# Characterising *RAG1* and *RAG2* with predictive genomics

Dylan Lawless, MSc[a,*], Jolan E. Walter, MD, PhD[b,c], Rashida Anwar, PhD[a],
Sinisa Savic, MD, PhD[d,e,*]

[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] *University of South Florida and Johns Hopkins All Children's Hospital, Saint Petersburg, FL, USA.*
[c] *Division of Allergy Immunology, Massachusetts General Hospital for Children, Boston, MA, USA.*
[d] *Department of Clinical Immunology and Allergy, St James's University Hospital, Beckett Street, Leeds, UK.*
[e] *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. The greatest hurdle in diagnosis of rare disease is validation for variants of unknown significance. RAG deficiency presents at an early age with a distinct phenotype of combined immunodeficiency with granuloma and/or autoimmunity. Allele frequency of a SNV in the general population is an indicator of the functional or structural importance of a particular amino acid residue. However, rare diseases are often attributable to variants in genes which are highly conserved. Mutation of a conserved residue does not confirm pathogenicity and functional validation must be confirmed to correctly identify a monogenic disorders such as RAG deficiency. We present protein variants in RAG1 and RAG2 which are most likely to be seen clinically as disease-causing. Our method of mutation rate residue fre-

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

quency builds a map of most probable mutations allowing pre-emptive functional analysis. We compare the accuracy of our predicted probabilities to previously established functional measurements.

## Acknowledgements

This work is funded by the University of Leeds 110 Anniversary Research Scholarship. Dr. Jolan Walter has received federal funding. The authors declare no conflict of interest.

## Key words

RAG1, RAG2, genomics.

## Abbreviations

CADD (combined annotation dependent depletion), GWAS (genome-wide association studies), $M_r$ (mutation rate), MFR (mutation rate residue frequency), (pLI) probability of being loss-of-function intolerant, $R_f$ (residue frequency), RAG1 (recombination activating gene 1).

# 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 the 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 variants of unknown 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. Although, 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 with functional studies. Our investigation aims to apply predictive genomics as a tool to identify pathogenic 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 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]. RAG deficiency is mesured by in vitro quantification of recombination activity. Hypomorphic RAG1 and RAG2 mutations with 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, 10, 11]. *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. Pre-emptive selection of residues for functional validation is a major challenge; a selection based on low allele frequency alone is infeasible. A shortened time between genetic analysis and diagnosis means that treatments may be delivered earlier. With such tools, patients with RAG deficiency may receive hematopoietic stem cell transplant [12] or be provided mechanism-based treatment [13].

GnomAD was queried 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 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 where the average MRF per 1% interval is displayed with a cut-off threshold at the 75th percentile.

## Results

*RAG1 and RAG2 conservation and mutation rate residue frequency.*

Fig 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 Fig 1(i) however; these clusters do not predict the likelihood of clinical presentation. Raw MRF scores are presented in Fig 1(ii). A histogram illustrates the MRF without Boolean scoring applied and

Fig 1(iii) presents a clearer visualisation. 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.

*MRF score versus known variant pathogenicity measure*

The functional validation of these predictions is presented in Fig 1(v). We have previously measured the recombination activity of RAG1 and RAG2 disease-causing variants in several patients [14]. We have combined the known functional activity from other extensive reports [15], to compare a total of 44 variants. RAG deficiency is measured by the level of recombination potential. We expected that damaging mutations (resulting in low recombination activity in vitro) would be identified with high MRF scores. MRF pathogenicity prediction correctly identified damaging mutations in RAG1 and RAG2 (Fig 1(v)). 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 1(v)). 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.

