above average while being measured as functionally benign (a false positives rate of 11.27%). RAG2 scored with only 3 variants out of 39 variants (7.69%) with an MRF above average while functionally benign. However, the 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. **Figure 4** illustrates the breakdown of functional testing carried out to date compared to the likelihood of such occurrences in disease cases.

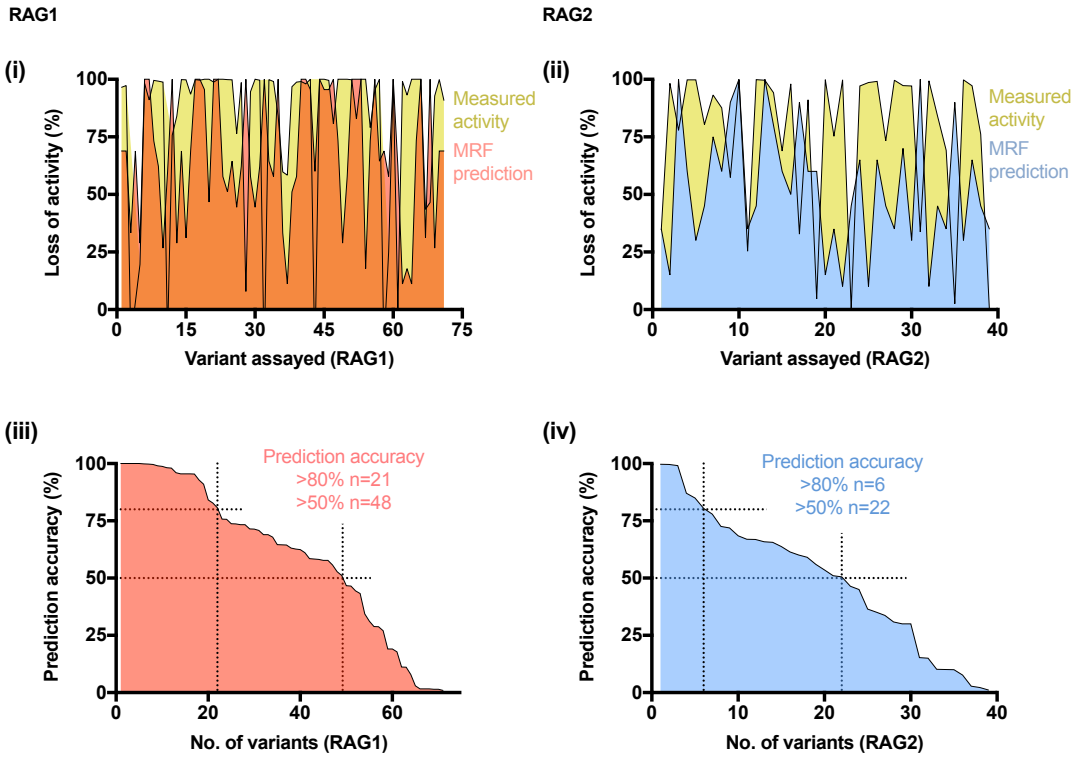

Figure 3: MRF score versus in vitro loss of protein activity. (i-ii) Predicted likelihood for disease-presentation (based on maximum and minimum MRF score as a percentage) is shown in red or blue for RAG1 and RAG2, respectively. In yellow, the functionally measured recombination activity of each variant where complete loss of protein activity is measured as (its inverse) 100% loss of activity. (iii-iv) Accuracy of MRF scoring compared to functionally validated recombination activity.

Top candidate variants require validation

Functionally characterising protein function 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* which were potentially damaging or were identified from PID patients as the most probable genetic determinant of disease. Functional assays for RAG deficiency in these cases, and generally, measure a loss of recombination activity as a percentage of wild type function (0-100%). MRF is not a predictor of pathogenicity, but a likelihood of variation occurring.

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 over 100 variants have been assayed in vitro, we calculate that only one quarter of those are most probable candidates for clinical presentation. **Figure 4** illustrates that while functional work targeted "hand picked" variants which were ultimately confirmed as damaging, many of those may be unlikely to arise based on population genetics data. On the left of **Figure 4**, the measured loss of protein activity is accompanied by an MRF heatmap for functionality validated mutations. To the right, the breakdown of MRF scores of the complete proteins are shown (i.e., for RAG1; 0.043-0.0048). We predict that only 21 of the top 66 most probable clinically relevant variants have been tested in *RAG1*.

