HHS Public Access

Author manuscript

Hum Genet. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Hum Genet. 2020 June ; 139(6-7): 759–768. doi:10.1007/s00439-019-02091-9.

Regulatory genome variants in human susceptibility to infection

Amalio Telenti, **Julia di Iulio**

Department of Integrative Structural and Computational Biology. The Scripps Research Institute, La Jolla, CA 92037, USA

Abstract

Genome studies have accelerated the discovery of common and rare genetic variants associated with susceptibility to infection and with disease severity. Genome wide association studies identified many common genetic variants associated with modest risk for infection. Over 80% of these common variants map to the non-coding genome and are thought to modulate the regulatory networks. Exome sequencing has rapidly expanded the number of recognized primary immunodeficiencies through the identification of rare coding variants. In contrast, less than 29 primary immunodeficiencies have causative rare variation mapped outside protein coding regions. In the future, whole genome sequencing will accelerate the identification of rare variants of substantial phenotypic impact that disrupt essential regulatory elements and the 3-dimensional structure of chromatin.

The two main strategies of study of the human genome are the sequencing of the exome, and the high-density genotyping arrays used in genome-wide association studies (GWAS). Both approaches have been successful for the identification of genetic determinants of susceptibility to infectious diseases. Exome sequencing has improved the characterization of primary immunodeficiencies (PIDs) while GWAS have identified common genetic variants associated with risk infection. GWAS also deepened the understanding of the role of the Major Histocompatibility Complex (MHC) in infection and immunity. Notably, GWAS studies, by mapping many of the associated variants outside of the coding regions have provided a first assessment of the role of regulatory regions in disease.

Recent efforts at extending the search space from the protein coding (exome) to the immediate regulatory regions show that new regulatory pathogenic variants of large effects can be identified in a small fraction of human disease cases (1–3). In parallel, there are recent examples of diseases that implicate genetic variation in distal enhancers and changes in the 3D genome structure (2, 4). Thus, the next milestones in the interpretation of the

Correspondence: atelenti@scripps.edu.

Conflict of interest

A.T. and J.dil. are current employees of Vir Biotechnology Inc. This employment is not a conflict of interst in regards to the content of the present paper.

human genome sequence may emerge from the analysis of functional consequences of rare genetic variants in the non-protein coding genome that includes the regulatory machinery.

Redundancy in regulatory networks

There is debate on the expected impact of non-coding, regulatory variants on human traits and diseases. The common belief states that there is considerable regulatory redundancy and that most regulatory variants will have limited functional consequences. Therefore, the concept of redundancy is central to current debate on the robustness of gene regulatory networks. Redundancy reduces the likelihood of severe consequences resulting from genetic or environmental challenges – including in infection (5). Work on developmentally expressed genes supports the concept of functionally redundant enhancers in mammalian genomes (6). However, enhancers can also be under purifying selection over evolutionary time and be relevant for organismal fitness under specific pressures because of their contribution to overall gene expression levels (6). We have in the past indicated that essential genes use proximal and distant regulatory elements that are constrained (conserved) intra-human – thus representing human-specific essentiality in the non-coding genome (7, 8). Such essential/constrained regulatory regions may be important for the fitness of the organism.

The identification of disruptive variants in regulatory elements with critical function may lead to disease. Redundancy in genomic regulatory circuitry should not be interchanged with biological redundancy where mutations affecting immunity genes may result in pleiotropic patterns of susceptibility to infections or result in no disease (9).

Defining putative essential regulatory elements

Interpretation of the non-coding genome requires the identification of landmarks, features and structures, the same principles that aid the interpretation of the coding genome. Genome-wide epigenomic maps (ENCODE, BluePrint Epigenome, Roadmap Epigenomics) have revealed hundreds of thousands of regions showing signatures of enhancers, promoters, and other gene-regulatory elements (10). However, the high-resolution dissection of functionally relevant nucleotides in the non-coding genome remains limited at present (11). Multiple sources of biochemical, genetic, and evolutionary data convey functional information on the non-coding genome (12). These data are used by different scoring algorithms (13–22) that aim at ranking variants according to their predicted deleteriousness. Increasingly, there are new sources of relevant genomic data that have not been included in previous analyses. These include the data on human-specific constraints that are revealed by population genomic analyses (7), analyses of 3D organization of the genome (23, 24), and high throughput screens via reporters or CRISPR Cas9 screens (25). On the front of analytical tools, there is a particular emphasis on the use of machine learning (ML) and deep learning (DL), approaches that are optimally suited to manage the sizable sources of information and data. We have described the basic principles for ML and DL for genomics in recent publications (26, 27). The reason to resort to ML and DL approaches for the analysis of the whole genome is because of the inherent flexibility of these computing tools to integrate the many signatures of promoters, enhancers, chromatic marks, topological

associated domains, conservation and constrain and the basic patterns of nucleotide sequence to model the functional genome.