Allele frequency is generally the single most important filtering method for rare disease in whole genome (and exome) sequencing experiments. *RAG1* and *RAG2* have probability of being loss-of-function intolerant (pLI) scores of 0.00 and 0.01, respectively. Mutations under pressure from purifying selection are more likely to cause disease than common variants. However, allele frequencies of rare variants reported on GnomAD cannot differentially predict likelihood of causing disease. This is particularly important for recessive diseases such as RAG deficiency. As such we find no significant difference between clinically damaging variants and those which have not been reported yet as disease-causing, illustrating the reasoning for our method design (Fig 1 (vi)). Many non-clinically-reported rare variants may cause disease; the MRF score identifies the top clinically-relevant candidates. Conserved residues with the highest MRF for both RAG1 and RAG2 are mapped onto the protein structure in Fig 3 and frequently show high MRF at DNA contact points. The accuracy for correctly identifying all disease-causing variants reported to date is shown in (Fig 1(vii). We found >80% accuracy for 21 known variants tested, >50% accuracy for 48 tested and <50% accuracy for only 23 tested. The raw values comparing functional pathogenicity and MRF scores are illustrated in Fig 2.

*False positives in Transib domains do not worsen probability prediction*

A set of conserved motifs in core *RAG1* are shared with the *Transib* transposase, including the critical DDE residue catalytic triad [16]. Ten *RAG1* core motifs are conserved amongst a set of diverse species including human [16]. To assess the influence of false positive effect on MRF prediction the conserved residues in this dataset are compared to GnomAD allele frequencies and MRF score. Fig 4 (i) plots the MRF (lacking the Boolean component $C$) for conserved *Transib* motif residues, non-conserved *Transib* motif residues, and non-*Transib* residues. Fig 4 (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. Combined Annotation Dependent Depletion (CADD) scoring [17] is an important bioinformatics filtering method. We compare MRF to the PHRED-scaled *RAG1* CADD scores for all possible SNVs (Fig 5). While CADD is a valuable scoring method its purpose is not to predict likelihood of variation.

## 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 occured 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. 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. Table S1 lists the values for calculated MRFs for RAG1 and RAG2.

We have presented a basic application of MRF scoring for RAG deficiency. Furthermore, we have suggested its genome-wide application with to the information retrieval method; term frequency, inverse document frequency ($tf - idf$). In this case the“term” will represent an amino acid residue $r$ while the "document" represents a gene $g$ 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 $g$.

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

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 ($\text{Score}(q, g)$) 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 released [18]. This study employed a method which included weighting by sequencing depth. We have not included this method as our analysis was gene-specific but implementation is advised when calculating genome-wide MRF scores.

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 [19]; 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. Autoimmune diseases had risk allele frequencies considered low at approximately 0.4 [19]. 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 [14]. 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 [20].

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 a a whole has made staggering advances in the field of genetics [21]. Challenges which remain unsolved, hindering the benefit of national or global genomics databases, include DNA data storage and random access retrieval [22], data privacy management [23], 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 [13].

## 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) [24] was queried for the canonical transcripts of *RAG1* and *RAG2* from population genetics data of approximately 146,000 alleles; ENST00000299440 (*RAG1*) 1495 vaiants (including filtered: 1586), GRCh37 11:36532259-36614706 and ENST00000311485 (*RAG2*) 786 varaitns (including filtered: 831), GRCh37 11:36597124 - 36619829. 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 (http://bioinformatics.org/sms2/mirror.html).

Input sets used GnomAD variant allele frequencies and reference sequences processed as cvs 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. 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. 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 patients ancestral population. Therefore, an allele frequency may be used to replace *C* in many investigations.

### *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 75th 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 cutoff threshold. GraphPad Prism was used for heatmaps and Adobe Adobe Illustrator and Photoshop were used for protein domain illustrations.

### *Validation of MRF against functional data*

The recombination activity of RAG1 and RAG2 was previously measured on 44 known pathogenic variants [14, 15]. 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 recombination signal sequences 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 MFR 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] Kapourani Chantriolnt-Andreas Stubbs Thomas M. Lee Heather J. Alda-Catalinas Celia Krueger Felix Sanguinetti Guido Kelsey Gavin Marioni John C. Stegle Oliver Reik Wolf Clark Stephen J., Argelaguet Ricard. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature Communications*, 9(1):781, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03149-4.

[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, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Mary Ellen Conley, Charlotte Cunningham-Rundles, Amos Etzioni, Steven M Holland, Christoph Klein, et al. Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency 2015. *Journal of clinical immunology*, 35(8):696–726, 2015.