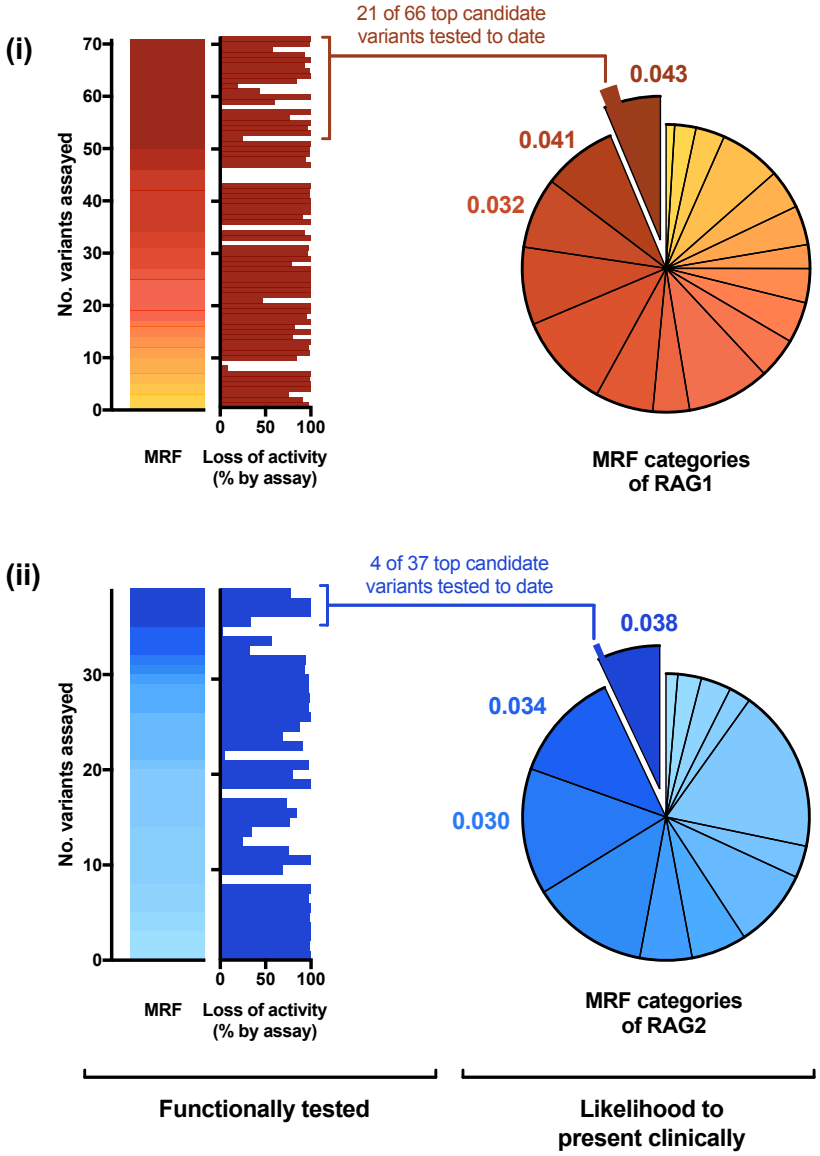

Figure 4: RAG1 and RAG2 MRF score categories and variants assayed to date. On the left, we rank the in vitro measurement of recombination activity (as its inverse; % loss of activity) by the MRF score per residue. On the right, the breakdown of MRF score categories are shown per protein. While many protein residues are critical to protein function, their mutation is less probable than many of the top MRF candidates. This indicates that pre-emptive clinically-relevant investigations may require tailoring using variance probability. MRF; mutation rate residue frequency.

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 [49, 50]. The *Transib* transposon is a 600 amino acid core region of RAG1 which 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 [51], 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" [52].