There are today different scoring algorithms that can predict the consequences of genetic variants in the non-coding genome (Table 1). There is however limited experience on the implementation of these tools, their accuracy and clinical usefulness – in part because of limited use of whole genome sequencing today. For all these considerations, the list of rare non-coding genome variants associated with susceptibility to infection, and PIDs – or with any other Mendelian or rare/de novo disease – is at present very limited.

Common non-coding variants modulating susceptibility to infection

Many of the common variants associated with infectious diseases have been identified through GWAS. As of March 2019, the GWAS catalog (https://www.ebi.ac.uk/gwas/home) lists 484 lead variants at a significance p value of $<5xl0^{-8}$ that are associated with 22 different infectious diseases and traits. Only 16 of the 484 (3%) lead variants are located in protein coding regions, 120 (25%) are in the MHC/HLA region – many of which are in regulatory regions; the remaining variants are broadly distributed across the non-coding, regulatory genome (Table 2). Recently, Ramsuran et al. reviewed the literature for reports of infectious disease associations with non-coding variants (28). Overall, the individual contribution (effect size) of these common variants to disease risk is small, and of unclear use to inform clinical care. Many variants would also need confirmatory follow-up studies that secure the validity of the original observation.

On the related field of immunology, where GWAS have identified hundreds of risk loci for autoimmunity, Fahr et al. (29) conducted one of the largest studies of the genetic and epigenetic fine mapping of GWAS common disease variants. Their work indicated that these variants tend to occur near binding sites for master regulators of immune differentiation and stimulus-dependent gene activation, but only 10-20% directly alter recognizable transcription factor binding motifs. Rather, most non-coding risk variants, including those that alter gene expression, affect non-canonical sequence determinants that are not well-explained by current gene regulatory models. Remarkably, 60% of the variants mapped near binding sites for immune-related transcription factors, but rarely altering their cognate motifs. There is a preference for stimulus-dependent T-cell enhancers that respond to immune activation by increasing histone acetylation and transcribing non-coding RNAs. The implicated pathways include antigen presentation, cytokine signaling, and NF-kB transcriptional regulation (29).

A complementary approach to the understanding of the role of common regulatory variants is through the study of expression quantitative loci (eQTL). These are a subset of common variants that are statistically associated with differences in expression of the dependent gene. The analytical approach generally starts by the identification of genes that are differentially expressed *in vitro* or *in vivo* during experimental or clinical infection. The next step is to identify the common variants that associate with interindividual differences in the level of response for the set of genes that are specifically modulated by infection; i.e., eQTLs. The regulatory effects may modify immune activation (30, 31), or the expression of other

regulatory elements such as miRNAs (32). Therefore, this approach combines aspects of candidate gene studies with common variant genetics. While the biological signals tend to be uncontroversial, less is known of the practical consequences of small interindividual differences in gene expression on clinical disease.

More advanced integration of eQTL, epigenetic, and transcriptomic profiling in human immune cells (ie., CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells) serves to demonstrate the occurrence of widespread genetic regulation of immune and host defense pathways overlapping disease loci and involving not only gene expression but also splicing and epigenetic modifications (33). Human genetic variation substantially affects parameters of immune cells, particularly the cell-surface expression of markers conventionally used to identify leukocyte differentiation or activation (34). In another publication, Nédélec et al. reported that a total of 9.3% of macrophage-expressed genes show ancestry-associated differences in the gene regulatory response to infection, and African ancestry predicted a stronger inflammatory response and reduced intracellular bacterial growth (35). More than 75% of ancestry effects on the immune response could be associated with a single cis- or trans-acting expression quantitative trait locus (eQTL). Differences in disease risk may result from regulatory variants exerting their effects only in the presence of immune stress (36). Related to these developments, recent work by Mohammadi et al. (37) introduced a new tool (ANEVA-DOT, analysis of expression variation/ dosage outlier test) to quantify genetic variation in gene dosage. Applied to transcriptome data from individuals with Mendelian muscular diseases led to several potential new diagnoses. This signals the interest to reinforce current diagnostic pipelines with tools that explore the contribution of gene dysregulation to severe disease.

In summary, there is a bulk of information on common non-coding genome regulatory variants that associate with infection disease risk, with immune activation and with interindividual differences in gene expression during infection. This knowledge has translated in a better modeling of regulatory networks during infection and inflammation and is now supporting new diagnostic clinical applications.

