ABSTRACT

Title of Dissertation:	INTERPRETING GENETIC VARIANTS FOR DISCOVERING DISEASE ETIOLOGY AND MECHANISMS
	Kunal Kundu, Doctor of Philosophy, 2020
Dissertation directed by:	Professor John Moult Department of Cell Biology and Molecular Genetics

High-throughput sequencing methods now provide extensive data on disease-related human genetic variants. New methods are required to maximally utilize these data for enhanced understanding and treatment of human diseases. This dissertation describes my work in addressing three aspects of this challenge: Determining disease-causative variants; representing mechanisms by which genetic variant(s) cause disease phenotypes; and quantitatively analyzing genetic disease mechanisms.

First, I developed a variant prioritization algorithm, VarP, and objectively tested it in CAGI (Critical Assessment of Genome Interpretation). It was ranked best in the CAGI challenge on interpreting panel sequencing data for 106 patients, determining which disease class each patient has and the corresponding causative variant(s). VarP correctly identified the disease class for 36 cases, including 10 where the original

clinical pipeline failed, and found seven cases with strong evidence of an alternative disease to that tested. Over-reliance on pathogenicity annotations in the HGMD mutation database led to several incorrect cases. Post analysis showed that protein structure data could have helped to interpret the impact of many prioritized missense variants.

Next, I co-developed and implemented MecCog, a web-based graphical framework to represent mechanisms by which genetic variants cause disease phenotypes. A MecCog mechanism schema displays the propagation of system perturbations across stages of biological organization, using graphical notations to symbolize perturbed entities and activities, knowledge gaps, ambiguities and uncertainties, and hyperlinked evidence. The web platform enables a user to construct, store, publish, browse, query, and comment on schemas. MecCog facilitates better comprehension of disease mechanisms, identification of critical unanswered questions on causal relationships, and possible new sites of therapeutic intervention.

Finally, I developed a framework to quantitatively represent and analyze mechanisms relating genetic variants to complex trait disease. It involves generating a computable circuit from MecCog schemas by assigning node functions and parameters to represent the behavior of the schema components. I demonstrate that such a circuit can be used to analyze the effect size of a variant contributing to disease risk as a function of the genetic background in an individual and the extent to which epistatic effects may be masked in population averages. I also show that the circuit functions

and parameters can be learned in a data-driven manner using a hybrid neural network approach.

INTERPRETING GENETIC VARIANTS FOR DISCOVERING DISEASE ETIOLOGY AND MECHANISMS

by

Kunal Kundu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor John Moult, Chair
Professor Lindley Darden
Professor Niklas Elmqvist
Associate Professor Stephen Mount
Assistant Professor Max Leiserson

Foreword

I made substantial contributions to the relevant aspects of the jointly authored work included in the dissertation.

Chapter 1 was published in the *Human Mutation*:

Kundu, K., Pal, L. R., Yin, Y., & Moult, J. (2017). Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge. *Human Mutation*, *38*(9), 1201–1216.

My contribution: computational experiment and data analysis.

Dedication

To my grandfather (Binoy Kumar Aush), grandmother (Chhabi Aush), father (Mukteswar Kundu), and mother (Ruma Kundu)

Acknowledgments

I would like to sincerely thank my advisor, John Moult, for giving me the opportunity in the first place to work with him. He is an amazing mentor, a brilliant scientific thinker, and a great human being, and so my Ph.D. journey with him has been a lifetime experience that I will always cherish. Through a variety of projects, he taught me how to do science. His supportive and caring attitude motivated me to take risks and never be afraid of failure. His passion for science is infectious that always gave me positive energy to be an avid learner and solve hard problems. As a whole, during the past five years, he gave me an environment to grow as a scientist.

I would like to thank Professor Lindley Darden for mentoring me and providing continuous valuable feedback on my progress during my Ph.D. years. She has been a friend, philosopher, and guide to me. She introduced me to the world of philosophy of biology and it was a terrific collaboration with her. She has always been a source of encouragement for me.

I would like to thank the members of my committee for all their support and valuable feedback on my projects. My deep gratitude to Professor Niklas Elmqvist, Professor Stephen Mount, Professor Hector Corrada Bravo, and Professor Max Leiserson. I would also like to thank my teaching coordinators Professor Hadiya Woodham and Professor Carmen Cantemir-Stone for giving me the opportunity for the phenomenal teaching experience at the University of Maryland.

I would like to thank Professor Osnat Herzberg for introducing me to life beyond the Ph.D. activities. She has always been a good listener and provided me valuable practical suggestions in life whenever I needed it. She always motivated me to work hard. She showed me the world of protein crystallography.

I would like to thank my lab members Dr. John Norvell, Dr. Lipika Ray, and Dr. Yizhou Yin for supporting me emotionally and intellectually in numerous instances during my Ph.D. journey. Also, a big thanks to the IBBR IT team members, Christian Presley, Maya Zuhl and Thomas Lane, for helping me with cluster computing issues.

I would like to thank my parents for making it all happen. They always tripled the joy of my success stories during the Ph.D. and motivated me to stay focused and work hard for doing better. A special remembrance and thanks to my late grandfather who had encouraged me to do Ph.D. My work in this dissertation is a tribute to him.

Lastly, thanks to the Starbucks café at Kentland, Gaithersburg for being my second home that kept me going!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Rare variants and human disease

1.1.1 Genetic variants in the human genome

Human genetic variants are differences found in DNA sequences between individuals within and among populations. Accumulation of DNA mutations due to uncorrected DNA replication errors, and exogenous and endogenous factors (such as chemicals, ionizing radiation, and oxygen free radicals) are major sources of genetic variants. Knowledge of the human genome sequence together with advances has made way for executing multiple large-scale initiatives on characterizing genetic variants in the human genome. These include the 1000 Genomes Project (1KG) (2504 individuals) (Auton et al., 2015), the Exome Sequencing Project (ESP) (7034 individuals) (Auer et al., 2016), and the Genome Aggregation Database (gnomAD) (141,456 individuals) (Karczewski et al., 2020). It has been found that a typical genome has 4.1 million to 5.0 million sites that differ from the reference human genome (Auton et al., 2015). The primary types of genetic variant are single nucleotide variants (SNVs), short indels, structural variants such as large deletions, copy-number variations (CNVs), and mobile insertion elements (MEIs). Based on 1000 genome data (Devuyst, 2015), typically a genome has ~4.31 million (median) SNVs, ~625K (median) indels, and an estimated 2100 to 2500 structural variants. Although SNVs and short indels forms >99.9% of the variant set, the structural variants affect more bases.

In the coding region of the human genome, SNVs may cause synonymous (no amino acid change) and non-synonymous (amino acid change resulting in missense, nonsense due to creation of termination codon, and stoploss due to loss of a termination codon) variations, and indel and structural variants may cause loss of function (LoF) variation such as frameshift (change in the reading frame of ORFs), and non-frameshift (no change in the reading frame of ORFs) variations. These coding variants may also cause aberrant splicing by changing regulatory splice sites such as exon splicing enhancer or silencer. In the non-coding region of the human genome, variants may alter splice sites or gene expression regulatory sites. In a typical genome, about 0.3% of genetic variants are missense (~12000), 0.3% are synonymous (~ 13000), 0.004% are LoF (~180), and 12% are in regulatory sites (such as promoter, insulator, enhancer, transcription factor binding sites) (~500000) (Auton et al., 2015). Genetic variants are also shown to affect long-range intra-chromosomal functional connections (Smemo et al., 2014). Minor allele frequency (MAF) analysis of the variants in populations shows that the majority of variants observed in a single genome are common (i.e. MAF > 0.5%) and only 1 to 4% are rare variants having MAF < 0.5%. Exome Sequencing Project (ESP) data shows that in an individual most coding variants (e.g. missense, synonymous or nonsense) are rare variants (MAF<1%) and the majority of these are missense (Auer et al., 2016).

As of August 2020, the two widely used human genetic variant databases, dbSNP (Sherry et al., 2001)and gnomAD (Karczewski et al., 2020), contain 700 million and 229.9 million variants respectively. These large counts are because of the total

amount of sequencing data being produced, currently doubling approximately every seven months, establishing human genomics as a big data domain (Stephens et al., 2015). Knowledge of these variants has provoked intense scientific interest in their use to obtain insights into human genetics and diseases.

1.1.2 Role of rare variants in monogenic diseases

Monogenic diseases are caused by mutations in one gene, so exhibiting a Mendelian inheritance pattern (therefore also referred to as Mendelian diseases). For example, Lynch syndrome, a hereditary nonpolyposis colorectal cancer, is caused by heterozygous mutations in any of the DNA mismatch repair (MMR) genes *MLH1, MSH2, MSH6,* or *PMS2* (Sehgal et al., 2014). These diseases are individually rare (incidence of 1 in 10000 live births on the upper end; based on data from Orphanet - https://www.orpha.net/) but impact millions of individuals and families (Baird et al., 1988; Carter, 1977), with over ~6200 distinct disease traits known to date (Amberger et al., 2019). The World Health Organization (WHO) reports that the global prevalence of all single-gene diseases at birth is very high (approximately 1/100) thus making this disease type of major scientific interest. Monogenic diseases are classified into three categories: (A) Dominant disease where one of the two copies of a gene is damaged, (B) Recessive disease where both the copies of a gene are damaged, and (C) X-linked disease where the defective gene is on the X chromosome. Data in the Online Mendelian Inheritance in Man (OMIM) (August 2020) databases show that for ~91% of the rare Mendelian diseases the inheritance pattern is not reported and in the set where it is reported, dominant diseases (670) are

the majority followed by recessive (101). Studies have revealed that these rare diseases can also be characterized by more complex modes of inheritance such as digenic inheritance (variants at two distinct loci required for trait manifestation), dual molecular diagnoses (variants at two distinct loci lead to two independent segregating traits), multilocus mutational burden (effect of a highly penetrant variant modified by variation at additional loci), and compound inheritance of rare and common variants (trait requires one rare and one common variant) (Posey, 2019). The OMIM data also shows that only ~20% (3949/~20000) of human protein-coding genes have been associated with one or more monogenic diseases. It is reported that genetic testing based on these disease genes yields a molecular diagnosis in only 24% of the tested patients (Lionel et al., 2018) and new disease genes are only being discovered at a rate of 263 a year (Posey et al., 2019). This indicates that there is a tremendous amount of research that remains to be done to elucidate the molecular etiology of monogenic diseases.

Many rare and novel (not previously seen) SNVs, short insertions/deletions, CNVs, and structural variants have been identified to underlie monogenic diseases. Commercial (such as the Human Gene Mutation Database (HGMD) (Stenson et al., 2017)) and public (such as ClinVar (Landrum et al., 2018)) databases have been built to collate medically important variants associated with monogenic diseases and susceptibility to complex trait diseases. HGMD (August 2020) catalogs ~275K variants, of which SNVs (including missense, nonsense, and splicing) are the majority (67%), followed by small deletions (14%), small insertions (6%), and complex

rearrangement and repeat variations (1%). ClinVar (August 2020) archives a smaller set of ~89K pathogenic and likely pathogenic variants but has a similar trend as HGMD with SNVs being the majority (70%), followed by deletions (17%), insertion/duplication (8%), copy number (2%) and structural variation (5%). To catalog somatic variants in cancer, expert-curated (such as COSMIC (Tate et al., 2019)) and community-driven (such as CIVIC (Griffith et al., 2017)) databases have also been built.

It has been estimated that clinical exome sequencing (CES, covering exomes and flanking regions) should capture about 95% of variants that cause genetic disorders (Shamseldin et al., 2017). This should make it possible to investigate the impact of variants on cis- control of expression, splicing, and protein level mechanisms using this type of data. Many such mechanisms have been elucidated. For example, a rare (MAF=0.007 in gnomAD) non-frameshift deletion variant rs113993960 (NM_000492:c.1521_1523del) in the *CFTR* gene is known to cause cystic fibrosis (CF) with a recessive inheritance pattern (Wang & Li, 2014). This variant is present in the 11^{th} exon (11/27) of the gene and causes a deletion of three base pairs that leads to the loss of amino acid, phenylalanine, at position 508 in the CFTR protein (1480 amino acids). Studies have reported that the mutation causes a twofold problem (https://www.ncbi.nlm.nih.gov/books/NBK540352/) that results in loss of chloride channel function: (a) A defect in the protein conformation due to misfolding results in the degradation of the mutant protein before it can reach to the cell surface (Cutting, 2015; X. R. ober. Wang & Li, 2014); (b) Misfolded protein that escapes degradation

has a reduced half-life compared to wildtype protein (Cutting, 2015). Understanding such molecular mechanisms aids in devising therapeutic strategies. For example, TriKafta (Bear, 2020), a triple combination drug containing tezacaftor, elexacaftor, and ivacaftor, is being used as one of the therapies for CF patients. Tezacaftor is a CFTR corrector that acts as a pharmacological chaperone that promotes forward trafficking of F508del-CFTR to the cell surface (Hanrahan et al., 2017). Elexacaftor is also a CFTR corrector that binds to an alternate site than tezacaftor to facilitate the trafficking of mutant CFTR to the cell surface (Ridley & Condren, 2020). Ivacaftor is a CFTR potentiator as it increases the probability of the CFTR channel open conformation, so increasing the chloride ion flow (Condren & Bradshaw, 2013; Hanrahan et al., 2017).

1.1.3 Interpretation of rare variants in clinical settings

Sequencing technologies (whole-genome, whole-exome, or targeted/panel sequencing) have recently become more available for clinical diagnostic testing of monogenic diseases and cancer. As of August 2020, the Genetic Testing Registry (https://www.ncbi.nlm.nih.gov/gtr/) contains entries for about 575 labs worldwide and so far a total of 76268 tests on 18695 genes for 16424 conditions. The current diagnostic yield ranges widely from 21 to 73%, depending on the phenotypes tested (Gilissen et al., 2014; Lionel et al., 2018; Posey, 2019; Soden et al., 2014; Taylor et al., 2015). Most clinical labs follow a semi-automated approach for variant interpretation, by first making use of available variant annotation and prioritization tools and then checking putative causative variants for association with the disease of

interest in databases and the literature (Sadedin et al., 2015) (https://blog.goldenhelix.com/golden-helix-end-to-end-architecture-for-clinical-testing-labs). There are dozens of commercial and open-source variant annotation and prioritization tools available that identify putative causative variants by considering factors such as inheritance pattern, minor allele frequency, genomic region (coding/non-coding), mutation type, and in silico impact prediction for missense/splicing mutations (Hu et al., 2019). However, it has been demonstrated that there are substantial discrepancies between existing tools (McCarthy et al., 2014; Pabinger et al., 2014). For instance, on comparing results between two widely used annotations tools (Annovar (K. Wang et al., 2010) and VEP (McLaren et al., 2016)) it was found that 35% of the LoF variants and 13% of all exonic variants had mismatched annotations, with splicing variants having the greatest discrepancies (McCarthy et al., 2014). The choice of transcript database (RefSeq or Ensembl) for an annotation tool caused 21% and 17% discrepancies for the Lof + missense and all exonic variants respectively. These discrepancies illustrate that there is scope for improved genome interpretation accuracy through further development of the tools and improved annotation.

To standardize assignment of variant pathogenicity in clinical labs, the American College of Medical Genetics and Genomics (ACMG) has developed guidelines for weighing the evidence of pathogenicity for a variant, with classification into one of five categories: pathogenic, likely pathogenic, uncertain significance, likely benign and benign (Richards et al., 2015a). In the current version of ClinVar (August 2020)

the majority of the variants are classified as ‘uncertain significance’ (41%) and with only 12% are classified as ‘pathogenic and likely pathogenic’. Additionally, there are 5% variants with ‘conflicting interpretation of pathogenicity’ for cases with conflicting ACMG classification outcomes from different sources. A recent analysis of variant reclassification over time showed that predominantly variants are being reclassified to ‘conflicting interpretation of pathogenicity’ from all other types (~5000 per year) and that there is very low (~150 per year) reclassification traffic towards ‘pathogenic or likely pathogenic’ types (Shah et al., 2018). These data show a need for improving methods for consistently and accurately assigning pathogenicity. One of the areas where there is most potential for improvement is the use of computational methods. These are currently down-weighted in the ACMG guideline compared to the experimental evidence because of low accuracy.

In order to objectively assess methods for interpreting the impact of genetic variants, John Moult and Steven Brenner started the Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/) in 2010 (https://doi.org/10.1038/news.2010.679). CAGI is an organization that conducts community experiments to test methods for relating genotype to phenotype. Participants are asked to predict particular phenotypes, given genetic variant information. The corresponding results are not released until all participants have submitted their predictions; thus, these are *bona fide* blind predictions. Independent experts assess the predictions and the outcomes are discussed at a CAGI conference. CAGI challenge datasets have included germline and somatic variants from whole-

genomes, whole-exomes, clinical gene panels, and phenotypes that have covered rare monogenic diseases, complex trait diseases, and particular types of cancer. One of these challenges is of particular interest in this dissertation. The aim was to determine which of 14 monogenic disease classes each of 106 patients has together with the corresponding causative variants, given each patient's gene panel sequencing data (https://genomeinterpretation.org/content/4-Hopkins_clinical_panel), obtained from a genetic testing laboratory. Motivated to build a more accurate variant interpretation method that can be used for the clinical diagnosis of rare monogenic diseases, I designed and implemented an open-source variant prioritization pipeline and assessed its performance on the CAGI gene panel challenge dataset. Chapter 2 describes the design and implementation of the prioritization pipeline and its assessment in CAGI.

1.2 Representation of genetic disease mechanisms

1.2.1 Genetic disease mechanisms

Variant annotation pipelines, when successful, provide insight into some of the low-level molecular mechanisms involved in the disease. Development of effective treatments, such as in the cystic fibrosis example above, is greatly facilitated by knowledge of the full succession of causal links across levels of biological organization by which a DNA change leads to a disease phenotype, not just the molecular stage steps. Advanced experimental model systems (such as cell lines (Pansarasa et al., 2018), organ-on-chips (Santoso & McCain, 2020), organoids (Lancaster & Huch, 2019), and model organisms), multi-omics data (such as epigenetic, transcriptomic, and proteomic), imaging techniques (Femminella et al.,

2018; Pegoraro & Misteli, 2017), and bioinformatics approaches (such as mathematical modeling and network analysis (Kikuchi et al., 2015; Zhang et al., 2013)) are facilitating the discovery of these complete causal mechanisms. But so far, as discussed below, the field lacks a comprehensive framework for describing and evaluating the mechanisms.

1.2.2 Mechanism representation in literature and digital platforms

To illustrate the need for tools for describing disease mechanism, consider the example of a mutation that causes Lynch Syndrome, a heritable form of colon cancer introduced earlier (Bartosova et al., 2003). The mechanism starts with a novel (not reported in the gnomAD or 1000 Genomes databases) heterozygous nonsense variant rs63750245:C>T in a DNA mismatch repair gene, *MSH2*, (Bartosova et al., 2003): This variant is in the sixth exon (6/16) of the gene and creates a premature termination codon in the mRNA. The codon position corresponds to the first half (p.Gln344Ter) of the MSH2 protein (934 amino acids). This causes nonsense-mediated decay of the *MSH2* mRNA and so leads to a decreased abundance of MSH2 protein in the cell. Normally, MSH2 protein interacts with MSH6 and MSH3 proteins to form MutSα and MutSβ complexes respectively. These complexes are involved in repairing single nucleotide variants (SNVs), and small (up to 13 nucleotides long) insertion/deletions in the genome (Acharya et al., 1996; Drummond et al., 1995; Gupta et al., 2012; Lang et al., 2011; Martín-López & Fishel, 2013; Umar et al., 1996). Perturbation of MSH2 protein abundance affects the abundance of these complexes and hence decreases the mismatch repair activity. That leads to the

accumulation of more variants, including those that alter the activity of the oncogenes/tumor suppressor genes, so increasing the chances of cancer.

This informal description of how a *MSH2* variant affects cancer risk is derived from considering the information reported in 17 publications. These publications contain varying combinations of structured and unstructured data and use many different diagrammatic representations. The scattered nature of information about the mechanisms by which human variants affect phenotypes is a general trend: querying for well-studied disease-related variants such as '*CFTR* rs113993960' (X. R. ober. Wang & Li, 2014), '*FTO* rs1558902' (Shimaoka et al., 2010) and '*NOD2* rs2066847' (Hugot et al., 2001a; Ogura et al., 2001) on LitVar (a PubMed and PMC search engine for genetic variant data) (Allot et al., 2018) returns ~100 – 200 publications. Multiple projects have addressed the resulting knowledge integration problem, including the building of disease-specific knowledge managements resources (for example alzforum.org (Kinoshita & Clark, 2007)), development of natural language processing (NLP) based texting mining methods (DARPA's Big Mechanism program (Cohen, 2015)), development of statistical methods for genotype-phenotype evidence integration (Konopka & Smedley, 2020), and the community-driven systems medicine disease maps project (Mazein et al., 2018). Each of these contributes elements of a solution, but none provides a systems-level representation of the disease mechanisms in a clear, precise, and comprehensive manner.