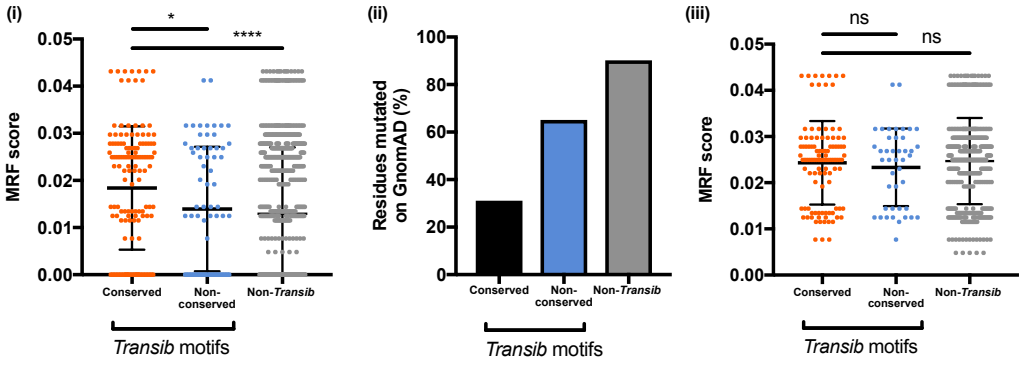

Figure 5: False positives in *Transib* domains do not worsen probability prediction. The *Transib* domains contain critical conserved protein residues. (i) False positives are simulated by scoring *Transib* domains MRF without their Boolean conservation weight C. (ii) Allele frequencies on GnomAD have inversely proportional conservation to simulated false-positive MRF scoring. (iii) When the Boolean component C is applied in MRF calculation the effect of false positives remains non-significant, illustrating the non-negative impact of MRF for pathogenicity rate prediction. 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) [53]. Ten *RAG1* core motifs are conserved amongst a set of diverse species including human [53]. 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 5 (i)** plots the MRF (lacking the Boolean

component C) for conserved *Transib* motif residues, non-conserved *Transib* motif residues, and non-*Transib* residues. **Figure 5 (ii)** shows the percentage of these which are reported as mutated on GnomAD. Removing reported variants by applying C, the resulting effect on incorrectly scoring MRF in the conserved *Transib* motifs remains neutral.

Conserved residues with the highest MRF for both RAG1 and RAG2 are mapped onto the protein structure in **Figure 6**. Structural mapping frequently shows high MRF scores at DNA contact points and protein-protein interaction sites; such residues have a logical involvement with protein function.