Rare non-coding variants as cause of primary immunodeficiencies

PIDs are characterized by recurrent and often life-threatening infections, autoimmunity and cancer (38). Although the most severe forms are identified in early childhood, the majority of patients present in adulthood, typically with no apparent family history and with a variable clinical phenotype of widespread immune dysregulation (38). Because PIDs represent a major diagnostic and therapeutic challenge, it is critical to uncover the genetic basis – both coding and non-coding.

Non coding genome variants may modify known genomic elements such as promoter and 5’ or 3’-UTR regulatory motifs, enhancers, splicing machinery resulting in novel cryptic splice sites, miRNA and long non-coding RNAs, epigenetic signals and 3D chromatin organization. There is a limited number of non-coding pathogenic variants reported across ClinVar, the Human Gene Mutation Database (HGMD) and compiled from recent publications (7, 39–42) that are associated with PIDs and other Mendelian inflammatory

disorders in 29 genes (Table 3). Consistent with a bias to sequence in proximity to exons, the median distance from those variants to the nearest splice site/exon is 109 nucleotides (range 10 to 1,776 bp). It is notable that some of the disease variants map to long non-coding RNAs (IncRNAs). Many IncRNAs are deregulated upon viral infection or IFN treatment, and some of them can modulate viral infection in an interferon-dependent or independent manner (43). Table 3 also provides the expectation of deleteriousness estimated by ncER (non-coding essential regulation score), a machine learning tool trained on functional, essentiality, conservation and structural features of the non-coding genome (44). The median ncER score for the set of variants associated with PIDs and other Mendelian inflammatory disorders in Table 3 is high – median of 93 percentile (the higherthe percentile the more essential the variant is) – with only 5 of 47 variants with percentile values below 80.

A recent report from Thaventhiran et al. (38) specifically addressed the respective role for common and rare variants in coding and non-coding sequences in a cohort of 974 sporadic and familial PID patients as part of the United Kingdom NIHR BioResource – Rare Diseases program. Analysis of coding regions of index cases found disease-causing mutations in known monogenic PID genes in 8.2%. Exploration of the non-coding space revealed deletions in regulatory regions which were thought to contribute to disease causation. It should be indicated that the non-coding genome analysis was limited to the identification of rare structural variants overlapping exon, promoter or 'super-enhancer' *cis*-regulatory elements of known PID genes. No homozygous deletion events affecting *cis*-regulatory elements were identified; however, the study identified 10 candidate compound heterozygotes. The study provided follow up analysis of three such events: a *LRBA* compound heterozygosis variant associated with impaired surface expression of CTLA-4 in Treg cells. The functional consequences of *DOCK8* compound heterozygosis was confirmed by flow cytometry (results were not shown). However, these structural variants encompassed both exon ad regulatory regions. A third study case described premature stop in *ARPC1B* and a 9Kb deletion spanning the promoter region including an untranslated first exon. Western blot analysis confirmed the complete absence of expression of ARPC1B. Overall, the authors defended the utility of whole genome sequencing for detecting compound heterozygosity – that includes a coding variant and a non-coding *cis*-regulatory deletion- for the investigation of PIDs

In the same study, Thaventhiran et al (38) also used a genome-wide association study (GWAS) to show co-localization of, and interplay between, novel high penetrance monogenic variants and common non-coding variants (specifically, at the *PTPN2* and *SOCS1* loci). They argued that their data shed light on the contribution of common variants to variable penetrance and phenotypic complexity in PIDs. We are increasingly aware of the overlap between loci of PIDs and those for inflammatory diseases (eg. autoimmune diseases and arthritis). Fodil and colleagues (45) quantified the intersection between genes mutated in PIDs and GWAS signals for 22 inflammatory diseases. In total, 80 of the 265 identified PID genes (30%) fell within boundaries of GWAS loci. This overlap was highly statistically significant and provided insights into the genetic architecture of infectious and inflammatory diseases. Based on this intersection, Fodli et al. (45) proposed that rare variants in these genes may cause severe disease (PIDs), while the subtle modulation by common regulatory and coding variants may contribute to chronic inflammation.

The number of non-coding variants associated with PIDs will certainly increase, as documented by recent reports. For example, Boisson et al. (46) described a recurrent deep intronic mutation underlying cell type-dependent quantitative NEMO (IKBKG) deficiency. The reported variant creates a new splicing donor site, giving rise to a 44-nucleotide pseudoexon that causes a frameshift and the impairment of NF-κB activation.