There have also been major technological advances in the development of tools to represent biological mechanism descriptions. These include graphical notations (SBGN (Systems Biology Graphical Notation) (Novère et al., 2009a)) and computer-readable languages (SBML (Systems Biology Markup Language) (Hucka et al., 2018), KGML (KEGG Markup Language) (https://www.genome.jp/kegg/xml/docs), RDF (Resource Description Framework), BEL (Biological Expression Language, https://bel.bio/)) to encode representations, software to draw/visualize models (GO-CAM (Thomas et al., 2019), PathWhiz (Pon et al., 2015), and Cytoscape (P. Shannon et al., 2003)), linked data formats such as Nanopublications (http://nanopub.org/) to organize provenance and metadata for scientific assertions (Mina et al., 2015), and database management systems to store and query graph-based representations (Neo4j - https://neo4j.com/). These tools have helped in the creation of pathway databases (such as KEGG (Minoru Kanehisa et al., 2016) and Reactome (Fabregat et al., 2017)), causal activity models (GO-CAM (Thomas et al., 2019)), causal biological networks (Boué et al., 2015), and knowledge graphs that integrate information about bio-entities (genes, compounds, and diseases) and their relationships (https://het.io/).

Most of these resources overlay disease mechanism information on depictions of 'normal' biological pathways. Two examples are shown in figure 1-1, for KEGG and Reactome depictions of Lynch syndrome related mechanisms. Figure 1-1 A is part of the KEGG disease pathway map of colorectal cancer (accession: hsa05210), showing the relationship between the inactivation of DNA mismatch repair genes (such as *MLH1* and/or *MSH2*) and genome instability in this cancer type. This type of KEGG

cancer pathway map is created by adding graphical indicators to represent the disease-related perturbations on top of the normal pathway representation. In this instance, the colorectal cancer map was created by combining parts of nine normal pathways, including those for the cell cycle, apoptosis, and p53 signaling.

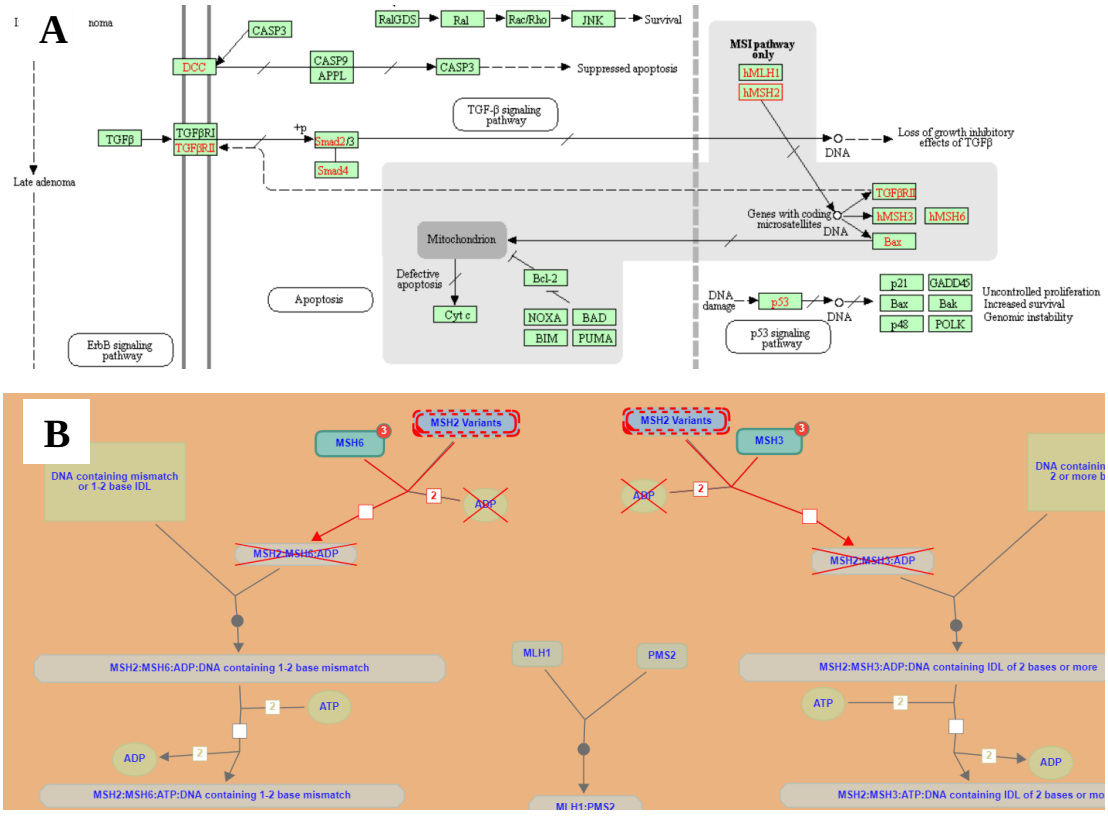


Figure 1-1. Disease mechanism representations in pathway databases. Both types of representation add disease mechanism symbols to ‘normal’ pathways and only show perturbations at the molecular level. Figure 1-1 A shows part of the colorectal cancer pathway in KEGG (accession: hsa05210) illustrating gene products (green boxes) and their interactions (e.g. protein-protein interactions) or relations (e.g. expression, repression) (black arrows). Red represents disease-associated gene products and

arrows with cut marks represent missing or reduced interactions due to mutations. Disease-related processes (such as anti-apoptosis and proliferation) are annotated in the pathway. Figure 1-1 B shows part of the Defective Mismatch Repair Associated With MSH2 pathway in Reactome (accession: R-HSA-5632928.1). Green boxes represent proteins, a red boundary box indicates the presence of a genetic variant, light blue boxes represent complexes, ovals represent small molecules, black arrows represent normal reaction types, red crosses represent perturbed entities, and red arrow represent perturbed reaction types caused by the genetic variants.

In the map, the 'disease genes' (for instance *hMSH2* and *hMLH1*) are colored red and interactions affected by mutations in them are indicated by adding cut marks to the corresponding 'normal' arrows. Figure 1-1 B shows part of the Reactome disease pathway for defective mismatch repair associated with *MSH2* (accession: R-HSA-5632928.1). This disease pathway was created by adding perturbation indicators to the normal state of the mismatch repair pathway (accession: R-HAS-5358508.1). Perturbed entities are indicated by red crosses, perturbed reaction types by red lines, and the presence of genetic variants is indicated by red boundary boxes.

These two pathway representations describe disease mechanisms as perturbations at the molecular level only and do not show how perturbations propagate through the higher stages of biological organization to cause a disease phenotype. This issue is partly addressed by the systems medicine disease map project: these disease maps provide an encyclopedic description of disease mechanisms involving signaling,

metabolic, and gene regulatory processes at different levels of granularity including molecular, subcellular, cellular, tissue, organs, and organism. But they do not represent the perturbation propagation across the levels (Mazein et al., 2018). Currently, disease maps have been generated for 14 diseases, for example, cystic fibrosis, Parkinson's disease, asthma, and lung cancer. The interactive graphical interface includes a feature to zoom in/out from cellular to molecular stages, showing the mapping of information across the stages.

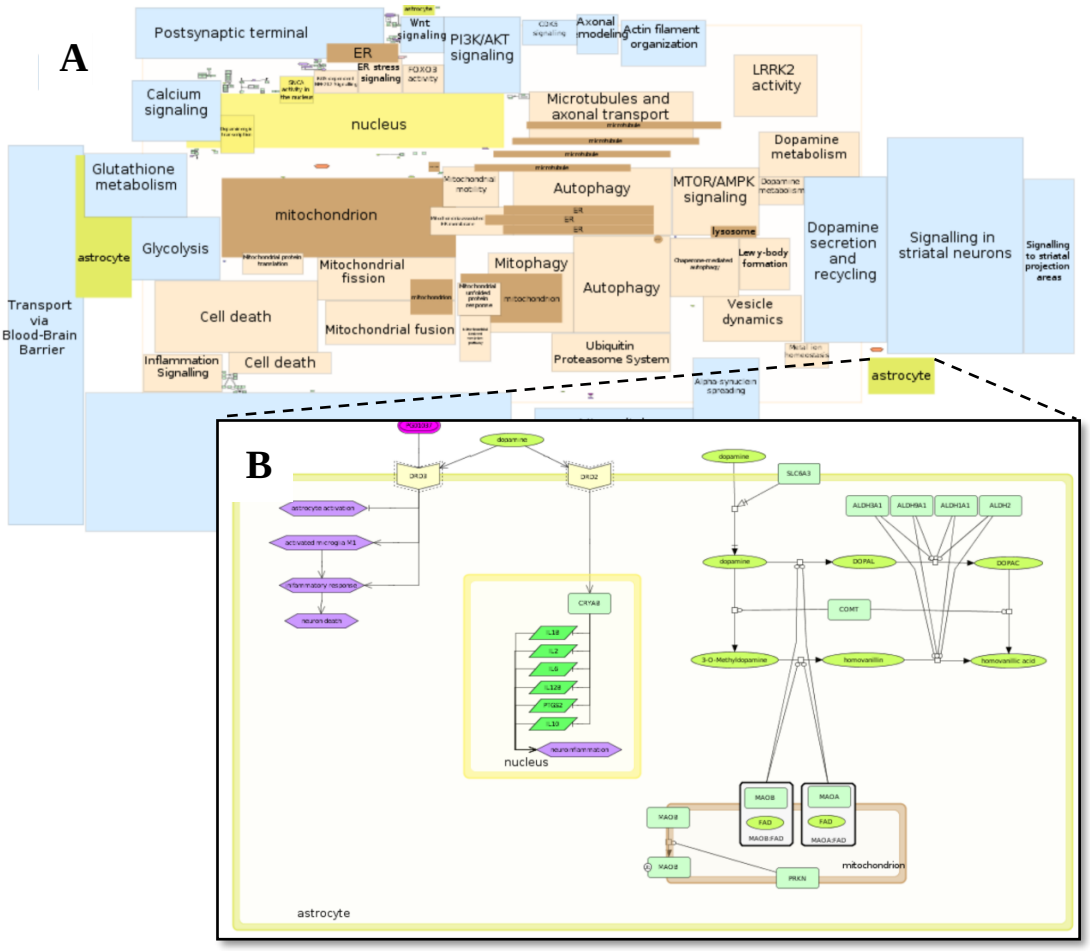


Figure 1-2. Parkinson's disease map. Figure 1-2 A shows the top-level view, displaying disease-related cellular components (such as mitochondria and the nucleus

of a neuron), cellular processes (such as autophagy, cell death, calcium signaling) and cell types (such as astrocyte). Figure 1-2 B shows the zoomed-in molecular level view inside the astrocyte. Green boxes represent proteins, green parallelograms represent mRNAs, green ovals represent small molecules, purple colored hexagons represent phenotypes, the pink oval shows a drug, and arrows represent reaction types (e.g. association, binding, dissociation, transition). Sub-cellular locations (such as the nucleus and mitochondria) are also shown.

Figure 1-2 shows a view of the Parkinson's disease map. The top-level view (Figure 1-2 A) displays a mosaic of disease-related cell types, and related cellular components and processes. Zooming into a specific tile displays the normal molecular interactions (Figure 1-2 B) but not the perturbed interactions of the diseased system. Uncertainties, ambiguities, and ignorance in mechanistic knowledge are not presented in the disease maps (only Reactome pathway diagrams label uncertain reaction types). This is a serious omission since such knowledge gaps exist in almost all disease mechanisms (Greenberg & Amato, 2004; Kametani & Hasegawa, 2018).

The two primary deficiencies of these disease mechanism representations–lack of means of displaying the propagation of disease-related perturbation across stages of biological organization and absence or inadequate inclusion of knowledge gaps, uncertainties and ambiguities–motivated us to develop a new graphical framework for disease mechanisms. Design goals for this framework include depiction of mechanism components across stages of biological organization; an as simple as

possible representation (only included features directly related to disease mechanism); an intuitive way to visualize ignorance, uncertainties, and ambiguities in the mechanistic information; and tight linkage to evidence in the literature and databases. Chapter 3 discusses the design and implementation of this framework.

1.3 Quantitative modeling of complex trait diseases

1.3.1 Quantitative modeling of biological systems

Biological systems are formally complex systems that have interconnected and interdependent components orchestrating non-linearly to produce system behavior (such as regulation of genes, or induction of immune response) in response to inputs (Hillmer, 2015). To understand the emergent properties of these systems, either a *reductionist* approach is used where small modular subsystems are discovered and studied (for example (Süel, 2011)), or a *synthesis* approach is used that involves approximating a complex system via a tractable number of components (Ellner & Guckenheimer, 2006) (for example, (Tsuda et al., 2009)). In both approaches, it may be possible to use descriptive sentences to represent relationships within or between system components, but exhaustively deriving the implications of these relationships is prohibitively laborious, error-prone, and imprecise. Therefore, quantitative models are built to numerically describe how system components cross-talk and control system output. These models are imperfect but provide a virtual version of a system that can be tested to see if it captures salient features (Hillmer, 2015). Successful models of this type also allow investigation of system response to varied external conditions and internal perturbations, insights that otherwise are far more labor-

intensive, costly, and sometimes impossible to achieve experimentally. Model building involves first choosing appropriate mathematical and/or computational methods that best capture the nature of the biological system and then following an iterative procedure (model train-predict-test-repeat all) to fine-tune the model properties. Quantitative modeling has a long history and has been used for studying a wide range of biological systems including tissue models of the human heart (Kohl & Noble, 2009), explaining the chemical basis of morphogenesis (Turing, 1952), and predicting disease spread (Keeling, 2005). Another important application of quantitative modeling has been to understand the mechanisms of complex trait diseases.

Complex trait diseases (for example hypertension, type 2 diabetes, and Crohn's disease) are caused by multiple genetic (DNA variants), epigenetic (such as DNA methylation), and environmental factors that perturb the interactions between components located within and across stages of biological organization so as to manifest the disease phenotype. Many computational modeling approaches have been devised to precisely (though not necessarily accurately) model the disease state and so obtain insights into the underlying mechanisms. Three primary approaches are network modeling, genome-scale metabolic modeling (GEM), and kinetic modeling: (1) In network modeling, construction of biomolecular (such as gene-gene and protein-protein) association networks is carried out based on large scale data (such as RNA-Seq co-expression, and affinity-purification – mass spectrometry (AP-MS)), sometimes for both the disease and healthy states of a system. The networks are then

analyzed to discover and characterize the biological significance of the highly connected genes/proteins often using packages such as WGCNA (Langfelder & Horvath, 2008), and also to identify differences in network structure between healthy and disease states using a differential network approach (Ideker & Krogan, 2012). Network modeling has resulted in many large-scale networks used to investigate the disease states, for example (Greene et al., 2015; Daniel S. Himmelstein & Baranzini, 2015; Thul & Lindskog, 2018). As these network models are built on data measured under a specific stimulus to a biological system, inferring the new network properties (such as edge weights) of the model for other instances of the stimulus, for which data does not exist or is hard to generate, can be challenging for this modeling approach. (2) In genome-scale metabolic modeling (GEM), organism-specific stoichiometric-based metabolic reaction networks are compiled and analyzed to predict metabolic fluxes using linear programming. The latest human GEM (Recon3D (Brunk et al., 2018)) contains 5835 metabolites, 10600 reactions, and 2248 genes. GEMs can be used to identify the impact of the disease-associated genes by knocking out the gene/reaction during flux balance analysis (N. E. Lewis et al., 2012). GEM models have been used to study types of cancer (such as breast cancer (Gámez-Pozo et al., 2017), prostate cancer (Asgari et al., 2018; Marín de Mas et al., 2018)), as well as chronic diseases (such as type 2 diabetes (J. Sarkar et al., 2019; Väremo et al., 2015)). Despite these developments, it has been reported that GEM-based simulations alone are insufficient to provide insights into disease mechanisms and there is a need for a computational framework that allows simultaneous simulation of material flow (metabolic network) and information flow (gene regulatory and signaling networks)

to capture the highly complex cues and cascading of signals in the diseased system (Gu et al., 2019).

(3) In kinetic modeling, in contrast to genome-scale modeling, a smaller set of reactions with known kinetic parameters (such as rate and affinity of reactions) are defined by nonlinear differential equations or partial differential equations to indicate changes in product concentration during a time period (Resat et al., 2009). These models have been developed to understand very specific dynamic aspects of a disease such as T cell autoreactivity in autoimmune diseases (Ramos et al., 2019), mononuclear phagocyte system function in systemic lupus erythematosus (Meryhew et al., 1986), and amyloid formation in prion disease (Come et al., 1993). Large kinetic models are also being built to capture the interactions for understanding system-wide properties (Bordbar et al., 2015). However, several issues (Miskovic et al., 2015) have been noted as the size of the model increases such as difficulties in estimating values of the large number of parameters, because of uncertainty in available data but also the intrinsic 'sloppiness' of these systems, implying a need for very high accuracy for some parameters in a kinetic model (Gutenkunst et al., 2007).

Although these modeling approaches are useful to study aspects of biological systems, none of them demonstrates the capacity to build large multiscale integrative models that can capture emergent properties of a complex disease at each scale (Tiwary, 2020). Recently, an integrated model for yeast cells (DCell, http://d-cell.ucsd.edu/) (J. Ma et al., 2018a) has been built using an old concept of a hybrid neural network (Psichogios & Ungar, 1992). In 1992, a hybrid neural network model

was first used to model a bioreactor, combining prior knowledge with a neural network to estimate the unknown process parameters. It was shown that the hybrid model had better properties than standard 'black-box' neural network models in terms of being easier to analyze and interpret, and requiring significantly fewer training data. In the yeast DCell model, prior knowledge is in the form of the Gene Ontology (Ashburner et al., 2000) hierarchical structure of 2526 biological subsystems in a eukaryotic cell. Those data are integrated with a deep neural network. The model is claimed to simulate growth phenotypes in response to gene knockout(s) as accurately as laboratory observations (J. Ma et al., 2018b). A key feature of the DCell model is interpretability: in spite of the use of a deep neural network, because of the biology-based architecture, genotype/phenotype relationships can be interpreted in terms of perturbed subsystems and their interactions. Thus, DCell provides a proof-of-concept on the feasibility and utility of this type of hybrid model for studying biological systems.

The fourth chapter focuses on developing a computational framework for building integrated quantitative models of complex diseases using a hybrid neural network. The approach takes as input results from the mechanism representation framework (MecCog (Darden et al., 2018)) described in the third chapter, where the mechanisms by which genetic variants cause disease phenotypes are represented as perturbation propagation across stages of the biological organization.

1.3.2 Genetic variants relating to complex trait disease

As sequencing and genotype technologies have advanced, methods for finding genetic variants (primarily single nucleotide polymorphisms (SNPs)) associated with the complex trait diseases have evolved. Initially, family-based linkage analysis was used to identify chromosomal regions containing the relevant genes. For example, the IBD1 risk locus on chromosome 16 for Crohn's disease (CD) was discovered in this way and later was more finely mapped to determine *NOD2* as the susceptibility gene (Hugot et al., 1996, 2001b). However, the overall results from this approach were poor because it could not be applied to finding all relevant genes across the whole genome. The advent of genotyping microarray technologies made it possible to screen hundreds of thousands of genetic variants in case and control populations allowing genome-wide association studies (GWAS). GWAS analyses have also been performed to screen for genetic variants associated with continuous traits, such as blood pressure (Yan Wang & Wang, 2018), body mass index (Willer et al., 2009), age at menarche (He et al., 2009), and height (Allen et al., 2010). A variety of statistical approaches have been devised to identify disease-associated variants in GWAS (Hayes, 2013). As far as possible, population structure and other confounding effects are taken into account in these approaches. The incidence of false positives is reduced by validation using an independent dataset. Because of linkage disequilibrium (LD), a GWAS variant found to be associated with a phenotype is not itself likely to be involved in disease mechanism, but is likely in LD with a variant that is. For example, the Illumina Human Omni2.5S-8 chip with only 2.5 million SNPs represents ~48% of the ~7.8 million SNPs in an Asian human genome (Ha et

al., 2014)). A fine-mapping process is often used to identify the genetic variants in the LD region around each GWAS associated variant that are most likely to causally influence the examined trait (Schaid et al., 2018). Approaches like GWASeq, a targeted re-sequencing follow-up to GWAS loci, (Salomon et al., 2016) are used for this purpose. Variant findings from the many GWAS studies are curated and maintained in the GWAS Catalog database (https://www.ebi.ac.uk/gwas/). As of September 2018, the database contains 71673 variant-phenotype associations based on 5687 GWAS studies (Buniello et al., 2019).