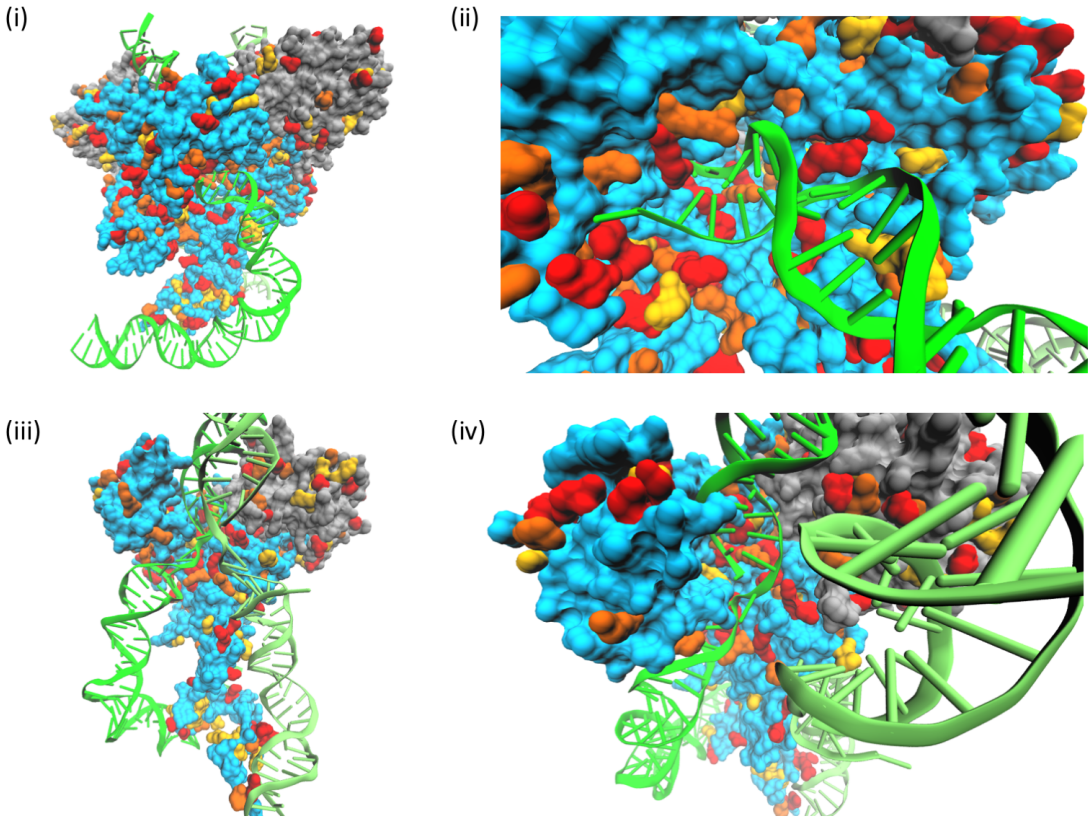

Figure 6: The RAG1 (blue) and RAG2 (grey) protein structure with top candidate MRF scores. (i) Protein dimers and (ii=iv) monomers illustrating the three highest category MRF scores for predicted clinically-relevant variants. Increasing in score the top three MRF categories (illustrated in **Figure 4**) for each protein are highlighted; yellow, orange, red. DNA (green) is bound by the RAG protein complex at recombination signal sequences. DNA contact points are integral to protein function. (PDB:3jbw)

Combined Annotation Dependent Depletion (CADD) scoring [54] is an important bioinformatics tool. 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 *RAG1* CADD scores for all possible SNV positions in RAG1 (**Fig. 7**) illustrating that pathogenicity prediction cannot account for mutation probability.

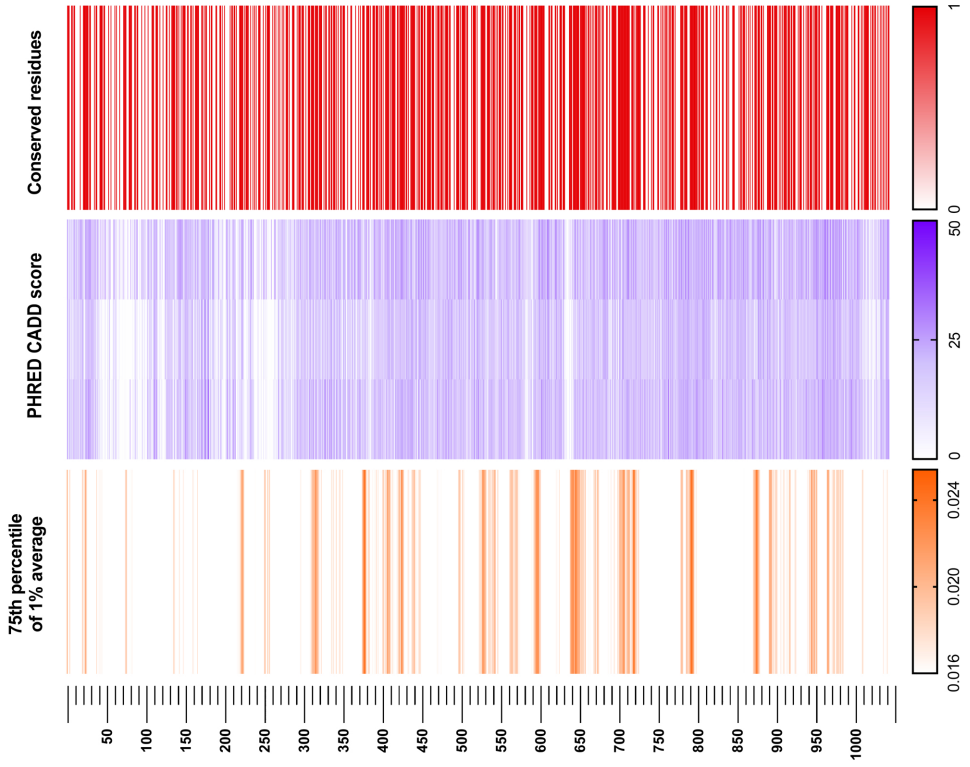

Figure 7: *RAG1* PHRED-scaled CADD score versus GnomAD conservation rate and MRF score. Allele frequency conservation rate (top) is vastly important for identifying critical structural and functional protein regions. The impact of mutation in one of these conserved regions is often estimated using CADD 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 75^{th} percentile with 1% averaging) highlights protein regions which are most likely to present clinically and may require pre-emptive functional investigation.