Fischer and Rausell (9, 47) classified PID genes within two major groups: those associated with vulnerability to infection, and those associated with immunopathology, including autoimmunity, inflammation, and allergy, with a partial overlap among both categories. Specifically, they assessed genes mutated in PID to correspond to defects in adaptive immunity (n=189), innate immunity (n= 59) and mixed/unknown mechanisms (n-=39). Their work also characterized monogenic diseases causing autoimmunity and inflammation that were classified as resulting in defects in adaptive immunity (n=75) and in innate immunity (n=58) (47). Against these numbers, it is intriguing that, for the set of disorders with variants that may disrupt gene regulation listed in Table 3, we observe a reverse distribution of the type of immune defects: mixed/unknown (n=12), innate immunity (n=11) and adaptive immunity (=6). We can only speculate on the significance of this observation in terms of the underlying immune regulatory mechanisms.

Future directions

Progress in determining the role of regulatory variants in susceptibility to infection is linked to the gradual deployment of whole genome sequencing, to improvements in the interpretation of the consequences of variation in the non-coding genome, and to the establishment of target populations and cohorts for a dedicated diagnostic effort. Certainly, the current 30% diagnostic success with the use of exome sequencing in rare pediatric disorders opens the door to the use of whole genome sequencing. However, we lack consensus or model that establishes the expectation for success on the 70% of probands for whom no diagnosis has been established. The proportion of disease-causing mutations not located in the exome may range from 5% up to 50%, depending on the primary immunodeficiency category (48). One of the first settings to promote the investigation of the non-coding genome is in the study of PIDs that are lacking the “expected” known coding defects – would there be a compound heterozygosis that implicates the significant disruption of a critical regulatory element or topological associated domain?

Another field of interest – both for a broader use of exome and whole genome sequencing – is the study of unexpected severity to a common infection. Examples include the possibility that the most severe presentations of influenza, rhinovirus or respiratory syncytial virus – ie., ICU, death – represent a PID (49–56). Indeed, there are a number of recognized defects of the interferon response associated to few probands, but not yet a large-scale analysis of the proportion of subjects that could have an underlying genetic defect. While many efforts target the pediatric population, there is increasing awareness that adult-onset diseases may also have an underlying genetic lesion. Adult-onset is a known characteristic of common variable immunodeficiency. More recent examples include defects of the inhibitory receptor CTLA4 that present as adult-onset severe immune dysregulation (57) and the description of adult-onset immunodeficiency with anti-interferon-gamma autoantibodies (58) characterized

by disseminated non-tuberculous mycobacteria, non-typhoidal salmonella, cytomegalovirus, *Penicillium marneffei*, and varicella zoster virus.

In the future, the use of whole genome sequencing, possibly combined with transcriptome analysis in the same subjects, new scoring metrics for predicting deleteriousness or pathogenicity in the non-coding genome, and large scale analysis of cohorts of individuals with the most severe manifestations of common infections, should contribute to the characterization of the contribution of regulatory variants to pediatric and adult onset immune disorders without a known etiology.

Acknowledgments

Work of A.T. was supported by the NIH Center for Translational Science Award (CTSA, grant number UL1TR002550).

References

1. Pena LDM, Jiang YH, Schoch K, Spillmann RC, Walley N, Stong N, et al. Looking beyond the exome: A phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases. Genet Med 2018;20:464–9. [PubMed: 28914269]
2. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature 2018;555:611–6. [PubMed: 29562236]
3. Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, et al. Paternally inherited cis-regulatory structural variants are associated with autism. Science 2018;360:327–31. [PubMed: 29674594]
4. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet 2015;24:R102–10. [PubMed: 26152199]
5. Casanova JL, Abel L. Human genetics of infectious diseases: Unique insights into immunological redundancy. Semin Immunol 2018;36:1–12. [PubMed: 29254755]
6. Osterwalder M, Barozzi I, Tissieres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 2018;554:239–43. [PubMed: 29420474]
7. di Iulio J, Bartha I, Wong EHM, Yu HC, Lavrenko V, Yang D, et al. The human noncoding genome defined by genetic diversity. Nat Genet 2018;50:333–7. [PubMed: 29483654]
8. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet 2018;19:51–62. [PubMed: 29082913]
9. Fischer A, Rausell A. Primary immunodeficiencies suggest redundancy within the human immune system. Sci Immunol 2016;1.
10. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30. [PubMed: 25693563]
11. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun 2018;9:5380. [PubMed: 30568279]
12. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 2014;111:6131–8. [PubMed: 24753594]
13. Torkamani A, Schork NJ. Predicting functional regulatory polymorphisms. Bioinformatics 2008;24:1787–92. [PubMed: 18562267]
14. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5. [PubMed: 24487276]

15. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 2016;48:214–20. [PubMed: 26727659]
16. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using gerp++. PLoS Comput Biol 2010;6:e1001025. [PubMed: 21152010]
17. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. Funseq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 2014;15:480. [PubMed: 25273974]
18. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 2015;31:1536–43. [PubMed: 25583119]
19. Smedley D, Schubach M, Jacobsen JO, Kohler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. Am J Hum Genet 2016;99:595–606. [PubMed: 27569544]
20. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931–4. [PubMed: 26301843]
21. Gussow AB, Copeland BR, Dhindsa RS, Wang Q, Petrovski S, Majoros WH, et al. Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. PLoS One 2017;12:e0181604. [PubMed: 28797091]
22. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet 2017;49:618–24. [PubMed: 28288115]
23. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 2013;503:290–4. [PubMed: 24141950]
24. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. Nat Genet 2015;47:598–606. [PubMed: 25938943]
25. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by starr-seq. Science 2013;339:1074–7. [PubMed: 23328393]
26. Telenti A, Lippert C, Chang PC, DePristo M. Deep learning of genomic variation and regulatory network data. Hum Mol Genet 2018;27:R63–R71. [PubMed: 29648622]
27. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51:12–8. [PubMed: 30478442]
28. Ramsuran V, Ewy R, Nguyen H, Kulkarni S. Variation in the untranslated genome and susceptibility to infections. Front Immunol 2018;9:2046. [PubMed: 30245696]
29. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337–43. [PubMed: 25363779]
30. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 2014;343:1246949. [PubMed: 24604202]
31. Kim-Hellmuth S, Bechheim M, Putz B, Mohammadi P, Nedelec Y, Giangreco N, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nat Commun 2017;8:266. [PubMed: 28814792]
32. Siddle KJ, Deschamps M, Tailleux L, Nedelec Y, Pothlichet J, Lugo-Villarino G, et al. A genomic portrait of the genetic architecture and regulatory impact of microrna expression in response to infection. Genome Res 2014;24:850–9. [PubMed: 24482540]
33. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell 2016;167:1398–414 e24. [PubMed: 27863251]
34. Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, et al. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. Nat Immunol 2018;19:302–14. [PubMed: 29476184]

35. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell 2016;167:657–69 e21. [PubMed: 27768889]
36. Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc Natl Acad Sci U S A 2018;115:E488–E97. [PubMed: 29282317]
37. Mohammadi P, Castel SE, Cummings BB, Einson J, Sousa C, Hoffman P, et al. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science 2019;366:351–6. [PubMed: 31601707]
38. Thaventhiran JED, Allen HL, Burren OS, Farmery JHR, Staples E, Zhang Z, et al. Whole genome sequencing of primary immunodeficiency reveals a role for common and rare variants in coding and non-coding sequences. https://wwwbiorxivorg/content/101101/499988v1 2018.
39. Esteller M Non-coding rnas in human disease. Nat Rev Genet 2011;12:861–74. [PubMed: 22094949]
40. Makrythanasis P, Antonarakis SE. Pathogenic variants in non-protein-coding sequences. Clin Genet 2013;84:422–8. [PubMed: 24007299]
41. Gordon CT, Lyonnet S. Enhancer mutations and phenotype modularity. Nat Genet 2014;46:3–4. [PubMed: 24370740]
42. Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. Am J Hum Genet 2016;99:595–606. [PubMed: 27569544]
43. Qiu L, Wang T, Tang Q, Li G, Wu P, Chen K. Long non-coding rnas: Regulators of viral infection and the interferon antiviral response. Front Microbiol 2018;9:1621. [PubMed: 30072977]
44. Wells A, Heckerman D, Torkamani A, Yin L, Ren B, Telenti A, di Iulio J. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat Commun 2019.
45. Fodil N, Langlais D, Gros P. Primary immunodeficiencies and inflammatory disease: A growing genetic intersection. Trends Immunol 2016;37:126–40. [PubMed: 26791050]
46. Boisson B, Honda Y, Ajiro M, Bustamante J, Bendavid M, Gennery AR, et al. Rescue of recurrent deep intronic mutation underlying cell type-dependent quantitative nemo deficiency. J Clin Invest 2019;129:583–97. [PubMed: 30422821]
47. Fischer A, Rausell A. What do primary immunodeficiencies tell us about the essentiality/redundancy of immune responses? Semin Immunol 2018;36:13–6. [PubMed: 29229549]
48. Meyts I, Bosch B, Bolze A, Boisson B, Itan Y, Belkadi A, et al. Exome and genome sequencing for inborn errors of immunity. J Allergy Clin Immunol 2016;138:957–69. [PubMed: 27720020]
49. Albright FS, Orlando P, Pavia AT, Jackson GG, Cannon Albright LA. Evidence for a heritable predisposition to death due to influenza. J Infect Dis 2008;197:18–24. [PubMed: 18171280]
50. Ciancanelli MJ, Huang SX, Luthra P, Garner H, Itan Y, Volpi S, et al. Infectious disease. Life-threatening influenza and impaired interferon amplification in human irf7 deficiency. Science 2015;348:448–53. [PubMed: 25814066]
51. Ciancanelli MJ, Abel L, Zhang SY, Casanova JL. Host genetics of severe influenza: From mouse mx1 to human irf7. Curr Opin Immunol 2016;38:109–20. [PubMed: 26761402]
52. Hernandez N, Melki I, Jing H, Habib T, Huang SSY, Danielson J, et al. Life-threatening influenza pneumonitis in a child with inherited irf9 deficiency. J Exp Med 2018;215:2567–85. [PubMed: 30143481]
53. Sologuren I, Martinez-Saavedra MT, Sole-Violan J, de Borges de Oliveira E Jr., Betancor E, Casas I, et al. Lethal influenza in two related adults with inherited gata2 deficiency. J Clin Immunol 2018;38:513–26. [PubMed: 29882021]
54. Zhou J, Wang D, Wong BH, Li C, Poon VK, Wen L, et al. Identification and characterization of gldc as host susceptibility gene to severe influenza. EMBO Mol Med 2019;11.
55. Lamborn IT, Jing H, Zhang Y, Drutman SB, Abbott JK, Munir S, et al. Recurrent rhinovirus infections in a child with inherited mda5 deficiency. J Exp Med 2017;214:1949–72. [PubMed: 28606988]