Unlike the rare causative variants of monogenic diseases, where each variant usually has a large impact on the function of a single protein, the functional effects of complex trait variants contributing to a complex trait are usually more subtle and often not yet known (Cleynen & Halfvarsson, 2019). Functional interpretation of these variants is challenging because: (a) the majority of associated variants are located in the non-coding region of the genome with only a small fraction in the coding region, making it hard to identify the affected genes and the ways in which their function is affected (Edwards et al., 2013; Giral et al., 2018), and (b) the variant effect can be influenced by gene-gene and gene-environment interactions. Currently, functional genomics datasets (Cano-Gamez & Trynka, 2020), chromatic organization datasets (Soskic et al., 2019), and epigenetic datasets (Tak & Farnham, 2015) are being used to assist in functionally interpreting the statistical association of these variants with the disease phenotype. For example, a previous study on predicting variant mechanisms (such as splicing, gene expression, or protein function altering) in

seven complex diseases (bipolar disorder, coronary heart, Crohn's disease, hypertension, rheumatoid arthritis, type 1 & 2 diabetes), using an expression quantitative trait loci (eQTL) dataset, and in silico tools to predict variant impact, found possible mechanisms for 76% of the 356 disease-associated loci (Pal et al., 2015).

1.3.3 Limitations of GWAS

Although GWAS has been extremely effective at identifying variants associated with complex traits, it has a number of limitations. One is the fact that observations are unavoidably made against the varied genetic background found in a human population. For example, the effect size of phenotype-associated variants is an average over all individuals in the sampled population (Stringer et al., 2011). In model organisms, many instances (Galardini et al., 2019; Vu et al., 2015) have been reported on the differences in single variant effect size as a function of genetic background. For example, in *Drosophila melanogaster* the severity of the retinitis pigmentosa disease phenotype (as measured by the eye size) caused by a missense mutation (G69D) in the rhodopsin gene (*Rh1*) has a strong Drosophila-strain dependent effect, with eye size varying from ~14K to ~28K pixels (Chow et al., 2016). In the fourth chapter, the variation in effect size of human GWAS variants as a function of genetic background and its consequences for GWAS are investigated using a quantitative model of complex disease outlined in the previous section.

A second GWAS limitation is that despite continuous efforts to discover and interpret GWAS variants, these variants explain only a small proportion of heritability – the portion of phenotypic variance in a population attributable to genetic factors (Kendler & Neale, 2009). Heritability of a disease phenotype is zero if it is fully dependent on environmental factors and is one if it is only determined by genetic factors. For the complex trait diseases, the heritability is between 0 and 1 and is often estimated from twin studies. Only a small proportion of the estimated heritability is explained by GWAS variants, for example (Manolio et al., 2009), only 20% for Crohn's disease (Barrett et al., 2008), 15% for systemic lupus erythematosus (Harley et al., 2008), and 6% for type 2 diabetes (Zeggini et al., 2008). Many hypotheses (Maher, 2008) have been put forward to explain the reasons for missing heritability. One explanation often advanced is the role of epistatic interactions in which one or multiple genes influence(s) the effect of another. These effects are not captured by GWAS. The fourth chapter investigates the extent of epistatic interactions using the quantitative model for complex disease.

1.3.4 Epistatic interactions between genetic variants

Three different types of epistasis have been proposed (Phillips, 2008):

(A) *Functional epistasis* describes the interaction between proteins, either directly in the form of protein complexes or indirectly by operating within the same pathway.

(B) *Compositional epistasis* describes the altering of the effect of one allele by an allele at another locus, in the presence of a specific genetic background. In model organisms, many systematic studies (Baryshnikova et al., 2013; Onge et al., 2007;

Sopko et al., 2006) have been conducted to analyze compositional epistatic effects between gene pairs. For example, using a Synthetic Genetic Array (SGA) analysis that enables large-scale construction and selection of yeast double-mutant strains, the majority of all possible yeast gene pairs (~18 million) revealed a network consisting of nearly one million genetic interactions (Costanzo et al., 2016). More recently, using CRISPR (clustered regularly interspaced short palindromic repeats)-based combinatorial loss-of-function screens, epistatic interactions in human cancer cell lines have been analyzed to (i) identify potential targets for synthetic lethal-based cancer therapy (Najm et al., 2018; Shen et al., 2017), and (ii) identify drug target genes for combinatorial therapies in cancer (Han et al., 2017). However, measuring compositional epistasis at the human population level is not possible because of the very varied genetic backgrounds.

(C) *Statistical epistasis* describes the average effect of combinations of alleles at different loci estimated over the diverse genetic background found in a population. Because of the dependence of the epistatic effect size on genetic background, the average will underestimate the phenomena in some individuals and overestimate it in others. For example, in *Escherichia coli*, it has been shown that the size of the epistatic effect between two beneficial mutations varies drastically across strains (Yinhua Wang et al., 2013).

Quantitative models of the relationship between GWAS risk variants and a phenotype such as those outlined earlier effectively incorporate functional epistatic effects – physical and pathway interactions between proteins. And these quantitative models

potentially can provide a way to determine the compositional epistasis for pairs of variants in each individual from GWAS data - the size of the epistatic effect in each specific genetic background. Results in the fourth chapter of this dissertation demonstrate this integration of functional, compositional, and statistical epistasis approaches, with an analysis of epistatic effects at the individual level across a human population.

1.4 Overview

The dissertation is organized as follows. In Chapter 2, I start by summarizing the use of targeted sequencing approaches in genetic testing for clinical diagnosis of monogenic diseases and highlight the inadequacy of the computational methods used to interpret the clinical significance of the genetic variants. I then introduce the CAGI gene panel challenge that our lab participated in, and for which I developed a variant prioritization pipeline for identifying causative variants from gene panel sequencing data. I provide a detailed CAGI assessment report of the variant prioritization pipeline performance and discuss how its performance may be improved. In Chapter 3, I summarize the inadequacy of existing representations for describing disease mechanisms at the system level. I then introduce the theory of the MecCog framework for graphically representing disease mechanisms. I describe the web-based implementation of MecCog and illustrate its use for qualitative representation of disease mechanisms. In Chapter 4, I summarize the use of the MecCog framework in constructing disease mechanism graphs and as a use-case describe the mechanism graph for the barrier integrity subprocess in Crohn's disease. I describe a quantitative

encoding technique to generate computable circuits from mechanism graphs and demonstrate the use of a hybrid neural network approach to learn properties of the circuit in a data-driven manner. I show a use-case of how such a disease mechanism-based circuit can be used to analyze epistatic interactions between genetic variants. In Chapter 5, I summarize the conclusions of the three projects and describe the future perspectives on improving genetic disease diagnosis, standardizing evidence of pathogenicity for disease causative variants, ways of scaling the disease mechanism representations in MecCog, and broader use of the quantitatively encoded disease mechanism graphs in complex trait disease risk assessment.

Chapter 2: Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge

Published:

Kundu, K., Pal, L. R., Yin, Y., & Moult, J. (2017). Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge. *Human Mutation, 38*(9), 1201–1216.

My contribution: computational experiment and data analysis

2.1 Abstract

The use of gene panel sequence for diagnostic and prognostic testing is now widespread, but there are so far few objective tests of methods to interpret these data. We describe the design and implementation of a gene panel sequencing data analysis pipeline (VarP) and its assessment in a CAGI4 community experiment. The method was applied to clinical gene panel sequencing data of 106 patients, with the goal of determining which of 14 disease classes each patient has and the corresponding causative variant(s). The disease class was correctly identified for 36 cases, including 10 where the original clinical pipeline did not find causative variants. For a further seven cases, we found strong evidence of an alternative disease to that tested. Many of the potentially causative variants are missense, with no previous association with

disease, and these proved the hardest to correctly assign pathogenicity or otherwise. Post analysis showed that three-dimensional structure data could have helped for up to half of these cases. Over-reliance on HGMD annotation led to a number of incorrect disease assignments. We used a largely *ad hoc* method to assign probabilities of pathogenicity for each variant, and there is much work still to be done in this area.

2.2 Introduction

Genetic testing in clinical laboratories is becoming increasingly common: As of March 2017, GeneTests.org contains entries for about 706 labs and 1,083 clinics worldwide performing a total of 67,187 tests on 5,926 genes for 4,963 genetic conditions. So far though, there has been only limited testing of method efficacy (Cornish & Guda, 2015; Hwang et al., 2015; McCarthy et al., 2014; Pirooznia et al., 2014). Many of the genetic tests use targeted gene sequencing panels for identifying variants in a set of genes or gene regions that are known to be associated with a disease (Kammermeier et al., 2014; Okazaki et al., 2016). In clinical laboratories specializing in specific diseases or classes of disease, panels provide high coverage data for genes of interest at relatively low cost, and also reduce the issues in reporting incidental findings to patients. A key and challenging step in all these tests is the ability to accurately interpret the genetic variants and assign a likelihood of pathogenicity (Richards et al., 2015a).

Potentially pathogenic sequence variants fall into three classes: (a) those almost certain to cause major loss of protein function (LoF), arising from the introduction of premature stop codons, frameshifts caused by small insertions or deletions, and direct hits on splice sites; (b) those that may or may not significantly affect gene regulation (such as regulatory variants at transcription factor binding sites) or protein function, particularly missense variants; and (c) those that are more likely benign, particularly synonymous, UTR, and deep intronic variants. The main challenge lies in understanding the phenotypic consequences of the large fraction of variants falling into the last two classes. Most clinical laboratories follow a semi-automated approach for variant interpretation, first making use of available variant annotation and prioritization tools and then checking the potential causative variants' association with the disease of interest in databases and the literature. For the first step, there are dozens of annotation and prioritization tools (open-source or commercial) available (for example, Wang et al. 2010; Cingolani et al. 2012; Sifrim et al. 2013; Robinson et al. 2014; McLaren et al. 2016), typically providing potentially causative variants based on inheritance pattern, allele frequency, genomic region of interest, mutation type and in silico analysis of the likely impact of missense mutations. It has been demonstrated that there are substantial discrepancies between existing annotation tools (McCarthy et al., 2014; Pabinger et al., 2014) so that there is a clear need to encourage and monitor advances in this field. In most clinical laboratories, standard guidelines such as those from the American College of Medical Genetics and Genomics (ACMG) (Richards et al., 2015a) are followed for variant interpretation and reporting. Although the guidelines accept computational predictions of

pathogenicity for variants, these are only considered a ‘supportive’ evidence. Other evidence is required to classify a variant as causative. As a consequence, the overall contribution of computational methods for variant classification is low and this motivates the development and testing of more accurate methods for variant interpretation.

CAGI (Critical Assessment of Genome Interpretation) is an organization that conducts community experiments to objectively assess computational methods for predicting phenotypic impacts of genomic variation (https://genomeinterpretation.org/). The most recent round of experiments (CAGI4) included a challenge to determine which of 14 disease classes each of 106 patients has and the corresponding causal variants, given each patient’s gene panel sequencing data (https://genomeinterpretation.org/content/4-Hopkins_clinical_panel). The gene panel dataset consists of exons with flanking regions and some complete intron sequencing data for 83 genes from each patient. Data were provided by the Johns Hopkins DNA Diagnostic Laboratory. The Laboratory is a CLIA and CAP certified, Maryland, New York, and Pennsylvania licensed clinical genetic testing laboratory specializing in rare, inherited disorder testing (http://www.hopkinsmedicine.org/dnadiagnostic/tests/).

The data were made available to registered CAGI participants, and all were required to deposit disease and variant assignments by a specified deadline. The anonymized submissions were assessed by John-Marc Chandonia

(http://enigma.lbl.gov/chandonia-john-marc/) and Shamil R. Sunyaev (http://genetics.bwh.harvard.edu/wiki/sunyaevlab/), and results were later discussed at the CAGI4 conference. A paper on the assessment is part of this CAGI special issue of Human Mutation (Chandonia et al., 2017).

The identification of causal variants requires a number of carefully controlled procedures for assessing the quality of the data, accurate variant annotation, handling of unphased genotypes, and an appropriate probability model that can prioritize primary and secondary disease findings. With these considerations in mind, we developed a new variant prioritization pipeline (implemented in Python) called VarP (https://github.com/kunduk/VarP) using a combination of open-source and in-house software tools for analyzing gene panel sequencing data. This pipeline was the most successful of those used in CAGI, in the sense that it resulted in the correct matching of the highest number of panel exomes to disease class. [https://genomeinterpretation.org/sites/default/files/protected_files/4-Hopkins_clinical_panel_assessor1_AAdhikari_remixable.pptx]. Nevertheless, the results are far from perfect. In this chapter , we describe the design and implementation of the variant prioritization pipeline and the results obtained.

2.3 Materials and Methods

2.3.1 Capture bed files, gene panel sequencing data, and the disease class

The Johns Hopkins DNA Diagnostic Laboratory panel sequencing procedure generates sequence for all exons plus a boundary of 50 bases up and down stream and some introns for 83 genes (1350 exonic and 39 intronic regions), covering 14 monogenic disease classes. 73 of these genes are known to harbor mutations for one of the 14 monogenic disease classes. The remaining ten genes are known to harbor mutations for two or more disease classes. Sequences had been captured using one of the two custom probe sets (Agilent SureSelectXT Target Enrichment Kit) and sequenced using Illumina MiSeq to generate paired-end reads (2X100 nt reads). Two capture bed files (v01, v02) describing the two probe sets were provided as part of the challenge. The Hopkins group called sequence variants and produced two VCF files for each patient, one a gVCF for single nucleotide variants (SNVs; using GATK UnifiedGenotyper, v2.7-4) and the other a VCF for insertion-deletion variants (Indels, GATK HaplotypeCaller, v2.7-4). For the challenge, all VCF files from 106 patients had been combined into two files, one each for SNVs and Indels.

2.3.2 Building the gene list for disease classes

All the genes annotated in the two capture bed files (v01 and v02) were extracted to compile a list of genes to examine. The description of 14 disease classes was provided on the challenge webpage (https://genomeinterpretation.org/sites/default/files/protected_files/4-Hopkins_clinical_panel_disorders.pdf). We made extensive use of the Hopkins' DNA

Diagnostic Laboratory website to map genes to disease class (http://www.hopkinsmedicine.org/dnadiagnostic/). The website lists a number of gene panel tests and also gives a detailed description of the genes associated with each disease as well as their inheritance pattern. Using this resource we were able to group 53 of the 83 genes to 12 disease classes and obtain the inheritance pattern. We used literature and the Genetic Home Reference Database (http://ghr.nlm.nih.gov/) to group another 24 genes to some of the disease classes and obtain the inheritance pattern. In total 77 out of 83 genes were grouped among the 14 disease classes as shown in Table 2-1. The remaining 6 genes (*DHODH, TRIM37, EFTUD2, AMACR, AGXT* and *CAT*) are associated with diseases that are not related to any of the 14 disease classes and therefore were excluded from any downstream analysis.

Table 2-1. The 14 disease classes and genes identified as relevant to each class. Genes associated with more than one disease class are indicated by an asterisk.

Disease Class	Gene List
Cystic Fibrosis and CF-related disorders	CA12, CFTR, *SCNN1A, *SCNN1B, *SCNN1G
Diffuse Lung Disease	ABCA3, AP3B1, CSF2RA, CSF2RB, *DKC1, FOXF1, HPS1, HPS4, NKX2-1, SFTPA2, SFTPB, SFTPC, SLC7A7, *TERC, *TERT, *TINF2
Primary Ciliary Dyskinesia	CCDC103, CCDC39, CCDC40, DNAAF1, DNAAF2, DNAAF3, HEATR2, DNAH11, DNAH5, DNAI1, DNAI2, DNAL1, HYDIN, LRRC6, NME8, RSPH4A, RSPH9
Peroxisomal Beta-Oxidation Defects	ACOX1, HSD17B4, SCP2
Rhizomelic Chondrodysplasia Punctata	AGPS, GNPAT, PEX7, PHYH
Zellweger Spectrum Disorders	DNM1L, PEX1, PEX2, PEX3, PEX5, PEX6, PEX10, PEX12, PEX13, PEX14, PEX16, PEX19, PEX26
Loeys-Dietz Syndrome	*TGFBR1, *TGFBR2, TGFB2
Marfan Syndrome	*FBN1, *TGFBR2
Thoracic Aortic Aneurysm and Dissection (TAAD)	*FBN1, ACTA2, MYH11, MYLK, SMAD3, *TGFBR1, *TGFBR2, COL3A1
Ataxia Telangiectasia	ATM
Liddle Syndrome	*SCNN1B, *SCNN1G
Pseudohypoaldosteronium Type 1	NR3C2, *SCNN1A, *SCNN1B, *SCNN1G
Telomere Shortening Disorders	CTC1, NHP2, NOP10, *TERC, *TERT, *TINF2, WRAP53, *DKC1
Treacher Collins and Related Syndrome	POLR1D, TCOF1, POLR1C, SF3B4

2.3.3 Gene Panel Sequencing data analysis pipeline

The method developed for this challenge (VarP - **Var**iant **P**rioritization) uses open-source and in-house software tools to analyze gene panel sequencing data with respect to rare genetic disorders in an automated manner. The method has four modules – Variant annotation, QC (quality control) analysis, Variant Prioritization, and estimation of the probability of each variant being causative for the disease. The four modules were executed in a sequential manner (Fig. 2-1). The inputs were the two VCF files and a gene configuration file that contained the genes associated with

each disease class and their inheritance pattern (autosomal dominant/recessive, compound heterozygous, pseudoautosomal dominant/recessive, X-linked dominant/recessive).

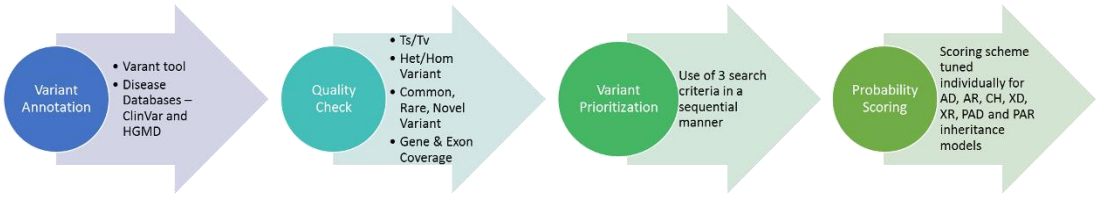


Figure 2-1. The **Var**iant **P**rioritization (VarP) Method. Circles represent the four modules. Modules are executed sequentially starting from Variant Annotation and ending with Probability Scoring. The 'Varant' tool in step 1 annotates variants with genomic region of occurrence, mutation type, minor allele frequency and prediction of pathogenicity for variants. Ts/Tv=Transition/Transversion, Het/Hom=Heterozygous/Homozygous AD=Autosomal Dominant, AR=Autosomal Recessive, CH=Compound Heterozygous, XD=X-linked dominant, XR=X-linked recessive, PAD=Pseudo Autosomal Dominant, and PAR=Pseudo Autosomal Recessive.

Variant Annotation: The two VCF files (one for SNVs and another for Indels) were annotated using Varant (http://compbio.berkeley.edu/proj/varant, doi:10.5060/D2F47M2C). Details on Varant are provided in the Appendix. Varant annotated each variant in the VCF files with region of occurrence (intron, exon, splice site or intergenic), observed minor allele frequencies (MAF) from ExAC (Lek et al.,

2016) and 1000 Genomes Phase-3 (Auton et al., 2015), mutation type (missense, nonsense, silent, frameshift and non-frameshift indels), predicted impact on protein function, and previously associated phenotypes reported in ClinVar (Landrum et al., 2016). Varant used dbNSFP (v2.9) (Jian et al., 2014) database to fetch the mutation impact predictions from PolyPhen-2 (v2.2.2) (Adzhubei et al., 2013), SIFT (release Jan, 2015) (Kumar et al., 2009) and CADD (v1.2) (Kircher et al., 2014). The RefGene (Pruitt et al., 2014) gene definition file was used for gene and transcript annotations. The principal isoforms of each gene were taken from the APPRIS database (Rodriguez et al., 2013). In addition, the VCF files were annotated with SNPs3D (May, 2015) (Yue et al., 2006) mutation impact predictions, HGMD (version June 2014) (Stenson et al., 2003) disease-related variants and with dbscSNV (Jian et al., 2014) variants that potentially alter splicing.