MRF predicts RAG deficiency amongst PID patients harbouring rare variants

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 slopes which were either negative or close to zero 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 is inferred in **Figure 8** [55]. R source and raw data can be found in supplemental material.

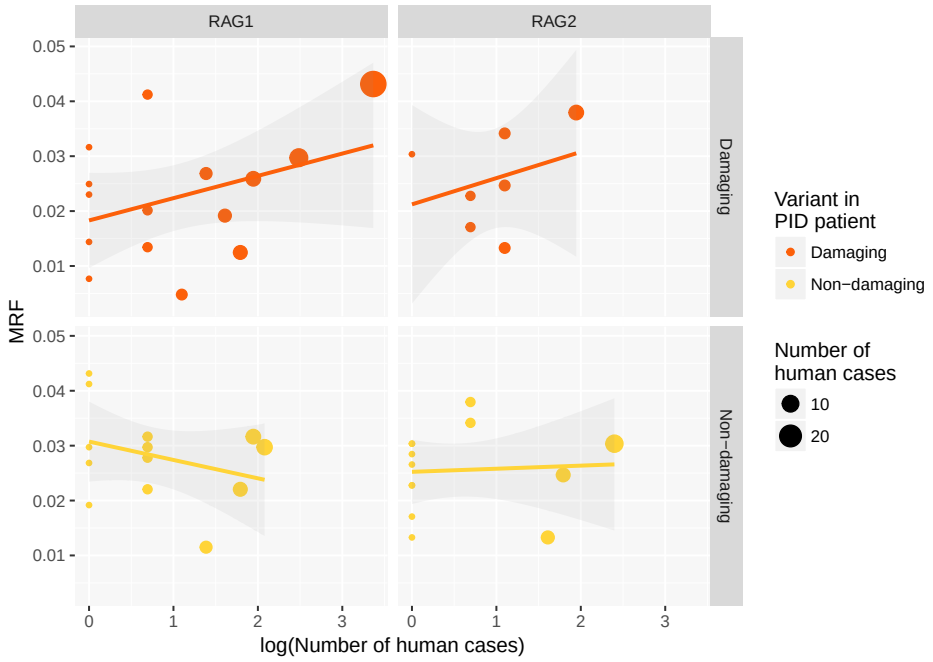

Figure 8: A linear regression model of RAG1/2 MRF scoring in cases of primary immune deficiency. MRF prediction correlates with disease. Damaging variants identified in confirmed RAG deficiency cases. Non-damaging variants sourced from cases of PID with rare variants but not responsible for disease. (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).

Discussion

Determining disease-causing variants for functional analysis typically aims to target conserved gene regions. On GnomAD 55.99% of *RAG1* (approx. 246,000 alleles) has no reported variants. Functionally validating unknown variants in genes with this level purifying selection is generally infeasible. Conserved regions are likely high importance regions, yet determining the likelihood of patients presenting with mutations in these clusters requires a scoring mechanism. An example of such clustering of highly scoring MRFs occurred in the RAG1 catalytic RNase H (RNH) domain at p.Ser638-Leu658 which is also considered a conserved *Transib* motif. Targeting clearly defined regions with high MRF scores allows for functional validation studies tailored to the most clinically-relevant protein regions. Weighting data may also be applied to this model to amplify the selectivity. Genome wide, this might included phenotypically derived weights to target candidate genes or tissue-specific epigenetic features. While many hypothetical variants with low MRF scores could be found 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 S1** lists the values for calculated MRFs for RAG1 and RAG2.

We have presented a basic application of MRF scoring for RAG deficiency. Future implementation of the MRF method can include genome-wide application 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;

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

We expand here briefly on the technical description of this method. Log weighting may offer clearer disease-causing variant discovery depending on the scoring method. 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 ($rf = rf_{r,g}$). 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.