56. Asgari S, Schlapbach LJ, Anchisi S, Hammer C, Bartha I, Junier T, et al. Severe viral respiratory infections in children with ifih1 loss-of-function mutations. Proc Natl Acad Sci U S A 2017;114:8342–7. [PubMed: 28716935]

57. Kuehn HS, Ouyang W, Lo B, Deenick EK, Niemela JE, Avery DT, et al. Immune dysregulation in human subjects with heterozygous germline mutations in ctla4. Science 2014;345:1623–7. [PubMed: 25213377]

58. Browne SK, Burbelo PD, Chetchotisakd P, Suputtamongkol Y, Kiertiburanakul S, Shaw PA, et al. Adult-onset immunodeficiency in thailand and taiwan. N Engl J Med 2012;367:725–34. [PubMed: 22913682]

59. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018;50:1161–70. [PubMed: 30038395]

60. Caron B, Luo Y, Rausell A. Ncboost classifies pathogenic non-coding variants in mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biol 2019;20:32. [PubMed: 30744685]

61. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell 2019;176:535–48 e24. [PubMed: 30661751]

Table 1.

Non-coding genome pathogenicity scores.

Score	Year	Data sources	Methodology	Ref
CADD	2014	• Ensembl Variant Effect Predictor • DNase hypersensitivity, TFBS • GC Content, CpG content, etc 64 features	Support vector machine	(14)
FunSeq2	2014	• Inter- and Intra-species conservation. • Loss- and gain-of-function events for transcription factor binding • Enhancer-gene linkage, etc 11 features types	Weighted scoring model	(17)
FATHMM	2015	• 46-way sequence conservation • ChIP-seq, TFBS, DNase-seq • FAIRE, footprints, etc 10 feature groups (up to 1,281 features)	Kernel-based classifier	(18)
Eigen	2016	• ENCODE and Roadmap Epigenomics projects data. • Conservation scores • Allelic frequency, etc 29 features	Unsupervised learning	(15)
ReMM	2016	• Conservation scores • allelic frequency • FANTOM5 enhancers, etc. 26 features	Random forest model	(42)
LINSIGHT	2017	• Conservation scores • predicted binding sites (TFBS, RNA) • FANTOM5 enhancers, etc 48 features	Generalized linear model	(22)
Orion	2017	• Independent from annotation and features. • Genetic human diversity	Observed versus expected site-frequency spectrum	(21)
CDTS	2018	• Independent from annotation and features. • Genetic human diversity	Observed versus expected heptameric variation rate	(7)
PrimateAI	2018	• common variants from population sequencing of six non-human primate species	Deep neural network	(59)
ncER	2018	• 3D chromatin structure • enhancer reporter assays • metrics of genome essentiality, etc 38 features	Supervised learning using gradient tree boosting (XGBoost model)	(44)
NCBoost	2019	• interspecies conservation • Recent and ongoing purifying selection signals in humans • GC and CpG content, etc 53 features	Supervised learning using gradient tree boosting (XGBoost model)	(60)
SpliceAI	2019	• pre-mRNA transcript sequences	Deep neural network	(61)

Table 2.

Common variants associated with infectious diseases.

Data from the GWAS catalog (March 2019) on the number of associated variants (n=484, strongest/lead risk allele) with 22 infectious agents and disease traits (*).