Quality control Analysis: Three types of QC analyses were run on the Hopkins' dataset. The first QC analysis is a comparison of Transition vs. Transversion ratio (Ts/Tv), Heterozygous vs. Homozygous variants (Het/Hom), no call sites vs. low quality sites and common vs. rare vs. novel variant counts across all 106 samples and with those in a control variant set from 2,504 samples in 1000 Genomes Phase-3 (Auton et al., 2015). No call sites (sites where neither reference nor alternate allele was called) and low-quality sites (sites not marked PASS and/or genotype quality less than or equal to 30) per sample were computed from the challenge gVCF file. A variant is considered novel if it was not present in the 1000 Genomes and ExAC (Lek et al., 2016) dataset and considered rare if present with an MAF of less than 5% in

both of these datasets. Other 1000 Genomes or ExAC variants were considered common. Only SNVs flagged as PASS in the VCF file and with a genotype quality (GQ) greater than 30 were included in the analysis. Scatter plots were generated to represent the results. The QC module also estimated which samples are of African ethnicity, to aid in interpretation of variant count differences. The ethnicity analysis used the population-specific allele frequency (AF) from the 1000 Genomes Phases-3 dataset to identify population enriched variants (i.e. variants that are common (AF > 0.05) in a population but rare (AF <= 0.05) in other populations). Samples whose African population enriched variant count was highest in number compared to other populations in 1000 Genomes (Admix American, South Asian, East Asian and European) were assigned African ethnicity. The second QC analysis is a comparison of the average read depth for 83 genes across 106 samples, using the read depth provided in the gVCF file. The module produced a heat-map of these data, allowing convenient visual inspection for anomalies. The third QC analysis identifies capture regions (exon or intron) with anomalous read depth with respect to other captured regions in the same gene, where the anomaly is found in at least 85% of the samples. Anomalous coverage was identified by first computing the average read depth across the gene (μ) and its standard deviation (σ), and then checking each region for significantly low ($< \mu - 2\sigma$) or high ($> \mu + 2\sigma$) coverage. The anomalous coverage regions were then visually inspected using gene coverage plots.

Identification of potentially causative variants: Only rare or novel variants rated high quality (marked PASS and with a GQ > 30 in the VCF files) were considered in the

search for causal variants. At this stage, a rare variant was defined as one reported in ExAC (Lek et al., 2016) with a minor allele frequency (MAF) less than or equal to 0.01 and a novel variant was defined as one not found in ExAC. Indels in low complexity regions (LCR) were excluded from the analysis, based on the LCR dataset computed for the human genome by Heng Li (Heng Li, 2014). For each sample, each QC qualified variant in each of the 83 genes was assigned to one of four categories, ranked by the likelihood that the variant is causative.

Category 1: Variants reported in HGMD with either DM (disease-causing mutation) or DP (disease-associated polymorphism) status, and/or reported in ClinVar with pathogenic or likely pathogenic clinical significance.

Category 2: Variants annotated as nonsense mutations, direct splicing mutations disrupting either a splice donor or acceptor site, frameshift or non-frameshift causing Indels, splice altering variants predicted in the dbscSNV database, and missense mutations predicted as damaging by one or more of SNPs3D, SIFT, PolyPhen-2 and CADD.

Category 3: Variants annotated as missense but not predicted to be damaging by any of the above methods, and UTR and intronic variants.

Category 4: All other variants (including synonymous and all with MAF > 0.1). These were not considered as potentially causative.

Each variant was also grouped by frequency based on its ExAC MAF: group 1 - novel, 2 - very rare (MAF <= 0.005), or 3 - rare (0.005 < MAF <= 0.01).

For each sample, the variant assigned to the lowest category was taken as the potentially causative variant. If there were two or more variants with the same category, the one in the lowest frequency group was selected. When there were two or more variants with the same category and frequency group, all were selected. Once a selection had been made, no other variants in that sample were considered. Category 1 variants were assumed to be of highest confidence, followed by category 2 and 3 variants and so selection was made in that order: If a suitable variant or variants were found in Category 1, no category 2 ones were considered, and similarly, if suitable variants were found in Category 2, no Category 3 ones were considered. No phase information was available for these data, so for non-homozygous variants where the inheritance model of the gene containing the selected variant required a second allele as part of a compound heterozygous pair, the next ranked variant in that gene was selected.

Thus, for each of the 106 samples, the output from the module was usually one (for dominant or homozygous recessive situations) or two (for compound heterozygous situations) potentially causative variants in a particular gene. Since each gene is associated with one or more of the 14 disease classes (shown in Table 2-1), identification of a gene implied one or in some cases two possible disease classes. For some samples, no potentially causative variants were found, or for compound heterozygous situations, only a single variant met selection criteria, and so no disease was identified.

Estimating probability for the disease: Table 2-2 lists the probability of pathogenicity assigned for each category of potentially causative variant. Category 1 variants (based on HGMD or ClinVar entries) were assigned a probability of 1.0, except for some missense variants where prediction methods suggested low impact. Category 2 missense variants were assigned a probability based on the extent of consensus among the four missense impact analysis methods used (SNPs3D, SIFT, PolyPhen-2, and CADD), utilizing a calibration from HGMD data and a control set of inter-species variants. That calibration shows a strong and approximately linear dependence of pathogenic probability on agreement between methods (Supp. Figure S1). Other variant types were subjectively assigned probabilities as shown in Table 2-2. For autosomal recessive situations, the combined probability of pathogenicity was taken as the product of probabilities for the two contributing variants. Those values were incremented by 0.2 for homozygous cases, as an *ad hoc* correction for increased confidence, and by 0.1 in compound heterozygous situations. Based on this scoring scheme, a probability of pathogenicity for a disease class was generated for all the samples in which one or more potentially causative variants were identified. For the cases in which a gene was associated with more than one disease class, equal probability was assigned for all the disease classes.

Table 2-2. Pathogenicity probability estimates for each variant type.

Variant Type	Probability Score
Reported in HGMD or ClinVar as pathogenic	1
Missense - Reported in HGMD or ClinVar as pathogenic and predicted damaging by only 2, 1 or 0 out of 4 methods	0.9
Missense – Predicted damaging by 4/4 methods	1
Missense – Predicted damaging by 3/4 methods	0.8
Missense – Predicted damaging by 2/4 methods	0.5
Missense – Predicted damaging by 1/4 methods	0.25
Missense - Not predicted damaging by any methods	0.15
Nonsense	1
Frameshift / Non-Frameshift Indel	1
Variant predicted to affect splicing	0.8
Variant close to Splice Donor site	0.2
Variant close to Splice Acceptor site	0.2
UTR Variant	0.05
Intronic Variant	0.05
All other variants	0

2.3.4 Post-challenge analysis

We performed many post-challenge analyses on the results in order to gain insight into the performance, strengths, and weaknesses of the method, and in doing so, made a number of observations. We assessed performance based on the official answer key provided by the Johns Hopkins DNA Diagnostic Laboratory group. For each patient, the key specified the disease class, the possibly causative variants (if any) found in the subset of the 83 genes examined, and a classification of each of these variants (pathogenic, likely pathogenic, VUS (variant of uncertain significance), likely benign and benign). The Hopkins classifications were based on the ACMG evidence rules (Richards et al., 2015a).

2.4 Results

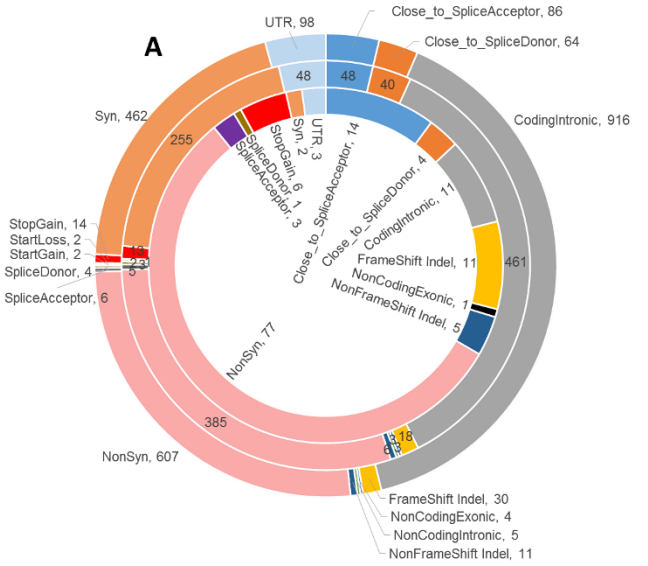
2.4.1 QC analysis summary

Supp. Figures S2 and S3 and Supp. Table S1 (in Appendix) together with accompanying text provide details of the QC analysis. Overall, transition/transversion ratios and heterozygosity/ homozygosity ratios are consistent with those found in 1000genome data, with the exception of one sample (P8) with excess homozygosity. There are a maximum of 2000 low quality and 940 no-call calls per sample in the v01 capture data and lower numbers in v02. We expect that any causative variants at these positions would be missed. Common, rare, and novel variant (SNV and Indel) counts for all the samples are consistent with 1000 genome data, except for two outlier samples identified as of African ethnicity which have larger rare Indel counts. The average read depth per gene per sample is high (greater than 100X) with the exception of two capture regions (Exon-53 and Exon-60 of *HYDIN* gene in Supp. Figure S4) where anomalous coverage could potentially result in causative variants being missed or in false positives.

2.4.2 Missense mutations are amplified in the potentially causative variant set

The VCF files provided for the challenge have a total 2311 unique variants across the 106 patients. This variant set consists of 40% intronic variants, 26% missense variants, 20% synonymous variants, and 14% of variants that are assigned as LoF (frameshift Indels, non-frameshift Indels, and nonsense), UTR, or splicing (Fig. 2-2A). After applying the PASS (PASS in VCF file), genotype (GQ > 30) and

frequency filters (MAF <= 1% in ExAC), the total number of variants was reduced by almost 50% to 1291, with 233 variants filtered because of low quality and 787 further variants filtered because of high MAF. Figure 2-2B shows that the frameshift and non-frameshift indels decreased the most (by 40% and 27%) on applying the PASS filter and NonSyn, Syn, UTR, CodingIntronic and 'Close to splice site' variants decreased the most (by 37 to 42%) on applying the frequency filter. After all filtering, 138 out of the 1291 variants were assigned as potentially causative by the prioritization procedure. In this set, the fraction of LoF variant is 16% and the fraction of missense variants is doubled to more than half (56%), while intronic variants drop to 8% and synonymous to 1%. The high fraction of potentially causative missense variants emphasizes the importance of correctly interpreting this class of mutation.



B

	Total variants	PASS variants	PASS and Rare variants	PASS, Rare and Potentially Causative variants
StartLoss	2	2	2	0
StopGain	14	14	13	6
NonCodingExonic	4	4	3	1
SpliceDonor	4	4	3	1
StartGain	2	2	1	0
Close_to_SpliceDonor	64	62	40	4
NonSyn	607	571	385	77
Syn	462	428	255	2
Close_to_SpliceAcceptor	86	76	48	14
CodingIntronic	916	799	461	11
UTR	98	82	48	3
SpliceAcceptor	6	5	5	3
NonFrameShift Indel	11	8	6	5
FrameShift Indel	30	18	18	11
NonCodingIntronic	5	3	3	0
Total	2311	2078	1291	138

Figure 2-2. Distribution of variant types for the gene panel sequencing data for 83 genes from 106 patients. Figure 2-2A: Distribution of variant types. The outer circle shows the distribution for all variants present in the VCF files provided as part of the

challenge. The middle circle shows the distribution of high-quality rare variants after applying PASS, GQ and frequency filters using data in the VCF file. The inner circle shows the distribution of potentially causative high quality rare and novel variants in 104 patients after applying the variant selection algorithm. Missense and loss of function variants are substantially enhanced in the latter set. Figure 2-2B: Changes in the variant type distribution during the filtering process, from total VCF variants, to those annotated as PASS, to those with low frequency, and finally those selected as potentially causative. The heat map indicates the percent decrease in variants on applying each filter (in the direction indicated by the arrows): the larger the decrease, the more orange; the smaller the decrease, the more green. The frameshift and non-frameshift Indel count decreased the most (by 40% and 27%) on applying the PASS filter and NonSyn, Syn, UTR, CodingIntronic, Close to splice site variants decreased the most (by 37 to 42%) on applying the frequency filter.

2.4.3 Matching individuals to disease class

Application of the categorization procedure described in Methods resulted in a non-zero probability for a specific disease class being assigned to 87 of the 106 patients. A further 17 patients were assigned a non-zero probability for two disease classes, as a consequence of a single gene being associated with two of the 14 disease classes. Two patients (P59 and P86) were not assigned to any disease class. P59 had the lowest average read depth for 50 genes out of 83 and next to lowest for the other 33 genes compared to other samples, suggesting that causative variants may have been missed.

2.4.4 Correct disease assignments also made by Hopkins

Overall, the assessors determined that we made correct disease assignments for 36 of 106 cases (Fig. 2-3A), in the sense that the highest probability was assigned to the disease class specified in the Hopkins answer key. The Hopkins group reported "pathogenic", "likely pathogenic, or "variant of uncertain significance" (VUS), based on ACMG variant classification, for 43 cases (Fig. 2-3B). The VarP pipeline assigned the maximum probability to the same disease class for 26 of these 43 cases, with the same variants assigned as causative. There are two primary reasons for our non-identification of the other 17 cases (row 6 and row 10 in Fig. 2-3A). First, for 10 of these patients, the Hopkins group found only one heterozygous variant in genes known to be associated with disease in a recessive inheritance pattern. Our method considered this insufficient evidence. Second, for the remaining seven patients we found an alternative disease that ranked higher in the variant categorization scheme. As noted in Methods, the selection scheme only considered the disease identified by the highest-ranked variants, and rejected all others. Had we considered diseases identified by lower confidence categorizations, five of these seven cases the Hopkins reported disease would have received 2nd ranking; one 3rd ranking; and one 4th ranking.

A

Row #	Prediction Performance	# of Patients
1	➢ Patients where VarP predicted potentially causative variants	104
2	○ Patients where VarP variant(s) and disease class matched Hopkins	**26**
3	○ Patients where VarP variant(s) did not match Hopkins	78
4	▪ Patients with different variants identified by VarP and Hopkins and the VarP disease did not match Hopkins diagnosis	16
5	▪ Patients where VarP found variant(s) but Hopkins did not and the VarP disease matched Hopkins diagnosis	**10**
6	▪ Patients where VarP found variant(s) but Hopkins did not and the VarP disease did not match Hopkins diagnosis	52
7	➢ Patients where VarP did not predict potentially causative variants	2
8	○ Patients where Hopkins found potentially causative variant	1
9	○ Patients where Hopkins not found pathogenic variant	1

B

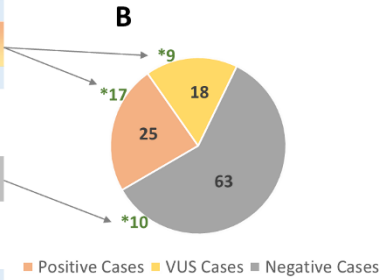


Figure 2-3. Disease assignment statistics for the 36 patients with correctly identified disease class. Figure 2-3A: Distribution of the number of patients across prediction performance. The highlighted numbers represent the patients with correct disease class assignment. The reasons for incorrect disease assignment are described in the text. Figure 2-3B: Distribution of Positive cases (orange) are those found by Hopkins to be carrying pathogenic or likely pathogenic variants, the VUS cases based on ACMG guidelines (yellow) are those carrying variants of uncertain significance and Negative cases (grey) are those in which no causative variant was found by Hopkins. * indicates the number of cases with correct disease class assignment. The VarP pipeline assigned the correct disease class for 26 of the Positive and VUS cases and also correctly assigned disease class (with potentially causative variants) in 10 of the Hopkins Negative cases.

2.4.5 Additional correct disease assignments

Out of the 63 patients for which the Hopkins analysis found no causative variants in the genes ordered as part of the clinical test, our method made 10 correct assignments of disease class and assigned potentially causative variants (row 5 in Fig. 2-3A). Seven of these patients were found to carry autosomal dominant or homozygous recessive variants and remaining three patients carried compound heterozygous variants. For nine of these 10 cases, the gene hosting the potentially causative variant was not analyzed by the Hopkins group, presumably because coverage was not selected by the requesting physician. For the remaining case, the Hopkins group did not report the potentially causative variant even though they analyzed the relevant gene. For the other 53 patients (row 6 and row 9 in Fig. 2-3A), neither our method nor the Hopkins group found any causative variants for the expected disease class. However, we found potentially causative variants for a different disease in four of these patients, suggesting alternative diagnoses (see the Alternative Diagnosis section).

2.4.6 Assignment of probability

In order to estimate the accuracy of our probability model, we checked how well the probability of pathogenicity scores correlated with incorrect disease class assignment. The dependence of incorrect disease assignment on assigned probability follows the correct trend, with a high fraction at low probability and a lower fraction at high probability (Fig. 2-4). However, there are 25 patients with incorrect disease class assignments and a probability greater than 0.8. We found the following reasons for

this: 1. High confidence given to DM status HGMD variants – 11 of the 25 anomalies are of this type. These are discussed below in the Selection section and listed in Table 2-4. 2. In five cases, there were pairs of Indels (frameshift or non-frameshift) close together (less than 10 bp apart, Supp. Table S2) in the CCDC40 gene and classified by us as causative compound heterozygous variants. Very likely, these are false Indels arising from alignment errors or errors near perfect repeat regions (Fang et al., 2014). 3. In two cases there are two heterozygous variants predicted damaging by three methods in genes associated with recessive disease. It is possible that these are the same copy of the gene (no phasing information was available). 4. In the remaining seven cases, we found possible alternative diagnoses. These are discussed in detail below.

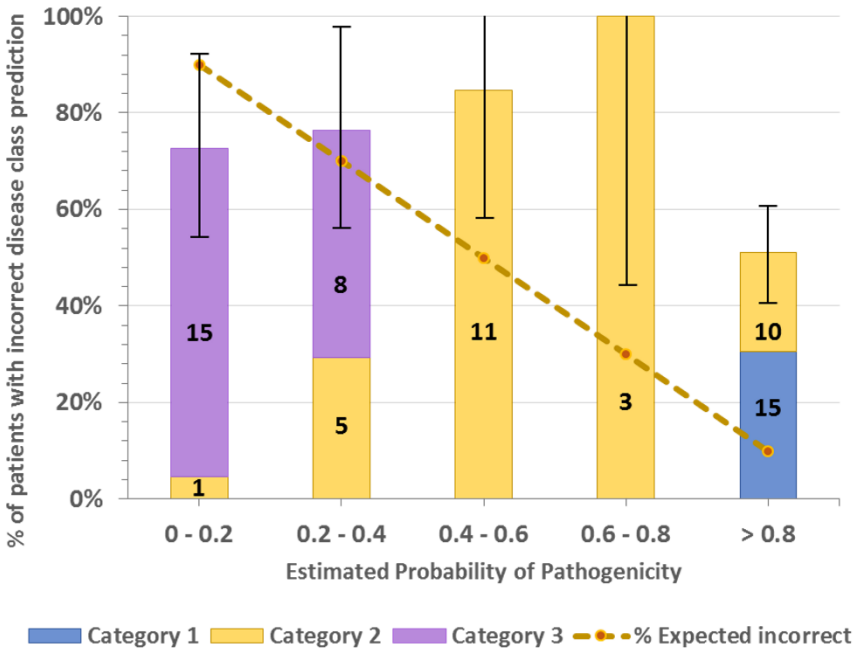


Figure 2-4. Distribution of patients with incorrectly assigned disease class versus estimated probability of pathogenicity. The dotted line shows the expected value in

each bin (e.g. in the 0.8 to 1.0 bin, 10% of disease assignments are expected to be incorrect). Bars show the % of patients in each bin that actually have incorrect assignments. Bar colors show the number of patients with assignments made in each category (Category 1, most confidence). The error bar for each bin is the standard deviation of the number of patients in that bin. As should be the case for a good probability algorithm, patients with a high probability of a correct disease assignment do have a lower rate of incorrect disease classes. However, the plot also shows that there are 25 patients with high probability scores (> 0.8) but incorrect disease class. 15 of these patients carry variant(s) reported as pathogenic (tagged as DM) in the HGMD database. Reasons for this are discussed in the text.

2.4.7 Variant assignment accuracy for each Selection Category

As described in Methods, we used a work flow to assign variants to one of three categories, ranked by likelihood of pathogenicity. Table 2-3 shows the percent of correct disease assignments for variants in each category. The highest fraction of cases (42%, 11 out of 26) agreeing with the Hopkins disease class were based on Category 1 variants. The corresponding fractions for Category 2 and Category 3 variants are 38% and 23% respectively. This trend is expected, since assignment confidence decreases with increasing category number.