During clinical investigations using personalised analysis of patient data, further scoring methods may be applied based on disease features. A patient with autoinflammatory features may require weighting for genes such as *MEFV* and *TNFAIP3*, whereas a patient with mainly immunodeficiency may have weighted scoring for genes such as *BTK* and *DOCK8*. A patient phenotype can contribute a weight based on known genotype correlations separating primary immunodeficiencies or autoinflammatory diseases [6]. However, validation of these expanded implementations requires a deeper consolidation of functional studies than is currently available. A method with similar possible applications for human health mapping constrained coding regions has been recently developed by Havrilla et al. [56]. Their study employed a method which 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 [57]. 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 potential to cause disease. It is not a predictor of pathogenicity. However, MRF may contribute to pathogenicity prediction as a component of Bayesian probability. While

beyond the scope of this investigation, we can outline the implementation of this approach here. 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 [58], 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, (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.

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 [59]; 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 [59]. 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 [60].

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 (<0.0001) will remain non-actionable until reliable predictive genomics practices are developed. Bioinformatics as a whole has made staggering advances in the field of genetics [61]. Challenges which remain unsolved, hindering the benefit of national or global genomics databases, include DNA data storage and random access retrieval [62], data privacy management [63], and predictive genomics analysis methods. Variant filtration in rare disease is based on reference allele frequency, yet the result is not clinically actionable in most cases. Development of predictive genomics tools may provide a critical role for single patient studies and timely diagnosis [23].

Conclusion

We provide the amino acid residue list for RAG1 and RAG2 which 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.

Methods

Population genetics

GnomAD (version r2.0.2) [21] was queried for the canonical transcripts of *RAG1* and *RAG2* from population genetics data of approximately 146,000 alleles;
ENST00000299440 (*RAG1*) 1495 variants (including filtered: 1586),
GRCh37 11:36532259-36614706 and ENST00000311485 (*RAG2*) 786 variants (including filtered: 831), GRCh37 11:36597124 - 36619829. However, any source of population-scale data can be substituted here. Data was filtered to contain the 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; 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 [64].

Input sets used GnomAD variant allele frequencies and reference sequences processed as csv files, cleaned and sorted to contain only coding amino acid residues, amino acid code, residue number, alternate variants, allele frequencies of variants, and a score (C) of 0 or 1 where 1 represented no reported variants. A score was also given where multiple alternate variants existed. A separate statistics report was generated from this processed input data. Statistics and calculation steps are listed in order in **Table S1** [Raw data calculation]. The percentage of conserved residues was calculated (55.99% of amino acids contained no reported variants in RAG1, 55.98% in RAG2). The count of variants per residue was found for both proteins. The ratio was also found per residue conservation rate / mutation rate. Basic protein statistics were generated using reference canonical transcript sequences of RAG1 and RAG2 with the Sequence Manipulation Suite [64]. The residue frequency was calculated based on the respective polypeptide chain length.

The calculated mutation rate value and residue frequency score together produce the mutation rate residue frequency as shown in **Table S1**. Our investigation used the Boolean C score of 0 or 1 to weight mutation rate residue frequencies. An important consideration for future application is whether to use this Boolean score or a frequency score. In the clinical setting, the likelihood of *de novo* mutations versus inherited mutations have different impact on recessive and dominant diseases. The likelihood of a patient presenting with a particular (predicted) variant is more likely if the variant exists even at a very low frequency in the patient's ancestral population. Therefore, an allele frequency may be used to replace C in many investigations, particularity when considering variants that exist at low rates.

Data visualisation

For our visualisation of MRF scores, small clusters of high MRF were of more significance 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). However, when using Boolean scoring C, this method should be applied before C. Alternatively, if using allele frequency scoring, this

visualisation method can be applied subsequently. Lastly, for a clear distinction of MRF clusters a cut-off threshold was applied at the 75^{th} percentile (0.0168 in RAG1).

A gene map for coding regions in *RAG1* and *RAG2* were populated with (1) Boolean C score from population genetics data, (2) raw MRF scores, and (3) MRF clusters with 1% average and cut-off threshold. GraphPad Prism was used for heatmaps. The data used for heatmaps is also available in the Supplemental zip "RAG MRF map" (Table S1 simplified) for automatic loading in the associated R source file to allow for alternative visualisations. An example of alternative output for non-R users is shown in **Figure S1**. Adobe Adobe Illustrator and Photoshop were used for protein domain illustrations.