Region	No. variants*	%
Intron	194	40
HLA**	120	25
Intergenic	96	20
Regulatory region	33	7
Coding	16	3
Non-coding transcript	9	2
TF-binding site	7	1.5
5UTR	5	1
3UTR	4	1

* Associated with: acne, bacterial meningitis, chickenpox, enteric fever, Epstein-Barr virus, Hepatitis B, Hepatitis C , HIV-1, HPV, Influenza A, leprosy, meningococcal disease, mononucleosis, mumps, nontyphoidal salmonellosis, pneumococcal bacteremia, rubella, malaria, shingles, *Staphylococcus aureus*, strep throat, tuberculosis, urinary tract infection.

** More than half of the HLA associations map to non-coding regions in MHC.

Table 3.

Non-coding variants associated with primary immunodeficiencies and inflammatory disorders.

Variants are compiled from Wells et al. (44). Inclusion of these variants in the list is not a definitive guarantee that these variants are confirmed as pathogenic. ncRNA=non-coding RNA, UTR=untranslated region. When multiple annotations overlapped, the priority was as follows: ncRNA > UTR > Promoter and Promoter Flanking > Intron > Histone marks. ncRNA variants are reported in association to a PID protein-coding gene because the variant is positioned in one of the non-protein coding isoforms of the protein coding gene. Such isoform are annotated in gencode as “processed_transcript”, “retained_intron” or “non-sense mediated decay”. ncER= non-coding essential regulation score. A score of “100” represents maximal estimate of deleterious functional consequences of variation. ncER values are available at https://ai-omni.com.

Chrom	Position (1based)	Distance to splice site	Annotation	ncER	Syndrome	Gene	Variant	Source	Category*	OMIM
1	16091760	31	Intron	90	Chronic multifocal osteomyelitis	FBLIM1	C>T	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/259680
1	92949058	86	ncRNA	90	Neutropenia, severe chronic	GFI1	T>G	HGMD DM High	Innate	https://www.omim.org/entry/613107
10	14966845	1776	Intron	61	Immunodeficiency, severe combined, atypical	DCLRE1C	C>G	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/602450
10	90770494	15	ncRNA	90	Autoimmune lymphoproliferative syndrome	FAS	A>G	HGMD DM High	Innate	https://www.omim.org/entry/601859
12	6443045	12	ncRNA	93	Periodic fever, autosomal dominant	TNFRSF1A	C>T	HGMD DM High	Innate	https://www.omim.org/entry/142680
13	51501530	12	Intron	99	Aicardi-Goutieres syndrome	RNASEH2B	G>A	HGMD DM High	Innate	https://www.omim.org/entry/610181
16	3306599	289	UTR	87	Mediterranean fever, familial	MEFV	G>C	HGMD DM High	Innate	https://www.omim.org/entry/249100
16	3306969	659	Histone marks	89	Mediterranean fever, familial	MEFV	G>C	HGMD DM High	Innate	https://www.omim.org/entry/249100
19	17943239	88	Intron	75	Immunodeficiency, severe combined	JAK3	C>T	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/600802
2	47249223	104	Intron	32	Gastrointestinal defects and immunodeficiency syndrome	TTC7A	T>A	Mendelian manually curated	Mixed/ Unknown	https://www.omim.org/entry/243150
2	47606078	13	Intron	91	Diarrhea 5 with tufting enteropathy congenital	EPCAM	A>G	Mendelian manually curated	Mixed/ Unknown	https://www.omim.org/entry/613217
2	98349927	79	ncRNA	93	Immunodeficiency 48	ZAP70	G>A	Mendelian manually curated	Adaptive	https://www.omim.org/entry/269840
20	25388397	135	ncRNA	85	Immunodeficiency 55	GINS1	A>G	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/617827