[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] 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.

[11] 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.

[12] 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 granulo-

matous disease and autoimmunity secondary to rag deficiency. *Journal of clinical immunology*, 36(7):725–732, 2016.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H OâĂŹDonnell-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.

Table S1: MRF data tables. The complete RAG1 and RAG2 amino acid residue MRF scores are provided along with known clinically pathogenic variant residues and raw data used for calculations.

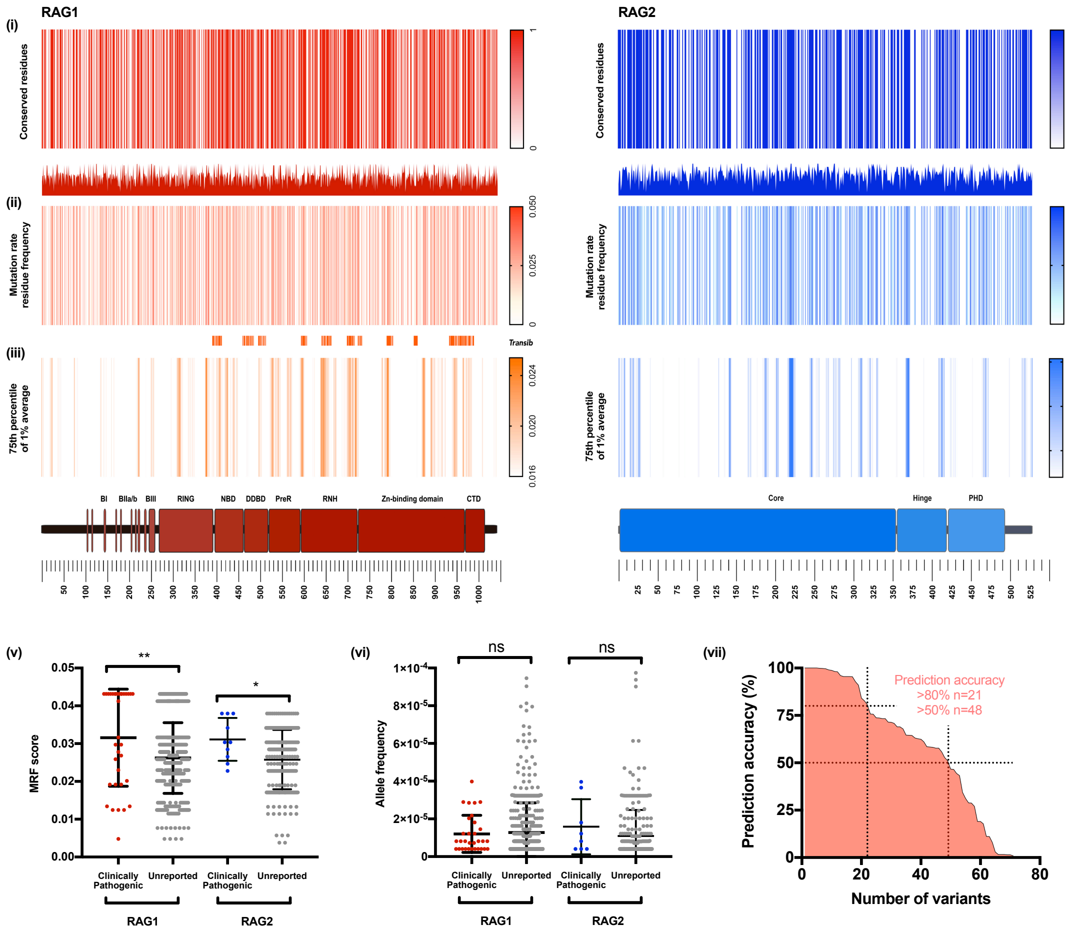

Figure 1: RAG1 and RAG2 conservation and mutation rate residue frequency. (i) Gene conservation score, non-conserved 0 and conserved 1. (ii) Histogram; raw MRF score. Heatmap; MRF prediction for conserved residues, graded 0 to 0.05. (iii) MRF score averaged with 1% intervals for each respective gene and cut-off below 75th percentile, graded 0 to 0.03 (Noise reduction method). (iv) Gene structure with functional domains. (v) Clinically damaging variants reported on GnomAD have significantly higher MRF scores than non-pathogenic variants. (Unpaired t test. RAG1 P value 0.002** RAG2 P value 0.0339*). (vi) GnomAD allele frequency <0.0001. No significant difference in allele frequency is found between clinically damaging variants and non-clinically reported. (vi) Accuracy of MRF scoring compared to functionally validated pathogenicity.