The crystal structure of DNA bound RAG complex was produced with data from RCSB Protein Data Bank (3jbw.pdb) [65]. Structures were visualised using the software VMD from the Theoretical and Computational Biophysics Group [66]. Imaged with Tachyon rendering [67] and colour mapped using MRF scores.

Validation of MRF against functional data

The recombination activity of RAG1 and RAG2 was previously measured on 44 known pathogenic variants [16, 68]. Briefly, the pathogenicity of variants in RAG1 and RAG2 are measured functionally *in vitro* by expression of RAG1 and RAG2 in combination with a recombination substrate plasmid containing RSS sites which are targeted by RAG complex during normal V(D)J recombination. Recombination events are assessed by quantitative real-time PCR using comparative CT. The inverse score of recombination activity (0-100%) is 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).

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] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] Paul Stothard. The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *University of Alberta, Education and Research Archive*, 2000. URL https://bioinformatics.org/sms2/mirror.html.

[65] 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.

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

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

[68] Yu Nee Lee, Francesco Frugoni, Kerry Dobbs, Irit Tirosh, Likun Du, Francesca A Ververs, Heng Ru, Lisa Ott de Bruin, Mehdi Adeli, Jacob H Bleesing, et al. Characterization of t and b cell repertoire diversity in patients with rag deficiency. *Science immunology*, 1(6), 2016.

Supplemental

Table S1: MRF data tables. **(1) MRF for RAG1 and RAG2** shows the complete amino acid residue MRF scores along with the known pathogenic variant residues used for calculations. Colour coding: MRF and 1% average - green heatmap; raw MRF before Boolean score - red heatmap; residue number, wild type amino acid, and variant reported - red indicates reported in general population and blue indicates a conserved site; sites with multiple variants are marked in beige [21]. **(2) Raw data calculation** lists all of the raw values used to calculate MRF score in ordered steps; (i) Percentage of mutated vs. unmutated variants, (ii) Number of times each residue is mutated, (iii) Ratio that each residue is mutated, (iv) Residues sorted by conservation rate, (v) Basic statistics using SMS2 [64], (vi) MRF per amino acid.

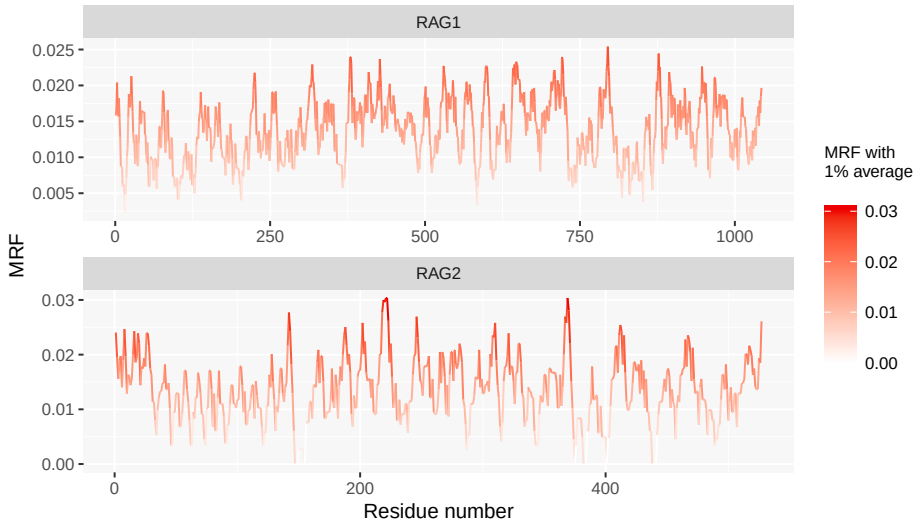

Figure S1: An alternative visualisation of MRF scores for RAG1 and RAG2 proteins. This figure is the output from the compressed "R source RAG MRF map" file. The data from Table S1 in column "Average over 1%" is displayed on both the y-axis and colour scale. An analysis-friendly long form csv of the Table S1 data is also provided in the compressed supplemental as file "Table S1 simplified".