Chrom	Position (1based)	Distance to splice site	Annotation	ncER	Syndrome	Gene	Variant	Source	Category*	OMIM
20	25388409	123	ncRNA	90	Immunodeficiency 55	GINS1	C>G	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/617827
20	31395557	10	Intron	99	Immunodeficiency, centromeric instability and facial anomalies syndrome	DNMT3B	G>A	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/242860
20	43248503	14	Intron	100	Immunodeficiency, severe combined	ADA	A>T	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/102700
21	34668714	226	ncRNA	64	Inflammatory bowel disease	IL10RB	C>T	HGMD DM High	Innate	https://www.omim.org/entry/612567
21	46321660	11	ncRNA	93	Leukocyte adhesion deficiency	ITGB2	A>C	HGMD DM High	Innate	https://www.omim.org/entry/116920
3	128202131	571	Intron	100	Immunodeficiency 21	GATA2	G>A	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/614172
3	128202163	539	Intron	100	Immunodeficiency 21	GATA2	del	Mendelian manually curated	Mixed/ Unknown	https://www.omim.org/entry/614172
5	35867853	287	Intron	32	Immunodeficiency, severe combined	IL7R	G>A	HGMD DM High	Adaptive	https://www.omim.org/entry/608971
5	40931143	22	Intron	85	Complement C7 deficiency	C7	T>A	HGMD DM High	Innate	https://www.omim.org/entry/217070
5	63258025	91	Intron	89	Periodic fever menstrual cycle dependent	HTR1A	del	Mendelian manually curated	Innate	https://www.omim.org/entry/614674
8	48844056	708	Intron	85	Immunodeficiency 26 with or without neurologic abnormalities	PRKDC	dupA	Mendelian manually curated	Mixed/ Unknown	https://www.omim.org/entry/615966
9	317025	17	Intron	95	Hyper-IgE recurrent infection syndrome	DOCK8	C>G	HGMD DM High	Adaptive	https://www.omim.org/entry/611432
9	317028	14	Intron	95	Hyper-IgE recurrent infection syndrome	DOCK8	T>G	HGMD DM High	Adaptive	https://www.omim.org/entry/611432
9	35657746	737	Promoter	83	Immunodeficiency, primary, Cartilage-hair hypoplasia	RMRP	A>G	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/250250
X	100609705	22	Intron	97	Agammaglobulinaemia	BTK	T>C	HGMD DM High	Adaptive	https://www.omim.org/entry/300300
X	100629415	108	ncRNA	88	Agammaglobulinaemia	BTK	G>T	HGMD DM High	Adaptive	https://www.omim.org/entry/300300
X	100629827	204	Intron	93	Agammaglobulinaemia	BTK	T>C	HGMD DM High	Adaptive	https://www.omim.org/entry/300300
X	100641212	163	UTR	100	Agammaglobulinaemia	BTK	T>C	HGMD DM High	Adaptive	https://www.omim.org/entry/300300

Chrom	Position (1based)	Distance to splice site	Annotation	ncER	Syndrome	Gene	Variant	Source	Category*	OMIM
X	135730217	347	Histone marks	96	Immunodeficiency X-linked with hyperIgM	CD40LG	A>C	Mendelian manually curated	Adaptive	https://www.omim.org/entry/308230
X	135736517	14	Intron	97	Immunodeficiency X-linked with hyperIgM	CD40LG	T>A	HGMD DM High	Adaptive	https://www.omim.org/entry/308230
X	153788599	22	Intron	93	Ectodermal dysplasia with immune deficiency	IKBKG	A>T	HGMD DM High	Mixed/ Unknown	https://www.omim.org/entry/300291
X	37639262	114	Promoter Flanking	100	Chronic granulomatous disease	CYBB	A>C	Mendelian manually curated	Innate	https://www.omim.org/entry/306400
X	37639264	112	UTR	100	Chronic granulomatous disease	CYBB	T>C	Mendelian manually curated	Innate	https://www.omim.org/entry/306400
X	37639266	110	UTR	100	Chronic granulomatous disease	CYBB	C>T	Mendelian manually curated	Innate	https://www.omim.org/entry/306400
X	37639267	109	UTR	100	Chronic granulomatous disease	CYBB	C>T	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	37641330	10	Promoter Flanking	98	Chronic granulomatous disease	CYBB	T>G	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	37654041	977	Intron	93	Chronic granulomatous disease	CYBB	G>T	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	37656474	1079	Promoter Flanking	93	Chronic granulomatous disease	CYBB	A>G	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	37656731	841	Promoter Flanking	80	Chronic granulomatous disease	CYBB	T>G	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	37657051	521	Intron	87	Chronic granulomatous disease	CYBB	A>G	HGMD DM High	Innate	https://www.omim.org/entry/306400
X	49114969	16	UTR	99	Immunodysregulation polyendocrinopathy and enteropathy X-linked	FOXP3	G>T	Mendelian manually curated	Innate	https://www.omim.org/entry/304790
X	70327278	494	ncRNA	97	Severe combined immunodeficiency X-linked	IL2RG	T>C	HGMD DM High	Adaptive	https://www.omim.org/entry/300400
X	70330553	14	Intron	93	Severe combined immunodeficiency X-linked	IL2RG	T>C	HGMD DM High	Adaptive	https://www.omim.org/entry/300400
X	70331494	149	Promoter	100	Severe combined immunodeficiency X-linked	IL2RG	G>A	Mendelian manually curated	Adaptive	https://www.omim.org/entry/300400

*Category is defined as in Fischer and Rausell (47).