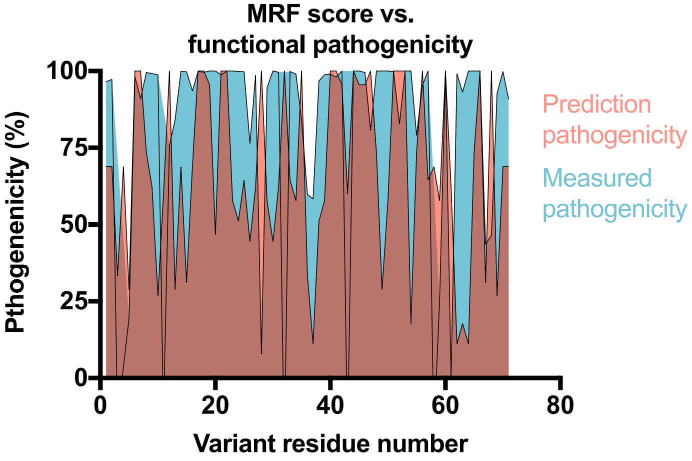

Figure 2: MRF score vs. known variant pathogenicity measure. Predicted pathogenicity likelihood (based on maximum and minimum MRF score as a percentage) is shown in red. In blue, the functionally measured recombination activity of each variant where complete loss of protein activity is measured as 100% pathogenicity. These values are summarised in Fig 1v(ii).

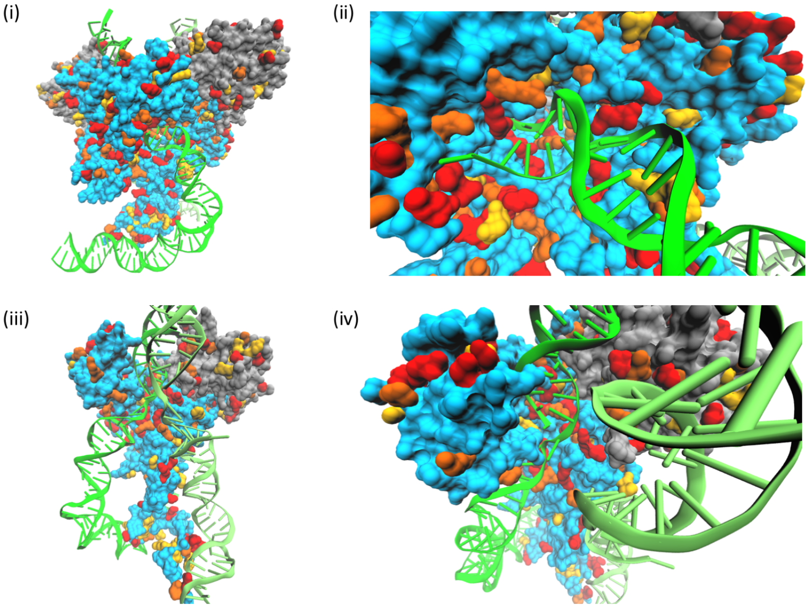

Figure 3: The RAG1 (blue) and RAG2 (grey) protein structure with MRF scores. (i) Protein dimers and (ii=iv) monomers illustrating the three highest category MRF scores for predicted clinically-relevant variants. Increasing in MRF score; yellow, orange, red. DNA contact points are integral to protein function and generally score as high MRF residues. (PDB:3jbw)