Table 2-3. Percentage of correct disease assignments in each of the three variant selection categories. As expected, accuracy is highest in Category-1, then Category-2, then Category-3. Novel variant assignments are more accurate than for rare variants.

Category	Variant Considered	Minor Allele Frequency			% Correct Assignment
		Novel	<= 0.005	<=0.01	
Category-1	In HGMD with DM, DP status and/or in ClinVar with Pathogenic or Likely pathogenic tag	4/4	7/19	0/3	11/26: 42%
Category-2	Missense (Predicted damaging either by SNPs3D, SIFT, PolyPhen2 or CADD) Frameshift / Non-Frameshift Indel NonSense Direct Splicing Any variant predicted damaging by dbscSNVs	9/14	7/28	2/5	18/47: 38%
Category-3	All other missense, UTR, and Intronic	4/17	2/12	1/2	7/31: 23%
		17/35: 49%	16/59: 27%	3/10: 30%	

As noted earlier, Category 1 variants are those annotated in HGMD and/or ClinVar as disease-causing. Further inspection showed that 11 of the 15 discordant assignments cases had conflicting database annotations and sometimes weak or no supporting evidence (Table 2-4). In seven cases, the corresponding variant is annotated ’DM’ (disease mutation) in HGMD but is annotated ‘benign’ or ‘likely benign’ in ClinVar. Consistent with the ClinVar annotation, a check of the supporting literature for these showed either no experimental support or no evidence favoring pathogenicity. For the other four cases, ClinVar had no relevant entry and there was no literature support. Seven of these 11 cases involved missense variants, and none of those were rated

high confidence pathogenic by our consensus method. With the wisdom of hindsight, we should have factored these considerations into the categorization and probability procedures, and placed less faith in HGMD. The remaining four discordant assignments have either functional validation of the variant as damaging or are annotated as pathogenic in ClinVar as well. As discussed later, these four patients may really have a different disease.

Category 2 variants are those selected because of being a LoF variant, the computational method assigning pathogenicity for missense variants, a direct hit on a splice site, or a prediction of an impact on splicing (Jian et al., 2014). The 18 correct assignment cases include seven compound heterozygous and 11 autosomal dominant or recessive cases. Seven out of these 11 cases carry a LoF (nonsense, frameshift or non-frameshift Indel) or direct splicing variant, and the remaining four carry missense variants (two predicted damaging by two methods and two predicted damaging by one method). The 29 cases with discordant disease class with respect to the Hopkins information in this category include 11 compound heterozygous cases and 18 autosomal dominant or recessive cases. For the 11 discordant compound heterozygous cases, the assumption that the two variants are appropriately phased is a likely cause of misassignment. The 18 other cases include one frameshift Indel and 17 missense variants. Of the 17 missense, only one was high confidence, predicted damaging by all four methods. Four were damaging by three methods (expected accuracy 0.8), 10 were damaging by two methods (expected accuracy 0.5), and two were damaging by only one method (expected accuracy 0.25).

Category 3 variants are missense mutations predicted benign by all four computational methods and those which are intronic or in a UTR. All were assigned low causative probability, ranging from 0.05 to 0.29. There are only seven out of 30 with correct disease class assignments that were assigned based on Category 3 potentially causative variants. Six of these seven cases carried intronic insertions or deletions close to a splice site (within 5 to 30 bases), suggesting proper treatment of this mechanism is important. The remaining case carries a missense mutation predicted benign by the four mutation impact prediction methods.

There is a marked dependence of level of agreement with the Hopkins disease class and the frequency of the potentially causative variants (Table 2-3): 49% of disease assignments made for novel variants agree with the Hopkins answer key, compared to 27-30% for the other, non-novel variants with less than 1% MAF.

Table 2-4. List of variants reported in HGMD with DM or DP status but not supported by other data and leading to an incorrect diagnosis. MAF: Minor Allele Frequency. These variants were present in 11 patients. Green: benign missense prediction, red: deleterious prediction.

Variant Chr:Pos:Ref:Alt	Potentially causative variants in # of Patients	MAF in ExAC	Mutation Type	Gene	cDNA Change / Amino Acid	Impact Predictions B=Benign, D=Damaging, PD=Possibly Damaging	ClinVar Clinical Significance	HGMD Status	HGMD Reported PMID	Comments
15:48748913:C:T	2	0.0048	Silent	FBN1	NM_000138.4:c.5343G>A (p.(V1781=))	CADD=17.74, D	Benign	DM	17627385	Reported in a table. No functional study reported.
15:48725102:C:T	1	0.0008	Missense	FBN1	NM_000138.4:c.6700G>A (p.(V2234M))	SIFT=0.218, B PolyPhen2=0.121, B CADD=16.35, D SNPs3D=0.8774, B	Other	DM	17253931	
3:30733044:T:A	1	0.0014	Missense	TGFBR2	NM_003242.5:c.1657T>A (p.(S553T))	PolyPhen2=0.942, D CADD=23.6, D	Likely Benign	DM	16791849	Reported in a Table. Has a normal transcript, no other data.
5:149740732:C:T	2	0.0023	Missense	TCOF1	NM_001135243.1:c.122C>T (p.(A41V))	SIFT=0.006, D PolyPhen2=0.139, B CADD=23, D SNPs3D=0.999, B	-	DM	12444270	Unknown Significance. In PMID: 19572402 this variant is reported benign.
7:117305631:A:G	1	0.0034	Intronic	CFTR	NM_000492.3:c.4242+13A>G	CADD=5.561, B	Benign	DM	15858154	Variant is not at all reported in the paper.
15:48722907:G:A	2	0.0033	Missense	FBN1	NM_000138.4:c.6832C>T (p.(P2278S))	SIFT=0.035, D PolyPhen2=0.59, PD CADD=25.6, D SNPs3D=0.72, B	Benign	DM	19293843	Reported as part of a double mutant in the paper. No functional study reported.
15:48818329:A:G	1	0.0023	Missense	FBN1	NM_000138.4:c.986T>C (p.(I329T))	PolyPhen2=0.015, B CADD=23.6, D	Likely Benign	DM		
16:23391725:C:T	1	0.0067	Intronic	SCNN1B	NM_000336.2:c.1543-17C>T	CADD=7.765, B	-	DP	15661075	Although close (17bp) to a splice site a study showed no splicing effect.

2.4.8 Alternative Diagnoses

There is an important difference between the Hopkins lab procedures and the CAGI challenge. In the lab, in accordance with clinical guidelines, for each patient, variant analysis was performed only on the subset of genes identified by the physician requesting the test, usually those for a single disease, and sometimes only a subset of genes for a single disease. On the other hand, the challenge required analysis of all genes for each patient. That led to a number of findings suggesting that in some cases, causative variants are overlooked in the clinic. Of the 70 cases where our disease assignments and the disease tested by the Hopkins pipeline differ, seven have strong evidence supporting assignment to a different disease (Table 2-5). In four of these cases, no variants supportive of the tested disease were found by ourselves or by Hopkins. In two further cases, the Hopkins pipeline reported only one variant in a recessive gene and for the remaining case (patient P8 in Table 2-5), there is evidence that the patient may have two diseases. These seven cases fall into three groups:

1. Three cases where the patient carried variants likely causative of a disease phenotype that has overlapping symptoms with the disease tested at Hopkins. One of these is a patient (P36) carrying a very rare (AF=0.0047 in ExAC) autosomal dominant missense mutation (rs5738:G>A, NM_001039.3:c.589G>A, p.(E197K)) in exon-3 of the *SCNN1G* gene. This mutation is reported in HGMD and ClinVar to be causative for Bronchiectasis with pathogenic clinical significance (Fajac et al., 2008). The patient was tested for Diffuse Lung disease and no variants with the required inheritance pattern was found in the relevant genes. Bronchiectasis has been

previously shown to be associated with Idiopathic Pulmonary Fibrosis, one of the diseases in the Diffuse Lung disease class. (International Consensus Statement of the American Thoracic Society and the European Respiratory Society. 2000; Bourke 2006).

2. One case where a patient (P8) carried a variant reported in HGMD and ClinVar to have pathogenic clinical significance for a disease other than that tested for, and where the tested and apparent diseases cannot be easily confused. P8 carries a very rare (AF=0.0051 in ExAC) homozygous recessive mutation (rs1800098:G>C, NM_000492.3:c.1727G>C, p.(G576A)) in the *CFTR* gene, consistent with the disease class 'Cystic Fibrosis and CF-related disorders'. A functional study found the mutation causes an increased amount of skipping of exon-12 during splicing (Pagani et al., 2003). This patient was originally tested for Peroxisomal Beta-Oxidation Defects and a homozygous recessive frameshift mutation was found in the relevant gene. We did not report that variant because of finding the *CFTR* variant which we categorize as higher confidence of pathogenicity. The data are consistent with the patient having both diseases.

3. Three cases where the patient carried variant(s) predicted damaging by all reporting computational methods or a LoF variant. For example, one of these is a patient (P46) to whom we assigned 'Treacher Collins and Related Syndromes' based on a very rare (MAF = 0.0002 in ExAC) missense mutation (rs538401137:C>T, NM_001135243.1:c.3029C>T, p.(T1010I)) in the *TCOF1* gene and assigned as

damaging by all four computational methods. This patient was tested for the Diffuse Lung disease class in the Hopkins pipeline, and no variants consistent with that phenotype were found by them or us.

Table 2-5. Patients carrying putative causative variants for an alternative disease. AD=Autosomal Dominant, AR=Autosomal Recessive, CH=Compound Heterozygous (for AR cases, the listed variant is homozygous). The table is divided into three case types. Green: benign missense prediction, red: deleterious prediction.

Patient ID	Variant Chr:Pos:Ref:Alt	MAF in ExAC	Inheritance Model	Mutation Type	Gene	Amino Acid / cDNA Change	Impact Predictions B=Benign, PD=Possibly Damaging, D=Damaging	ClinVar Clinical Significance	HGMD Status, PMID	Predicted Disease Class	Tested Disease class
# TYPE-1: Patients carrying variants likely causative of a disease phenotype that has overlapping symptoms with the disease tested at Hopkins.											
P36	16:23200963:G:A	0.0047	AD	Missense	SCNN1G	NM_001039.3:c.589G>A (p.(E197K))	PolyPhen2=0.003, B CADD=13.76, D SNPs3D=0.7541, B	Pathogenic	DM, 18507830	Bronchiectasis	Diffuse Lung Disease
P48	16:23200963:G:A	0.0047	AD	Missense	SCNN1G	NM_001039.3:c.589G>A (p.(E197K))	PolyPhen2=0.003, B CADD=13.76, D SNPs3D=0.7541, B	Pathogenic	DM, 18507830	Bronchiectasis	Pseudohypoaldosteronism type 1
P7	5:1293767:G:A	0.0022	AD	Missense	TERT	NM_001193376.1:c.1234C>T (p.(H412Y))	SIFT=0.046, D PolyPhen2=0.897, PD CADD=4.908, B SNPs3D=1.915, B	Pathogenic	DM, 15814878	Pulmonary Fibrosis and/or Bone marrow failure	Cystic Fibrosis and CF-Related
# TYPE-2: Patient carrying a variant reported in HGMD and ClinVar with pathogenic clinical significance, for a disease other than that tested at Hopkins.											
P8	7:117230454:G:C	0.0051	AR	Missense	CFTR	NM_000492.3:c.1727G>C (p.(G576A))	SIFT=0.258, B PolyPhen2=0.697, PD CADD=13.18, B SNPs3D= -0.125, D	Pathogenic	DM, 1545465	Cystic Fibrosis and CF-related disorders	Peroxisomal Beta-Oxidation Defects
# TYPE-3: Patient carrying LoF variant or missense predicted damaging by all reporting computational methods.											
P40	10:13337608:T:C	0.0001	CH	Splice Acceptor	PHYH	NM_006214.3:c.135-2A>G	CADD=24.7, D	Pathogenic	DM, 9326939	Zellweger Spectrum Disorders	Cystic Fibrosis and CF-Related
	10:13336522:GA:G	Novel		Frame-shift		NM_006214.3:c.319del (p.(S107Rfs*11))					
P46	5:149767634:C:T	0.0001	AD	Missense	TCOF1	NM_001135243.1:c.3029C>T (p.(T1010I))	SIFT=0.0, D PolyPhen2=0.993, D CADD=26.2, D SNPs3D= -1.284, D			Treacher Collins and Related Syndrome	Diffuse Lung Disease
P75	17:78032436:G:A	0.0018	CH	Missense	CCDC40	NM_001243342.1:c.1303G>A (p.(E435K))	SIFT=0.008, D PolyPhen2=0.99, D CADD=28.4, D SNPs3D= -0.328, D			Primary Ciliary Dyskinesia	Diffuse Lung Disease
	17:78064014:A:ACAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCAC	0.0007		Frame-shift		NM_001243342.1:c.c.2909_2910insCAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCAC (p.(K970Nfs*144))					

2.4.9 Protein structure coverage for potentially causative variants

In principle, information from three-dimensional structure and on the detailed functional roles of residues, motifs, and domains should be of considerable value in evaluating the impact of missense variants. In practice, it is often ignored, and indeed we did not use it in this challenge. What difference might it have made? To investigate this, we considered only potentially causative missense variants that are not included in HGMD or ClinVar. Current ACMG guidelines (Richards et al., 2015a) would place a low weight on computational analysis of these, and thus they would likely be reported as VUSs. There are 47 such missense variants distributed over 41 patients. ~50% (23/47) of these are either included in an experimental structure or can be included in a homology model based on 22% or higher sequence identity to an experimental structure (Fig. 2-5A). Three of these mutations are in proteins with experimental structure (X-ray structure). We use these three cases to illustrate how protein structure could be used to: (a) supplement the sequence analysis methods to increase confidence in a pathogenic or benign assignment and (b) understand the pathogenicity mechanism at the protein level. Two of these mutations have correctly assigned disease classes and causative variants in our submission. One of those is of a novel mutation (NM_000901.4:c.1807T>A, p.(C603S)) at a highly conserved position in the Mineralocorticoid receptor. This protein is associated with Pseudohypoaldosteronism Type 1. Although we correctly identified this mutation from sequence information, only two of the four (SNPs3D, SIFT, PolyPhen2, CADD) methods assigned it as pathogenic, and the other two did not. Thus additional evidence would have improved confidence in the assignment. Inspection of the

structure (PDBID: 4TNT) showed that the wild-type amino acid (CYS-603) is a zinc ligand in a zinc finger domain (Fig. 2-5B). Many other zinc ligand mutations in these domains cause loss of function of the corresponding proteins (Kambouris et al., 2014; Vincent et al., 2014), providing additional evidence of pathogenicity. The second case with correct disease assignment is of another novel mutation (NM_000492.3:c.3849G>C, p.(R1283S)) at a highly conserved position in the second nucleotide binding domain of the *CFTR* protein. Mutations in *CFTR* cause Cystic Fibrosis, one of the disease classes in the Hopkins dataset. This mutation is predicted damaging by three out of four (SNPs3D, SIFT, PolyPhen2, CADD) sequence methods. Inspection of the protein structure (PDBID: 3GD7) hosting this mutation shows the wild type side chain (R1283) makes two charge-dipole interactions with main chain carbonyl groups of L1258 and R1259, providing a helix cap (Hol et al., 1981), consistent with significant destabilization of the structure (Fig. 2-5C). Loss of protein stability has been shown to be the most common cause of monogenic disease (Wang and Moult 2001; Yue et al. 2005). A different mutation at this position (rs77902683, NM_000492.3:c.3848G>T, p.(R1283M)) has previously been found in CF patients (Cheadle et al., 1992) and has been reported as pathogenic in ClinVar and HGMD.

The third mutation with experimental structure coverage is one where we made an incorrect disease assignment on the basis of just one of the four missense analysis methods predicting deleterious. Although that was already a low confidence prediction, further evidence would be useful. This is a very rare (MAF=0.0005 in

ExAC) variant (rs147398624:G>A, NM_000901.4: c.2578G>A, p.(V860I)) in the Mineralocorticoid receptor, with an autosomal dominant pattern disease inheritance pattern. The mutation is located on the protein surface (PDBID: 2AA5) and is not part of any known interface, providing further evidence the mutation is benign.

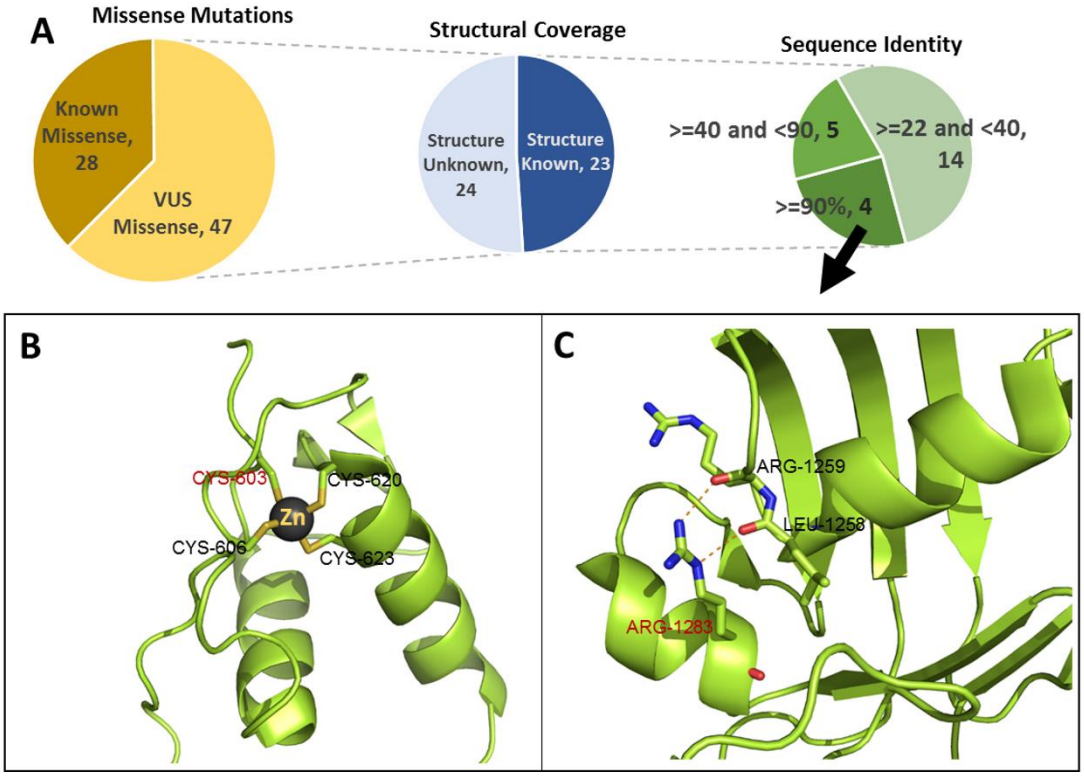


Figure 2-5. Structural coverage of prioritized missense mutations. Figure 2-5A: Missense variant distribution: 1) Known (variant reported in HGMD or ClinVar) versus VUS variants, 2) Structural coverage for VUS variants, 3) Number of mutations in different sequence identity ranges between the protein hosting the mutation and the closest available homologous protein in the PDB. Figure 2-5B and 2-5C show two examples of structure assisting mutation interpretation. Figure 2-5B: Part of a zinc finger domain of the Mineralocorticoid receptor protein (PDB: 4TNT)

including the mutation C603S found in a Hopkins patient, showing that C603 is one of the Zn ligands. Analogy to other zinc coordinating mutations in zinc fingers provides strong evidence structure and hence function will be disrupted. Figure 2-5C: Mutation R1283S, found in one of the Hopkins patients, is predicted deleterious by the three out of four computational methods. Inspection of the structure shows disruption of two charge-dipole interactions forming a helix cap, expected to significantly destabilize the structure.

2.5 Discussion

The CAGI4 challenge based on panel sequencing data provided by the Johns Hopkins DNA diagnostic laboratory has allowed a blind test of current methods for identifying causative variants in clinical rare disease sequence data. Participants were asked to match each of 106 patients to one of 14 classes of disease. To address this challenge, we developed an analysis pipeline, VarP, designed to identify potentially causative variants. Using this pipeline, we were able to correctly match 36 patients to the reported disease class. The analysis provided a number of insights into issues related to gene panel testing, including the relationship between data quality and success in finding causative variants, variant prioritization procedure limitations, inconsistencies in databases, and cases of possible alternate diagnosis.

2.5.1 Undiagnosed cases

Even with full knowledge of the reported disease class, the Hopkins pipeline could only find potential causative variants for 43 cases, leaving 63 with no causative

variants. As discussed below, we were able to find variants correctly matching a further 10, but that still leaves half (53) of the cases where neither we nor Hopkins could find variants. There are three major factors that may contribute to the high fraction of undiagnosed cases. First, a limitation in all studies of this type is data quality. Our QC analysis suggests the Hopkins data are generally of high quality. Read depth per gene per sample is high (between 107X to 983X) and each sample has only about 2000 positions with no call or a low-quality variant call. But there are some particular sample level properties in the data that may affect analysis. For example, sample P8 (tested for an autosomal recessive disease, Peroxisomal Beta-Oxidation defect) has an abnormally high fraction of homozygous variant calls compared with heterozygous ones, increasing the chances of finding an apparently causative homozygous variant. Our pipeline identified a potentially causative homozygous missense variant in *CFTR*, consistent with cystic fibrosis, and annotated as pathogenic in both HGMD and ClinVar, whereas the Hopkins pipeline found a homozygous frameshift variant in HSD17B4, consistent with the tested disease. There are also some areas of low coverage, for instance 78 samples have zero coverage of Exon-60 in *HYDIN*. Variants in this gene may cause Primary ciliary dyskinesia. Overall though, sequencing data quality does not appear to make a large contribution to missing diagnostic variants.

A second factor contributing to non-identification of causative variants is that there may be other, unknown, genes where variants cause the disease phenotype. Many new monogenic disease genes are still being discovered (more than 67 genes in a two-

year period, Beaulieu et al. 2014). Thirdly, the causative variants may have been not covered in the panel, which consists of mostly exon sequence. Missing variants may include those affecting the expression of a relevant gene, CNVs, and larger scale structural genomic changes. In some rare disease analyses using whole genome sequence (WGS), such as in the SickKids Genome Clinic (http://www.sickkids.ca/CGM/genome-clinic/index.html), the latter type of variant has been found to make a significant contribution (Stavropoulos et al., 2016). However, those patients mostly exhibit major developmental disease phenotypes, and may not be typical of rare disease patients in general.

2.5.2 Correct diagnosis for cases where Hopkins pipeline did not find causative variants

For 10 cases we were able to identify the reported disease class even though Hopkins reported no potentially causative variants. In nine out of these ten cases, the Hopkins pipeline did not include analysis of the gene carrying the diagnostic variant(s). Apparently this is because the requested test did not include the gene, a choice made by the referring physician. As noted earlier, Hopkins is only permitted to analyze the requested gene set. For the 10th (a compound heterozygous case where one of the variants is missense predicted damaging by four methods and other is an intronic variant close to a splice acceptor), Hopkins did not report the potentially causative variants even though they analyzed the relevant gene.

2.5.3 Missed diagnoses

There are 17 cases where we did not identify the correct disease class, but the Hopkins analysis did find potentially causative variants. For 10 of these, the Hopkins variants are in genes expected to have a recessive inheritance pattern, and only one heterozygous variant was present – not sufficient for our evidence rule. Had we used such a weak criterion for inheritance model filtering many more false positives would have been generated. Thus these should not be regarded as failure of the VarP approach but rather an appropriate filtering strategy used in VarP. In the other seven cases where Hopkins found variants, VarP found stronger evidence for a different disease class. For two of these, as discussed below, we consider the evidence that the patients have the VarP identified disease very strong, and if so, these also are not errors. For the other five, we made two sorts of errors. One was placing too much trust in HGMD that affected three cases – in each of these cases the HGMD annotations were incorrect and contradicted or not supported by ClinVar or experimental data. The other source of error was for two compound heterozygous cases where one of the partner variants was a low impact missense (predicted benign by 1/4 methods) or an intronic variant and so provided very weak evidence. In retrospect, the procedure of taking just the most likely causative variant(s) and ignoring all other variants in a patient was sub-optimal. A better procedure would probably be to use all variants in each gene to assign a probability of pathogenicity and to use those probabilities to infer disease class.

2.5.4 Incorrect diagnoses

For 25 patients VarP made high confidence (probability score > 0.8) incorrect disease class assignments. A primary factor was again over-reliance on HGMD annotation, accounting of 11 of the 25 cases. A further five cases involved pairs of Indels very close to each other (less than 10 base pairs apart), and consistent with a compound heterozygous cause for a recessive disease. In fact, these Indel pairs are probably coupled alignment errors. There are two cases where the assumption that a pair of recessive variants are on different copies of the gene may be incorrect (there was no phasing data available). In seven of the remaining cases, we found high confidence pathogenic variants in genes associated with a different disease from that in the Hopkins answer key. As discussed later, the evidence for some of these is sufficiently strong that they may not be errors.

2.5.5 Distinct potentially causative variants that led to disease classification

VarP identified 105 potentially causative variants each of which occurs once in a total of 78 patients. A further 14 potentially causative variants were seen in two or more of the other 28 patients (Supp. Table S3). We also considered accuracy in terms of the fraction of these 119 distinct variants which led to correct and incorrect disease assignments. By this measure, correct disease identification increases from 34% (36/106) to 36% (33/91). The improvement occurs because the majority of repeat variants are present in cases where an incorrect disease was assigned, and we speculate that some of these may reflect sequencing artifacts.

2.5.6 Reliability of probability for disease assignments

In the clinic, perhaps more important than having an accurate method of determining pathogenicity is having an accurate method for assigning a probability of correctness to a pathogenic assignment. The CAGI challenge required participants to also provide these probabilities, and so it was possible to evaluate how effective our approach was. We used a largely *ad hoc* probability scale in this analysis. Although there is a reasonable overall correlation between these quantities (Fig. 2-4), there were a substantial number of variants assigned a high probability that were not in fact pathogenic. There were two primary reasons for that – first, as noted earlier, we misjudged the reliability of HGMD assignments of disease mutations. Had we used a model that included disagreements between HGMD and ClinVar, these cases would have had more appropriate probabilities. Second, as discussed below, in a number of cases we consider the evidence strong that these patients had a different disease.

2.5.7 Reliability of missense probability estimates

As described in the Results, overall, the estimated probabilities of pathogenicity shows qualitative though not quantitatively correct properties. The majority of potentially causative variants are missense, so improved confidence in assigning a probability of pathogenicity to these are of particular importance. As described earlier, we assigned a probability based on the fraction of four different missense analysis methods reporting deleterious. The method was calibrated (Yin et al., 2017b) using a set of HGMD mutations (all assumed pathogenic) and a set of interspecies variants (assumed benign). There are a number of limitations to this dataset, and so

we were interested to see to what extent the estimated probabilities were useful. Interpretation of the results is complicated by the alternative diagnosis cases and by compound heterozygous cases, involving two different variants. Supp. Figure S5 shows the relationship between estimated probabilities and correct disease class assignment, omitting those cases. Counts here are too small to draw firm conclusions. A high proportion of mutations assigned with a probability of less than or equal to 0.5 are incorrect, consistent with expectations. However, more than half of the mutations with probabilities higher than 0.7 are also incorrect, not as expected. Further analysis Yin *et al.* (ref to Yin *et al.* CAGI issue paper when available) suggests that a probability method based on more than four missense impact prediction methods would have yielded better results. But clearly a more extensive blind test is needed to evaluate this approach.

2.5.8 Apparent cases of alternative diagnoses

Using quite stringent criteria we identified seven cases where the data are consistent with patients having a different disease class than that provided in the Hopkins answer key. Four of these patients carry variants for the alternative disease class that are reported in HGMD and ClinVar as pathogenic. The remaining cases carry missense variants predicted damaging by all reporting methods, frameshift or non-frameshift indels, or variants directly affecting splicing. In three cases, symptoms of the answer key disease and the alternative overlap, so it is possible that there was a misdiagnosis in the referring clinic. The other cases are more puzzling. Since we have no information as to why a particular test was ordered (and in many cases the

Hopkins group may not either), it is difficult to comment further. But it is concerning that in a number of cases there could be confusion of some sort as to what disease patients have. In these seven cases, the Hopkins pipeline did not report any variant for four cases, reported only one variant in a recessive gene for two cases and reported a homozygous frameshift mutation in the remaining case. The pipeline was prevented from discovering the possible alternatives by the current guidelines, which require that only requested genes for a specific disease test be examined. On the basis of these limited data, it is not clear whether on balance this practice is in the patients' best interest.

2.5.9 VarP performance improves when the patients' clinical indications are known

Clinical laboratories typically have information on each patient's disease phenotype, and variants are evaluated with that knowledge. In that aspect, the CAGI Hopkins challenge creates an artificially harder problem, since disease class is not known to participants. If the disease classes were known, would VarP identify the variant(s) reported by Hopkins pipeline? We tested this scenario by searching for potentially causative variant(s) only in genes associated with each patient's diagnosed disease class, using the VarP pipeline. On this basis, VarP identifies potentially causative variants for 61 patients, 18 more cases than the Hopkins pipeline. However, there are still nine cases where Hopkins identified potentially causative variants and VarP does not. As discussed earlier, these patients each carry only one heterozygous variant in a recessive gene, which we considered insufficient evidence.

2.5.10 Better results have been obtained not using HGMD

As noted earlier, 11 of the 25 incorrect disease class assignment cases with a probability of pathogenicity higher than 0.8 are a result of accepting HGMD annotations of pathogenicity. Such a high error rate from a single cause suggests that it might be better to ignore HGMD altogether and just use ClinVar for pathogenicity information. We tested this by running the VarP pipeline again, omitting HGMD. The success rate (correct match to disease class) increases from 36 to 40 (Supp. Table S4).

2.5.11 Lessons learned

Going forward, how would we now improve performance of the VarP analysis pipeline? As noted earlier, a suboptimal feature of the procedure was terminating the variant search once a suitable candidate had been found, rather than finding all possible causative variants and assigning each a probability. As also noted earlier, over-reliance on HGMD was a cause of errors and this can be corrected by considering ClinVar and HGMD annotations together, and, where appropriate, include missense impact analysis in assigning a probability to these Category 1 variants. Structure also has the potential for contributing to the discovery of causative variants and providing mechanistic insight. However, full automation of that analysis will require the development of new methods. In general, much more work must be done to provide a reliable probability of pathogenicity, not only for missense but for all types of variants.

2.6 Acknowledgements

We are grateful to Drs Bethany Buckley, Molly Sheridan, and Garry R. Cutting, The Johns Hopkins University for making the challenge data available. This work was supported in part by NIH R01GM104436 to JM. The CAGI experiment coordination is supported by NIH U41 HG007446 and the CAGI conference by NIH R13 HG006650.

Chapter 3: MecCog: A knowledge representation framework for genetic disease mechanism

3.1 Abstract

Experimental findings on genetic disease mechanisms are scattered throughout the literature and represented in many ways, including unstructured text, cartoons, pathway diagrams, and network graphs. Integration and structuring of such mechanistic information will greatly enhance its utility. MecCog is a graphical framework for building integrated representations (mechanism schemas) of mechanisms by which a genetic variant causes a disease phenotype. A MecCog mechanism schema displays the propagation of system perturbations across stages of biological organization, using graphical notations to symbolize perturbed entities and activities, hyperlinked evidence tagging, a mechanism ontology, and depiction of knowledge gaps, ambiguities, and uncertainties. The web platform enables a user to construct, store, publish, browse, query, and comment on schemas. MecCog facilitates the identification of potential biomarkers, therapeutic intervention sites, and critical future experiments.

3.2 Introduction

Findings from experimental studies of disease mechanism are reported across multiple publications in varying combinations of structured and unstructured data and many different diagrammatic representations. A number of projects have addressed different aspects of the resulting knowledge integration problem. These include

building disease-specific knowledge management resources (for example alzforum.org (Kinoshita & Clark, 2007)) and ontologies (ADO (Malhotra et al., 2014), PDON (Younesi et al., 2015), CVDO https://bioportal.bioontology.org/ontologies/CVDO); compiling disease etiology databases (HGMD (Stenson et al., 2017), ClinVar (Landrum et al., 2018), CIVIC (Griffith et al., 2017), PanelApp (Martin et al., 2019)); development of biomedical text mining methods (DARPA's Big Mechanism program (Cohen, 2015)); development of statistical methods for evidence integration and assessment (Konopka & Smedley, 2020); and community-driven expert systems medicine disease maps projects (Mazein et al., 2018). Each of these contributes elements of a solution, but a major omission is an integrated representation of mechanism knowledge in a clear, precise, and comprehensive manner.

There have also been major technological advances in the development of tools to support mechanism descriptions, such as graphical notations (SGBN (Systems Biology Graphical Notation) (Novère et al., 2009b)) and languages (SBML (Systems Biology Markup Language) (Hucka et al., 2018), KGML (KEGG Markup language - https://www.genome.jp/kegg/xml/), BCML (Biological Connection Markup Language) (Beltrame et al., 2011), BioPAX (http://www.biopax.org/), BEL (Biological Expression Language - https://bel.bio/)) to encode representations; software to draw and visualize models (GO-CAM (Thomas et al., 2019), PathWhiz (Pon et al., 2015), Cytoscape (Paul Shannon et al., 2003)); linked data formats such as Nanopublications (Mina et al., 2015) to organize provenance and metadata for

scientific assertions; and databases to store and query graph-based representations (Neo4j - https://neo4j.com/, (Daniel Scott Himmelstein et al., 2017)).

With the help of these tools; pathway, network, and disease map representation types have been created to describe aspects of biological system mechanism and in some cases disease mechanisms as well. For instance, KEGG (Minoru Kanehisa et al., 2016), and Reactome (Fabregat et al., 2017) pathways represent normal and perturbed molecular interactions that are part of cellular or metabolic processes. STRING (Szklarczyk et al., 2018) and GeneMANIA (Franz et al., 2018) networks represent integrated information on protein-protein interactions and associations that are part of normally functioning biological systems. Gene ontology (GO) causal activity models (Thomas et al., 2019) integrate GO annotations to generate larger models of normal biological function (such as 'pathways') in a semantically structured manner. The Disease Maps Project (Mazein et al., 2018) provides an encyclopedic description of disease-related signaling, metabolic, and gene regulatory processes. Although together these representations aptly describe the normal working of biological systems, representation of the disease-related perturbations is limited. In the existing representations (such as KEGG or Reactome disease pathways), disease state perturbations and consequences are added locally to the depiction of the normal state of the biological system. Adding disease perturbation information to already complex pathway diagrams can be useful, but limits clarity. Also, uncertainties, ambiguities, and ignorance in mechanistic knowledge are not presented in most representations (with the exception of Reactome pathway diagrams, but these label uncertain reaction

types only). Such knowledge gaps exist in almost all disease mechanisms (Greenberg & Amato, 2004; Kametani & Hasegawa, 2018).

These considerations led us to propose a graphical framework with an integrated representation of genetic disease mechanisms from gene to phenotype. Our design goals were that the representation framework depict mechanism components across stages of biological organization; display perturbation propagation; make use of standard biomedical ontology terms wherever possible to name the components; provide an intuitive way to visualize ignorance, uncertainties, and ambiguities; and allow tight linkage to evidence in the literature and databases. The MecCog mechanism representation framework (Darden et al., 2018) incorporates all these features. The representation formalism is based on the analysis of biological mechanisms developed in the philosophy of biology (Craver & Darden, 2013): Mechanisms are characterized as entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. In MecCog, a mechanism by which a genetic variant causes a disease phenotype is represented as a *mechanism schema* that displays the propagation of entity and activity perturbations across biological organizational stages (DNA→RNA→Protein→Complex→Organelle→Cell→Tissue→Organ→Phenotype) in the form of a graph (nodes are biological entities; directed edges are causal and labeled with productive activities) constructed from information in the biomedical literature in addition to established biological concepts. The schema structure uses graph properties such as branching, merging, and looping of sub-paths.

In this article, we describe the implementation of the MecCog framework as a web platform with a collaborative environment to manually construct, store, publish, browse, query, and comment on mechanism schemas for genetic diseases. The schema building tool in MecCog is supported by specially designed graphical notations, curated ontology-informed terminology for the annotation of mechanism components (entities and activities), an interactive graphical user interface (GUI) to construct the schema drawings, application programming interfaces (APIs) to fetch reference information and scientific figures, tight integration and hyperlinking of evidence to the graphics, and a secure server to save schemas as JSON (JavaScript Object Notation) objects. The platform supports edit, version, and share operations on each schema to facilitate collaborative work. Mature schemas can be published on the platform, thereby adding to the collection of disease mechanisms available for browsing by MecCog web-site visitors. Sketchier schemas with gaps, ambiguities, and uncertainties can also be published to indicate where additional work needs to be directed.

3.3 Methods and Results

3.3.1 Mechanism schema representation structure

In MecCog, a mechanism by which a genetic variant causes a disease phenotype is represented by multiple steps. Each step consists of a triplet with an input substate perturbation, a mechanism module, and an output substate perturbation (SSP-MM-SSP). A substate perturbation represents a perturbed biological entity (e.g. a DNA

base change, altered stability of a protein, altered abundance of a molecular complex, altered state of a cell). A mechanism module represents the productive activity (e.g. transcription, translation, or protein-protein interaction) by which the input sub-state perturbation produces the output sub-state perturbation. The succession of overlapping SSP-MM-SSP triplets represents perturbation propagation across stages of biological organization (DNA, RNA, Protein, Complex, Organelle, Cell, Tissue, Organ, and Phenotype), and together form a mechanism schema. In MecCog, a schema is represented as a graph where the nodes are SSPs and edge labels are MMs, as illustrated in Figure 3-1.

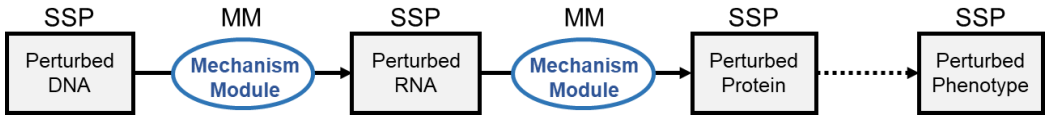


Figure 3-1. Principles of a mechanism schema. SSP: Substate Perturbation; MM: Mechanism Module. Each SSP represents a perturbed biological entity and each MM represents a productive activity (or a group of entities and activities) that produce an output SSP.

Evidence about SSPs and MMs is curated from the literature. Ambiguities in a mechanism and possible alternative mechanisms are represented in a schema by branching. Branch points may be labeled with the logical operators AND, OR, or AND/OR. The degree of confidence as to whether each SSP and MM is part of a schema is indicated by an evidence strength color code (red least confidence to green most confident) for the corresponding symbol. In addition to SSPs and MMs, five

other types of mechanism components are defined in MecCog: 1. *Unknown mechanism module* to represent ignorance about a mechanism component; 2. *Biomarker* to represent entities correlated with a disease phenotype; 3. *Environmental factor* to represent relevant external factors; 4. *Hypothetical therapeutic intervention site*; 5. *Known therapeutic intervention site.*

3.3.2 MecCog platform web-architecture

Figure 3-2 shows the web-architecture of MecCog. On the server-side, Node.js (an open-source JavaScript runtime environment) is used as the web-server, Sails.js is used to build the model-view-controller compliant web-application, and a MySQL relational database is used to store data on users and mechanism schemas. The MySQL database is connected to the web-application by the Object-relational mapper (ORM), Waterline, in Sails and all the database transactions use REST APIs secured by CSRF (Cross-site request forgery). The front-end GUI of MecCog is implemented using HTML, CSS, and Javascript, and is made responsive by Bootstrap.js javascript library. The schema building and visualizing GUI is powered by the Rappid Diagramming Javascript library (https://www.jointjs.com/). Rappid also provides a feature for converting diagrams to JSON format and for communicating with the database via AJAX requests. An open-source version of the IntenseDebate commenting system (https://www.intensedebate.com/) is used to render commenting forms.

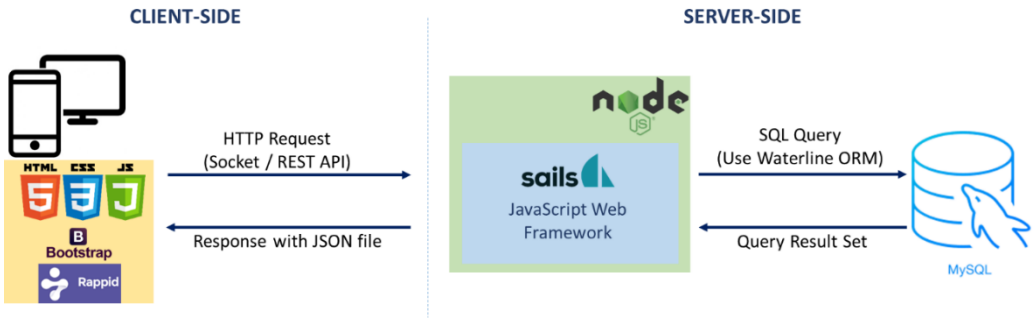


Figure 3-2. MecCog Web-Application Architecture. HTTP=Hypertext Transfer Protocol, REST API = Representational State Transfer Application Programming Interface, JSON=JavaScript Object Notation, SQL = Structured Query Language, ORM= Object-relational mapping.