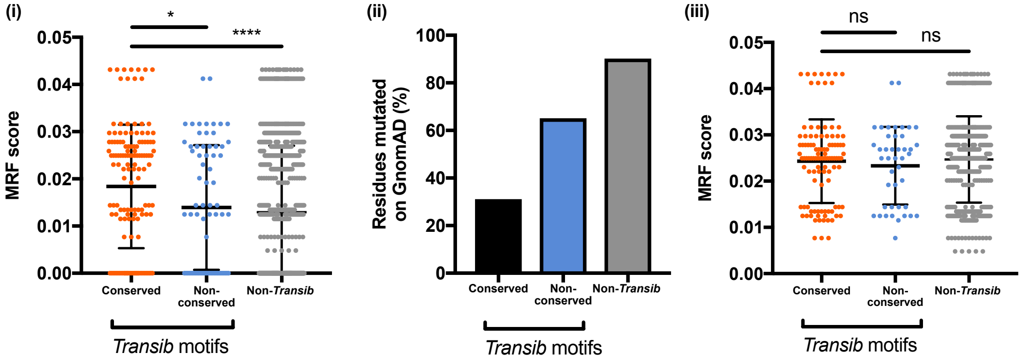

Figure 4: False positives in *Transib* domains do not worsen probability prediction. The *Transib* domains contain critical conserved protein residues. (i) False positives are simulated by scoring *Transib* domains MRF without their Boolean conservation weight $C$. (ii) Allele frequencies on GnomAD have 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.

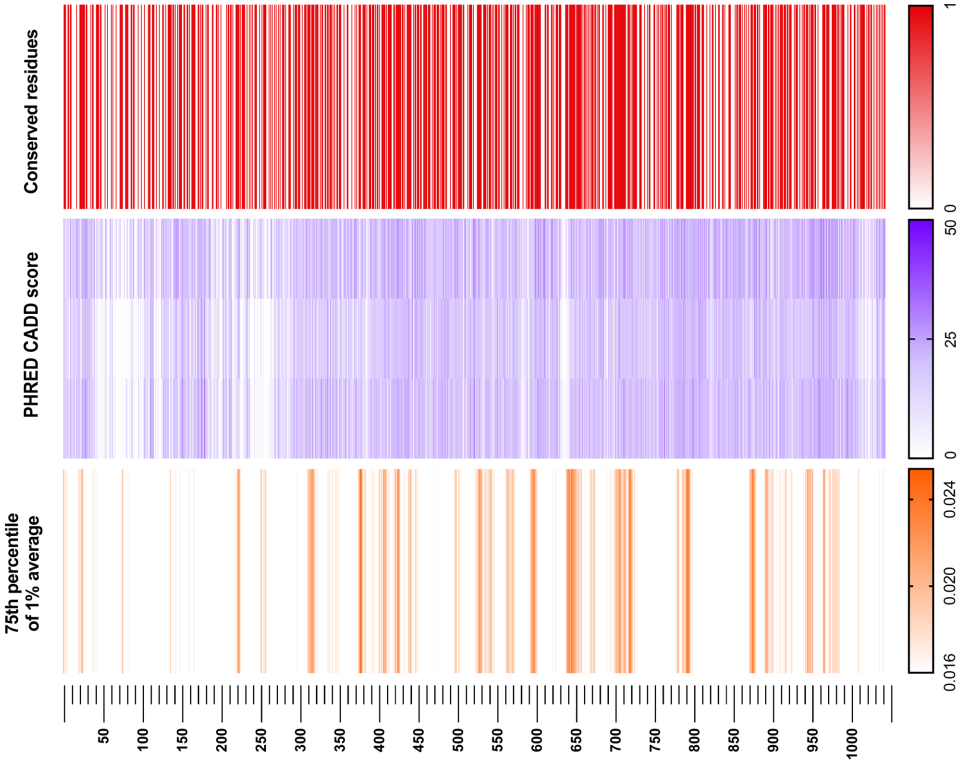

Figure 5: *RAG1* PHRED-scaled CADD score versus GnomAD conservation rate and MRF score. Allele frequency conservation rate (top) is vastly important for identifying critical structural and functional protein regions. The impact of mutation in one of these conserved regions is often estimated using CADD scoring (middle). 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.