3.3.3 Graphical notation of mechanism components in MecCog

Graphical notations symbolize components of a mechanism schema (Figure 3-3). An SSP (substate perturbation) is represented by a rectangle containing three types of information – the biological stage where the SSP occurs, the perturbation class name (curated from standard biomedical ontologies wherever possible), and the instance of that perturbation class. For example, a truncated *NOD2* protein can be represented by an SSP with *Protein* as the stage name, *Truncated Protein* as the SSP perturbation class name (from BioAssay Ontology (Visser et al., 2011)), and *NOD2:1007fs* as the instance of the SSP class. A biomarker is a special case of an SSP and is represented by the same shape but with a different color. An environmental factor is represented by a cloud icon. Known and hypothetical therapeutic intervention sites are represented by pink and blue octagons respectively. A known mechanism module is represented by a clear oval displaying the MM class or instance name, such as

transcription or protein folding. An unknown mechanism module is represented by a black oval.

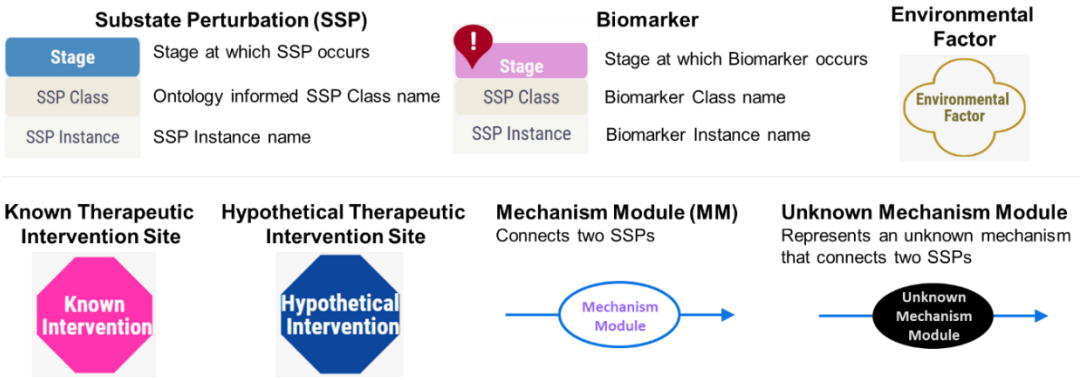


Figure 3-3. Graphical notations for components in a mechanism schema.

3.3.4 Mechanism schema meta-information and schema component annotations in MecCog

Table 3-1 summarizes the mechanism schema data model. Table 3-1A shows the meta-information of a schema. Each schema in MecCog is identified by a unique accession number automatically generated by the platform. Schema authors provide a schema name, a schema caption, genes that are part of the schema, keywords relevant to the mechanism, names of authors who constructed the schema, and the name of the curator who publishes the schema, monitors comments, and approves changes. Authors also provide a schema description with scientific background information.

Table 3-1B shows the mechanism component annotations. All mechanism components in a schema are annotated with a unique component identifier. Nine stage names may be assigned to an SSP component notation – DNA, RNA, Protein,

Complex, Organelle, Cell, Tissue, Organ, and Phenotype. For the molecular stages (DNA, RNA Protein, and Complex), a set of stage-specific SSP perturbation class names, such as SNV, mRNA abundance, or protein stability, have been compiled. Molecular stage MM classes, such as transcription, translation, and protein folding, are also defined (Table 3-2). Whenever possible the SSP and MM class names are curated from existing biomedical ontologies. Currently used ontologies are listed in Table 3-2. Where required, ontology terms may be prefixed with a modifier – increased, decreased, or altered. A MecCog schema builder may choose from the curated set of classes for a step, or may add new class names if needed. SSP and MM instance names are in free text. We are in the process of developing a disease mechanism ontology, based on the class names. Such an ontology is potentially useful for automatic text mining of SSP, MM, and triplet information from the literature, so speeding schema building by identifying relevant papers and sections of papers. Environmental factor names are in free text. Therapeutic intervention site components may be annotated with a potential therapy name or known drug name.

For all mechanism components, five types of evidence annotation are defined (Table 3-1C): 1. *For Evidence PubMed IDs* of papers that contain data supporting a component's role in a mechanism; 2. *Against Evidence PubMed IDs* for papers that provide data suggesting a mechanism component is incorrect; 3. *Links to figures in PMC* that illustrate aspects of a schema by summarizing experimental results and evidence for spatial and structural features; 4. User assigned *Confidence scores* with five levels (also used to automatically encode a component's confidence color) based

on the strength of the available evidence; 5. *Evidence Comments* - brief free text comments that summarize the evidence.

Table 3-1. Data Model of mechanism schema and component annotations. Text in parentheses indicates the data type.

Table 3-1A. Mechanism schema meta-information

Accession number (*Automatically generated, versioned, alphanumeric*)
Schema Name (*Free text – 300 character limit*)
Schema Caption (*Free text – 500 character limit*)
Schema Description (*Free text*)
Genes (*Free text*)
Keywords (*Free text*)
Curators (*Free text*)
Authors (*Free text*)

Table 3-1B. Mechanism component annotations

Mechanism Component	**Component Specific Annotation**
SSP (Substate Perturbation)	Component ID (Format: SSP#)
	Stage Name (*Predefined list*)
	SSP Class Name (*Predefined list with the facility to add new names*)
	SSP Instance Name (*Free text*)
MM (Mechanism Module)	Component ID (Format: MM#)
	MM Class Name (*Predefined list with the facility to add new names*)
	MM Instance Name (*Free text; Optional*)
Biomarker	Component ID (Format: BM#)
	Stage Name (*Predefined list*)
	Biomarker Class Name (*Predefined list with the facility to add new names*)
	Biomarker Instance Name (*Free text*)
Unknown Mechanism Module	Component ID (Format: MM#)
Environmental factor	Component ID (Format: EF#)
	Factor name (*Free text*)
	Component ID (Format: TT#)

Known therapeutic intervention site	Drug or Therapy name (*Free text*)
Hypothetical therapeutic intervention site	Component ID (Format: TT#)
	Potential therapy name (*Free text*)

'#' represents an integer number denoting the order of the schema component (for example SSP1, MM3, BM1, TT2)

Table 3-1C. Evidence annotations for the mechanism components

For Evidence PubMed IDs (*PMID number*)
Against Evidence PubMed IDs (*PMID number*)
Figure links from PubMed Central (*PMC figure number*)
Confidence score (*Predefined integer score range 1 to 5*)
Evidence comment (*Free text*)

Table 3-2. Curated class names for Substate Perturbations (SSP) and Mechanism Modules (MM) at the molecular stages. The class names are curated from biomedical ontologies and are prefixed with the ontology name abbreviation.

Mechanism Component	**Stage Name**	**Number of Classes**	**Class Names**
SSP	DNA	10	**SO**:SNV, **NCIT**:IN/DEL, **NCIT**:Methylation Sites, **NCIT**:Chromosomal Rearrangement, **VARIO**:Copy Number Variation, DNA Repeats, **NCIT**:DNA Structure, **NCIT**:Holliday Junction, DNA Supercoiling, DNA Curvature
	RNA	12	**SO**:SNV, **NCIT**:IN/DEL, **VARIO**:Edited RNA, Fused mRNAs, RNA Repeats, mRNA Abundance, Splicing Isoform

			Abundance, Edited RNA Abundance, Other RNA Abundance, **EDAM**:RNA Secondary Structure
	Protein	16	**NCIT**:Missense Mutation, **NCIT**:IN/DEL, **BAO**:Truncated Protein, **NCIT**:Post-Translational Modification Site(s), **NCIT**:Phosphorylation Site(s), Fused Proteins, Protein Sequence Repeats, **PLOSTHES**:Protein Abundance, Splicing Isoform Abundance, Post-Translational Modified Protein Abundance, **BAO**:Phosphorylated Protein Abundance, **MESH**:Protein Conformation, **NCIT**:Protein Dynamics, **MESH**:Protein Stability, **MI**:Allostery, **MESH**:Quaternary Protein Structure
	Complex	12	**GRO**:Protein-RNA Complex Abundance, **CRISP**:Spliceosome Abundance, DNA-RNA Complex Abundance, **BIPON**:RNA-RNA Complex Abundance, **GO**:Protein-DNA Complex Abundance, Transcription Complex Abundance, **GO**:Transcription Factor Complex Abundance, DNA-Scaffold Complex Abundance, DNA Replication Complex Abundance, DNA-Histone Complex Abundance, **EDAM**:Protein-Ligand Complex Abundance, **ADMO**:Protein-Protein Complex Abundance.
MM	-	24	**IXNO**:Cleavage Rate, **NCIT**:Synthesis Rate, **GO**:Transport Rate, **CRISP**:Protein Degradation, **NCIT**:RNA Degradation Rate, **NCIT**:Protein

			Folding Rate, **NCIT**:Nonsense-Mediated Decay Rate, **NCIT:** Transcription Rate, **GO**:Translation Rate, DNA Internal Interactions, RNA Internal Interactions, Protein Internal Interactions, DNA-RNA Interaction, RNA-Ligand Interaction, **IOBC**:RNA-RNA Interaction, **GRO**:DNA Protein Interaction, **GO**:Scaffold Protein Binding, **GO**:Basal Transcriptional Machinery Binding, **GO**:Histone Binding, **NCIT**:RNA-Protein Interaction, **NCIT**:Protein-Protein Interaction, **NCIT**:Ligand Binding, Le Chatelier, **GO**:Signaling.

SO: Sequence Ontology (Eilbeck et al., 2005); NCIT: National Cancer Institute Thesaurus (Sioutos et al., 2007); VARIO: Variation Ontology (Vihinen, 2014); EDAM: EMBRACE Data and Methods (Ison et al., 2013), MESH: Medical Subject Headings (https://meshb.nlm.nih.gov/), MI: Molecular Interactions (https://www.ebi.ac.uk/ols/ontologies/mi), CRISP: Computer Retrieval of Information on Scientific Projects Thesaurus (https://bioportal.bioontology.org/ontologies/CRISP), GRO: Gene Regulation Ontology (https://bioportal.bioontology.org/ontologies/GRO), BIPON: Bacterial interlocked Process Ontology (Henry et al., 2017), GO: Gene Ontology (The Gene Ontology Consortium, 2019), ADMO: Alzheimer Disease Map Ontology (Malhotra et al., 2014), PLOSTHES: PLOS Thesaurus (https://bioportal.bioontology.org/ontologies/PLOSTHES), BAO: BioAssay Ontology (Visser et al., 2011), IXNO: Interaction Ontology

(https://bioportal.bioontology.org/ontologies/IXNO), IOBC: Interlinking Ontology for Biological Concepts (https://bioportal.bioontology.org/ontologies/IOBC).

3.3.5 Rules for constructing mechanism schemas in MecCog

1. Each schema begins with a genome perturbation and ends in perturbation of a disease-related phenotype such as greater risk of a disease.

2. Overall, the sequence of SSPs in a schema progresses through successive stages of biological organization, from DNA, through RNA, proteins, macromolecular complexes, organelles, cells, tissues, organs, and finally to a phenotype. There may be one or more or no SSP at any particular stage of organization and the order of the stages need not follow a prescribed order. For instance, the schema for Lynch syndrome (http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3), where a causative mutation results in decreased DNA mismatch repair, reverts to the DNA stage after stages involving macromolecular complexes.

3. Each pair of SSPs is linked by an MM. The granularity of an MM may be a single activity (such as splicing, protein-protein interaction, ligand binding, or protein folding) or may represent telescoped combinations of entities and activities (such as protein synthesis, or cell-cell signaling). If an activity is unknown, the black oval unknown mechanism module notation is used.

4. Class names for the SSPs and MMs at the molecular stages (DNA, RNA, Protein, and Complex) can be selected from the pre-compiled list shown in Table 3-2. If existing names are inadequate, new names can be used. At higher organizational

stages, class names are user-provided. Wherever possible these should be part of existing biomedical ontologies. The NCBO BioPortal site (https://bioportal.bioontology.org/) is a source for ontology terms. Since most ontologies describe the normal state of a system, a user may select one of the in-built modifiers (increased, decreased, altered) to prefix a class name so as to represent a perturbed state.

5. An evidence-based confidence score (on the scale of 1 to 5, where one indicates low confidence and five indicates high confidence) should be assigned to each SSP and MM. Evidence on which a confidence score is based should be recorded in the form of supporting/contradicting PMIDs and PMC figure URLs, together with appropriate free text commentary.

6. Two or more possible sub-paths can exist in a schema either because of ambiguity due to conflicting evidence, or alternative sub-mechanisms. Branch points should be labeled with OR, AND or AND/OR.

7. Schemas should explicitly include steps only where there is a perturbation from the normal system. Where the function of a portion of a schema is unperturbed, for example, representing the standard activity of transcription operating on a perturbed input DNA sequence or a standard cell signaling process operates with more or less input signal, that section of the schema should be telescoped into a single mechanism module.

3.3.6 Steps in constructing, managing, and publishing mechanism schemas

Before beginning schema building a new user must register on the MecCog platform. A registered user may select the "Build Schema" tab to initiate building a new schema or the "My Schemas" tab to access the workspace for managing and editing their existing schemas. Figures 3-4A and 3-4B show the two interfaces used in schema construction: A. The *Initiate Mechanism Schema* form used to enter meta-information about a schema, and B. The *Schema Builder* GUI used to draw a schema. In the schema builder, mechanism components can be dragged and dropped from the mechanism component catalog panel to the drawing board panel. Clicking on a component displays five associated control icons: i) Icon to connect to other components; ii) Icon to adjust the component size; iii) Icon to clone the component; iv) Icon to show the pop-up box; and v) Icon to delete the component. Clicking on a component also renders a component-specific annotation form on the rightmost panel of the interface (labeled in Figure 3-4B). This form is used to enter the stage, class, and instance name of the components, prefix class names with a perturbation type if needed (increased, decreased, or altered), and record the evidence annotations (listed in the Evidence Annotations Table 3-1C). This is a dynamic form that automatically provides predefined possible perturbation class names for the selected stage (listed in Table 3-2) and creates fields for adding new PubMed IDs and PMC image URLs. The NCBI E-utilities application programming interface (API) is used on the server-side to fetch publication details for the PubMed IDs. All the evidence annotations are transferred to the current component pop-up box together with hyperlinked PMIDs and PMC image URLs. The pop-up box can be visualized by clicking the '*i*' shaped

control icon of the component. Confidence score values selected in the annotation form are used to automatically apply the appropriate color to the current schema component (red: score 1, orange: scores 2, 3, 4 and, green: score 5). The color of the edge connecting two components is inherited from the target component, so indicating causal confidence. Schemas are saved to the database using the *Click to save* button. For each schema, a unique accession number is automatically generated in the database. The accession number format has a section indicating the version (default is .1) of the schema. The schema builder GUI also has a panel of interactive buttons to undo, redo, clear page, zoom, auto-layout, export (in SVG and PNG formats), and print schema diagrams.

Figure 3-4C shows the view of a registered user's workspace. Each schema can be versioned, edited, shared with other MecCog users, published, or deleted, using operation-specific buttons. The workspace has three sections: i) The Unpublished Mechanism Schemas section catalogs work-in-progress schemas. ii) The Published Mechanism Schemas section catalogs published schemas. A button to remove each of these from the public collection is provided. iii) The Shared Mechanism Schemas section catalogs schemas that have been shared with the current user. For schemas with edit access privilege, the *Copy to My Space* operation is enabled, allowing the creation of a copy for the user to work on independently. All the schema accession numbers in the workspace page are hyperlinked to the schema specific landing page (described in the next section).

Published schemas are available for browsing via the main webpage of the MecCog site (as shown in Figure 3-4D) without the need for logging in. There is a search bar that allows schemas to be searched by gene name, keyword, or any component class/instance name. MYSQL's FULLTEXT indexing (https://dev.mysql.com/doc/refman/5.6/en/innodb-fulltext-index.html) feature is used to support the search operation.

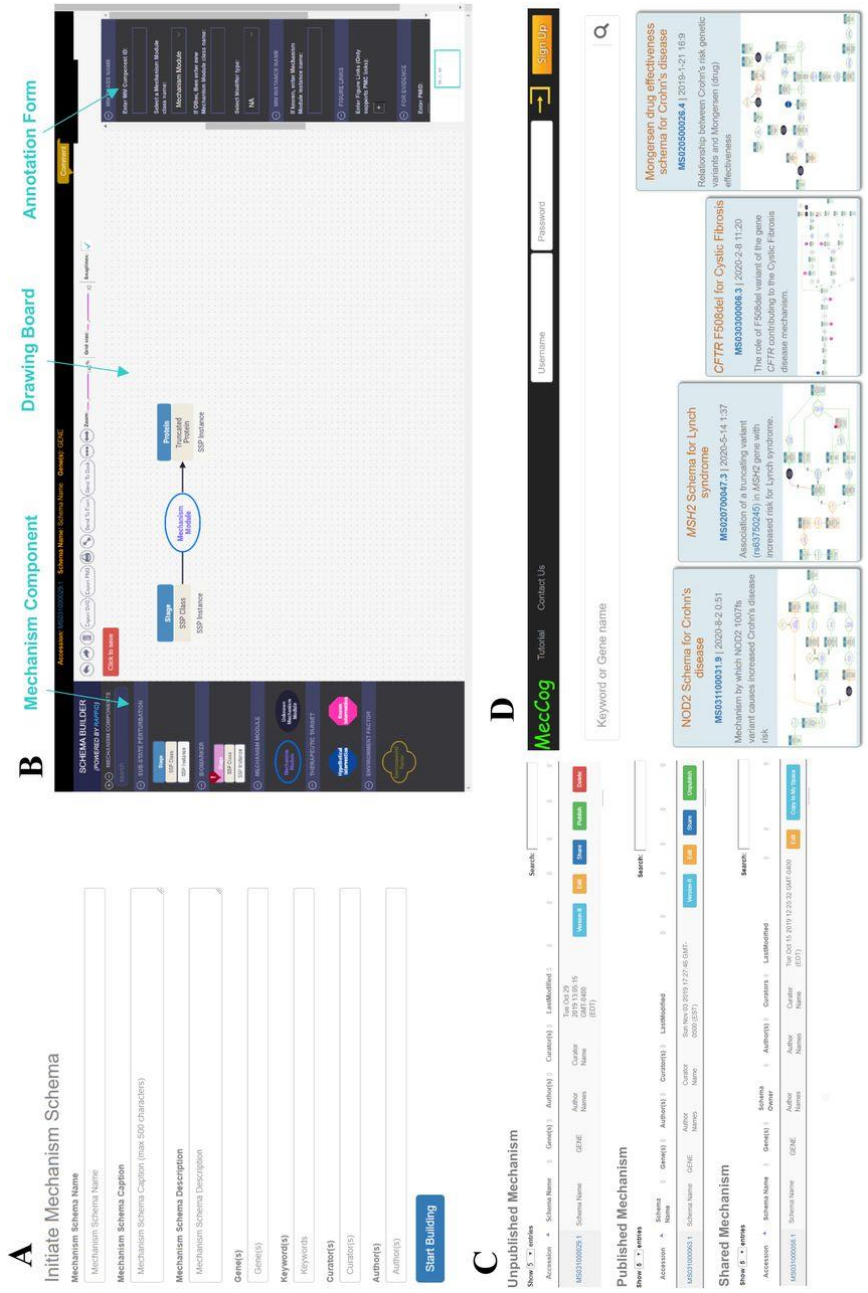


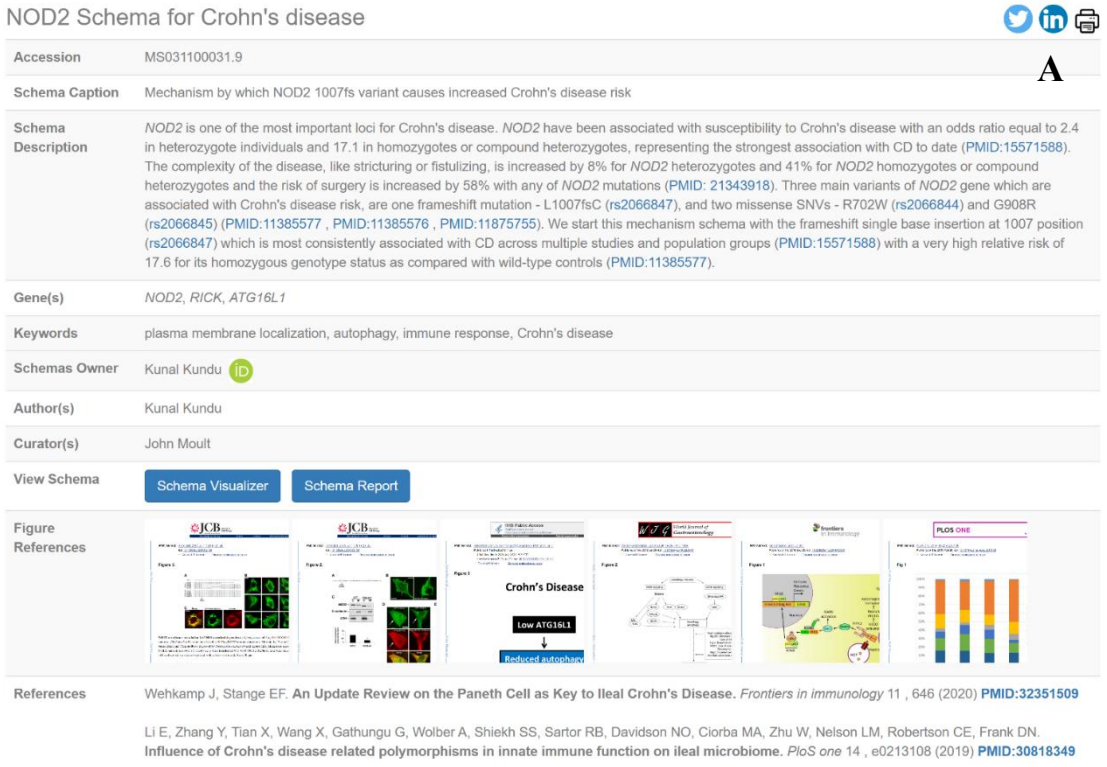
Figure 3-4. Graphical User Interface (GUI) to construct, manage, and browse mechanism schemas. Figure 3-4A shows the form for entering meta-information on a schema. Figure 3-4B shows the schema builder interface, including the mechanism component catalog panel (left), the drawing board panel, and the annotation form panel (right). Figure 3-4C shows the user workspace interface. Figure 3-4D shows a portion of the MecCog main webpage, designed to facilitate browsing publicly available mechanism schemas.

The structured organization of mechanistic knowledge in MecCog allows this search to be used to find common entities or activities and common classes used in different schemas. On the main page, schemas are presented in a masonry layout view. Each tile in the view displays the schema name, schema caption, hyperlinked accession number of the schema linking to the corresponding schema landing page (described in the next section), and a hyperlinked schema image linked to an interactive web-based visualization of the schema (also described in the next section).

3.3.7 Schema landing page, schema visualizer, and schema report

A schema landing page displays schema meta-information in a tabular layout (Figure 3-5A). A novel feature of this page is the display of references and hyper-linked PubMed IDs providing evidence for each aspect of the schema as well as PMC images selected to illustrate aspects of the mechanism. The *Schema Visualizer* button on the landing page directs a user to the GUI for interactively navigating the mechanism schema (as shown in Figure 3-5B). The visualizer inherits all the

interactive features of the schema builder GUI (described previously). A unique feature of the visualizer is the tight integration of the graphical notations for the mechanism components and the associated evidence information (presented in the pop-up box). The pop-up box (yellow-colored box in Figure 3-5B) displays hyperlinked ontology sources for the SSP/MM class name (if the term is from an ontology), a brief builder-provided commentary on the evidence, and hyperlinked PMIDs and PMC figure IDs. There is a help icon ('?') in the visualizer to display the mechanism schema key. Clicking on the Comment button in the visualizer opens a modal box to view or enter comments about the schema. The *Schema Report* button on the landing page generates a narrative report in which the meta-information, mechanism components, and evidence annotations about the schema are presented in a structured format. The schema content can be downloaded as a JSON file from the landing page using the download icon. The page also has social media share icons.



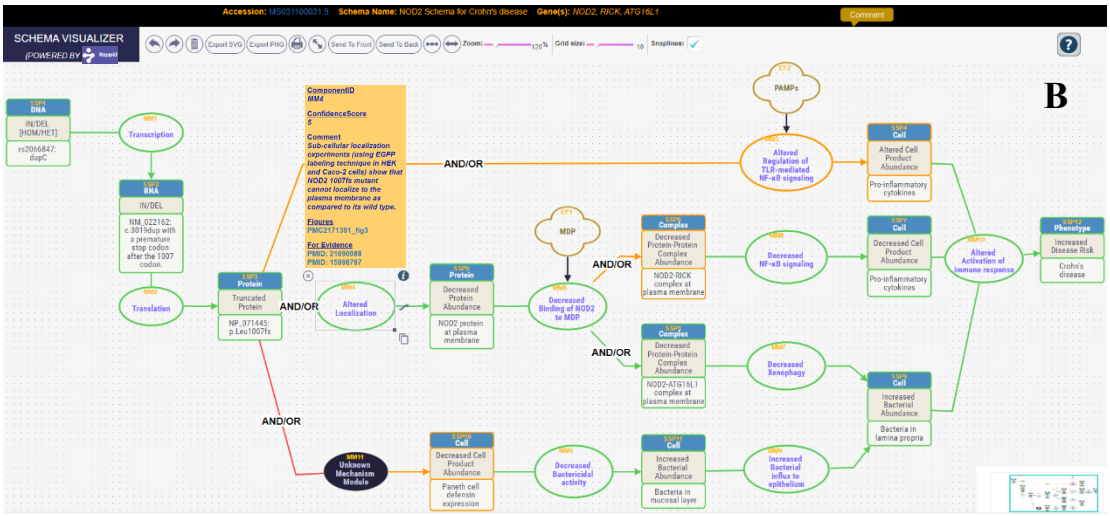


Figure 3-5. NOD2 mechanism schema entry in MecCog. This schema describes the known mechanisms by which a frameshift mutation (rs2066847) in the NOD2 gene causes an increased risk for Crohn's disease. Figure 3-5A shows part of the landing

page of the NOD2 schema, displaying the meta-information in tabular format. This page includes the collection of thumbnails of PMC figures selected to illustrate aspects of the mechanism and the list of references with PubMed IDs from which evidence was derived (the list is truncated here - there are 26 references). Figure 3-5B shows the schema visualizer GUI used for interactive navigation of the schema. For this schema, four possible submechanisms with varying levels of evidence (indicated by the confidence colors – red=low, orange=medium, and green=high) are included. The example yellow pop-up box in Figure 3-5B displays hyperlinked evidence for the associated MM. The comment button on the top can be used to open a modal box, allowing a user to view and add comments. Details of the mechanism are described in the text.

3.3.8 An example MecCog Schema: Known mechanisms by which a frameshift mutation in the NOD2 gene causes an increased risk of Crohn's disease

Figure 3-5 shows two pages of a MecCog mechanism schema (http://www.meccog.org/mchain/showpubchain?accession=MS031100031.9) describing the mechanism by which a frameshift mutation (rs2066847; NM_022162.3:c.3019dup (p.Leu1007fs)) in the *NOD2* gene causes an increased risk of Crohn's disease (CD). *NOD2* is the first gene for which variants were found to be associated with altered CD risk (Hugot et al., 2001b; Ogura et al., 2001; Yamamoto & Ma, 2009) and the 1007fs mutation is most consistently associated with CD across multiple studies and population groups (Economou et al., 2004) with a very high relative risk of 17.6 for its homozygous genotype status as compared with wild-type

controls (Ogura et al., 2001). Figure 3-5A shows the landing page displaying the meta-information of the schema and the list of references used as evidence in the schema, together with figures used to illustrate aspects of the mechanism.

Figure 3-5B shows the view of the NOD2 1007fs schema in the interactive visualizer. This schema was constructed using information about mechanism reported in 26 research articles. The left-most SSP (SSP1) represents the DNA stage perturbation (i.e. the single base insertion of cytosine - rs2066847 in the *NOD2* gene). The paths in the schema show how the effect of this perturbation propagates through the RNA, protein, complex, and cell stages (represented by the stage-specific SSPs and MMs) so causing the increased Crohn's disease risk phenotype (SSP12). At the RNA stage (SSP2), the rs2066847 variant causes the insertion of a cytosine after the first nucleotide of codon 1007, so introducing a premature downstream stop codon. This leads to the protein stage perturbation, a truncated NOD2 protein (SSP3) missing the last 33 amino acids of the wild-type sequence (Lécine et al., 2007). Following this, the schema branches represent the multiple submechanisms by which the truncated NOD2 protein may lead to the increased Crohn's disease risk phenotype by altering the activation of the immune response (MM10) (Negroni et al., 2018; Park et al., 2007; W. Strober & Watanabe, 2011). All the branches are labeled 'AND/OR' since none has fully compelling supporting evidence. The submechanism of each branch is outlined below.

A) The top branch (SSP3→MM3 → SSP4) shows a potential alteration to NOD2-dependent regulation of Toll-like receptor (TLR) mediated NF-κB signaling that produces pro-inflammatory cytokines in response to the pathogen-associated molecular patterns (PAMPs) such as lipopolysaccharide (LPS), or muramyl dipeptide (MDP). This path is sparse and labeled medium confidence (orange) because the mechanism of interaction between NOD2 and TLRs is not known, nor is it clear how that interaction normally results in increased production of pro-inflammatory cytokines (Underhill, 2007). Different models have been proposed to describe the mechanism: synergistic production of TNF-α by NOD2 and TLR4 (Wolfert et al., 2002); activation of the inflammasome by NOD2 via RICK to produce IL-1β from pro-IL-1β generated as the result of TLR signaling (A. Sarkar et al., 2006); and MDP (the primary agonist for NOD2 (Grimes et al., 2012)) dose-dependent TNF-α production by NOD2 and TLR2 (Borm et al., 2008). Further, for none of these possibilities has the effect of the NOD2 100fs variant been investigated. These details are provided in the pop-up box for MM3.

B) The middle branch shows that truncated NOD2 protein has lost its ability to localize to the plasma membrane (MM4 → SSP5) (Barnich et al., 2005; Morosky et al., 2011) where binding to incoming MDP normally produces an activated state of the protein (Al Nabhani et al., 2017). In turn, activated NOD2 forms complexes with RICK and with ATG16L1 (Barnich et al., 2005; Travassos et al., 2010). The schema shows these effects as lower abundance of the NOD2-RICK complex (MM5 → SSP6) (Barnich et al., 2005) and the NOD2-ATG16L1 complex (MM5 → SSP8)

(Travassos et al., 2010). There is no experimental evidence of the NOD2 1007fs protein's impact on complex formation. Therefore the MM5 → SSP6 step in the schema is labeled medium confidence (orange). Following this step, the lower abundance of the NOD2-RICK complex alters downstream NF-κB signaling (SSP6→MM6→SSP7) (Barnich et al., 2005; Caruso et al., 2014; Girardin et al., 2003; Lécine et al., 2007), resulting in lower pro-inflammatory cytokine production, so contributing to an altered activation of the immune response (MM10) (Negroni et al., 2018; Park et al., 2007; W. Strober & Watanabe, 2011; Vilela et al., 2012). The perturbation of the NOD2-ATG16L1 complex affects the xenophagy process (autophagy against bacteria) (MM7) (Travassos et al., 2010) so leading to an increase in the abundance of bacteria in the lamina propria (SSP9) (Sidiq et al., 2016) and thereby likely contributing to a more aggressive response from other components of the immune system, as indicated by the altered activation of immune response (MM10). This sub-path is labeled high confidence (green) as its mechanism components are well understood based on the available evidence in the literature. The yellow pop-up box for MM4 shows an example of an evidence commentary with an associated hyperlinked PMC figure and PMIDs.

C) The lower branch of the schema provides examples of the representation of a gap in knowledge and of overall low confidence. Commensal bacteria are largely prevented from penetrating the gut wall by an outer mucosal barrier and the epithelial cell layer. Paneth cells situated in the gut epithelial layer produce a range of antibiotic defensin peptides to aid in preventing commensal bacteria from traversing the

mucosal layer. Some data suggest that this process is partly dependent on MDP binding to NOD2 in these cells, likely signaling that significant numbers of bacteria are getting through to the epithelial cell layer, and so triggering an increased response. Data supporting that view come from an experiment showing stimulation of NOD2 by MDP binding induces production of defensin HNP-1 (human neutrophil peptide 1) in Caco-2 cells (Yamamoto-Furusho et al., 2010). It has also been shown that the NOD2 1007fs protein fails to induce the production of defensin hBD2 (human β-defensin-2) in several epithelial cell lines (Voss et al., 2006). Hence the link between the presence of the 1007fs variant (SSP3) and increased defensin production (currently SSP10). But the mechanism by which MDP binding to NOD2 normally causes defensin production is unknown, hence the black oval (MM11) linking those two SSPs. There is also evidence from other studies that do not support the mechanism represented by this schema path: In two out of four CD cohort studies (Hayashi et al., 2016; Simms et al., 2008; Jan Wehkamp et al., 2005), and in NOD2 deficient mouse organoids (Wilson et al., 2015), the association between NOD2 and defensin was not reproduced. Hence this branch is labeled low confidence (red). Further along this schema branch, the decrease in defensin production (SSP10) leads to an increased abundance of bacteria in the mucosal layer (SSP11) due to decreased bactericidal activity (MM8). In turn, this contributes to increased bacterial abundance in the lamina propria due to increased bacterial influx from the mucosal layer (MM9) and finally leads to the altered activation of the immune response (MM10).

3.3.9 Representation of biomarker and therapeutic intervention sites in MecCog

Figure 3-6A shows an example of the use of the biomarker symbol, in part of the Lynch syndrome schema (http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3). In this schema, microsatellite instability is a diagnostic biomarker (Vilar et al., 2014) for Lynch syndrome, resulting from defective base mismatch repair machinery, in turn a consequence of a mutation (rs63750245: C>T) in the *MSH2* gene. Figure 3-6B shows an example of a putative therapeutic intervention site in a Crohn's disease schema (http://www.meccog.org/mchain/showpubchain?accession=MS020500019.2) describing the mechanism by which a missense variant (rs3197999: G>A; R703C) in the *MST1* gene (coding for Macrophage Stimulating Protein, MSP) increases disease risk. The missense variant causes a lower abundance of the MSP-RON protein complex by one or both of two mechanisms: a weakened protein-protein interaction (Chao et al., 2014; Gorlatova et al., 2011) and reduced MSP abundance. Lower abundance of the complex is expected to result in reduced intracellular signaling affecting macrophage activation (Häuser et al., 2012; L. Kretschmann et al., 2010; M. H. Wang et al., 2002) and/or epithelial cell survival and growth (Danilkovitch et al., 2000; Neurath, 2014). An appropriate compound that bridges the structural interface between MSP and RON could restore wild-type abundance of the complex and hence signaling strength and so eliminate the downstream consequences. (Of course, many factors affect whether this is in fact an effective therapeutic strategy.)

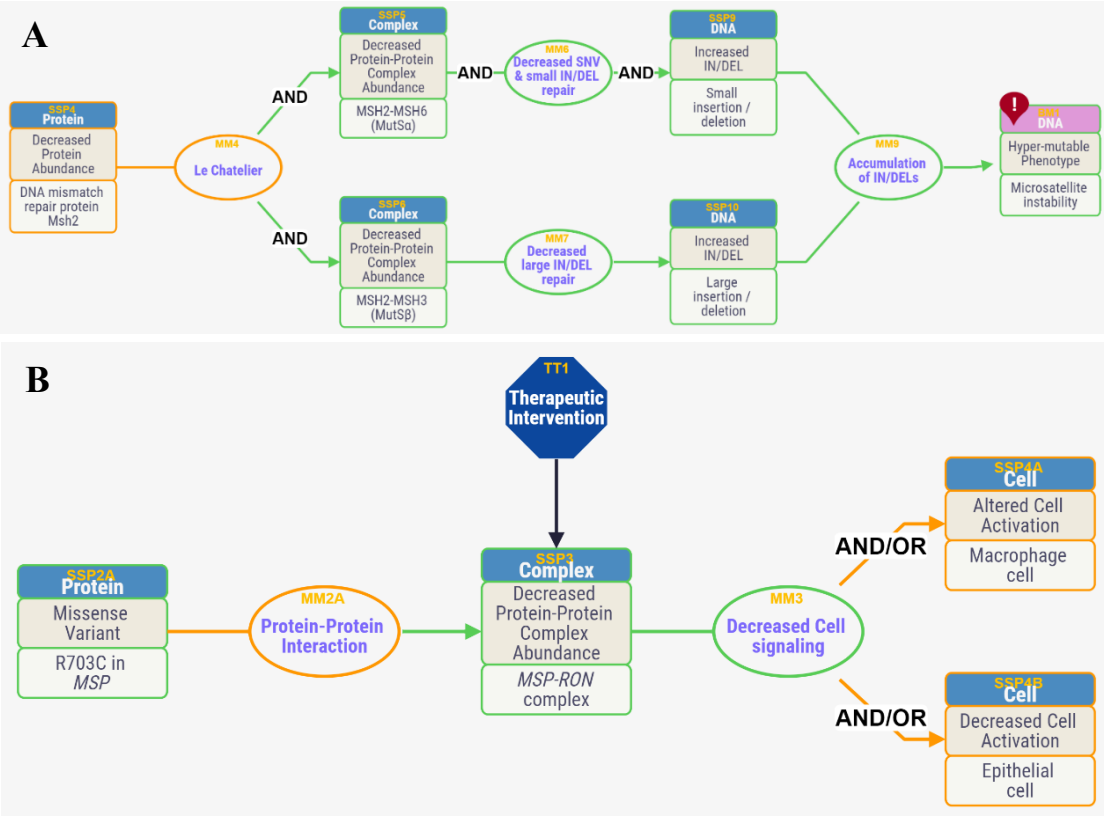


Figure 3-6. Biomarker and Therapeutic intervention site representation in MecCog. Figure 3-6A shows part of a Lynch syndrome schema (http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3) where the presence of a nonsense mutation (rs63750245: C>T) in the MSH2 gene causes Microsatellite instability (MSI), a known biomarker (symbolized by the red location icon on the SSP) for the Lynch syndrome. Figure 3-6B shows part of a Crohn's disease schema (http://www.meccog.org/mchain/showpubchain?accession=MS020500019.2) where the decreased abundance of the MSP-RON protein complex (SSP3) is a hypothetical therapeutic intervention site, indicated by the blue octagon. In this case, an

appropriate small molecule binding across the protein-protein interface might restore the wild-type abundance.

3.3.10 Validation of the MecCog representation framework

Eleven MecCog mechanism schemas (nine for Crohn's disease, one for cystic fibrosis, and one for Lynch syndrome) have so far been published, with additional schemas in progress on breast cancer and Alzheimer's disease. Validation and improvement of MecCog content is obtained by soliciting feedback from specialists in the disease described in each schema. Feedback on the representation technique and platform can be provided using MecCog's *contact us* form, encouraging users to provide suggestions and report problems. MecCog has also been used as an educational tool for senior undergraduate students in a Human Genetics class at the University of Maryland, providing valuable feedback, for example, linking PMC figures to aspects of schemas.

3.4 Discussion

We have developed MecCog, a graphical knowledge representation framework, to describe genetic disease mechanisms in a structured mechanism schema format. MecCog facilitates the assembly of mechanistic information in terms of perturbation propagation across stages of biological organization, evaluation of the evidence related to that information, and identification of uncertainties, ambiguities, and ignorance. The MecCog web platform provides functionalities to create, store, browse, and search schemas. Graphical notations are annotated with ontology-

informed class terms so as to consistently and intuitively represent types of mechanism components found in schemas. The schema interactive visualizer in MecCog tightly integrates the graphics, text, and hyperlinks to evidence sources.

Each schema in MecCog describes mechanisms by which a single genetic variant contributes to the increased risk for the disease phenotype. For complex trait genetic disease and cancer, multiple genetic variants contribute to disease phenotypes (Lilyquist et al., 2018; Peter et al., 2011). Further, contributions from variants may not be independent, as reflected by evidence of epistatic effects between pairs of variants for complex trait disease (Y. Li et al., 2020; Lin et al., 2017). The MecCog formalism also supports mechanism schemas with multiple input genetic perturbations. Interactions between these inputs results in a mechanism graph. An example for Crohn's disease is a barrier integrity mechanism graph constructed by combining schemas on loci relevant to bacterial penetration of the gut-lining mucosal layer (Figure 3-7). This graph incorporates a number of non-additive interactions between mucin gene variants affecting mucosal-layer integrity (*MUC1*, *MUC2*), variants affecting the unfolded protein response (*XBP1*, *ORMDL3*), and variants affecting autophagy (*NOD2*, *ATG16L1*, *LRRK2*, *IRGM*).

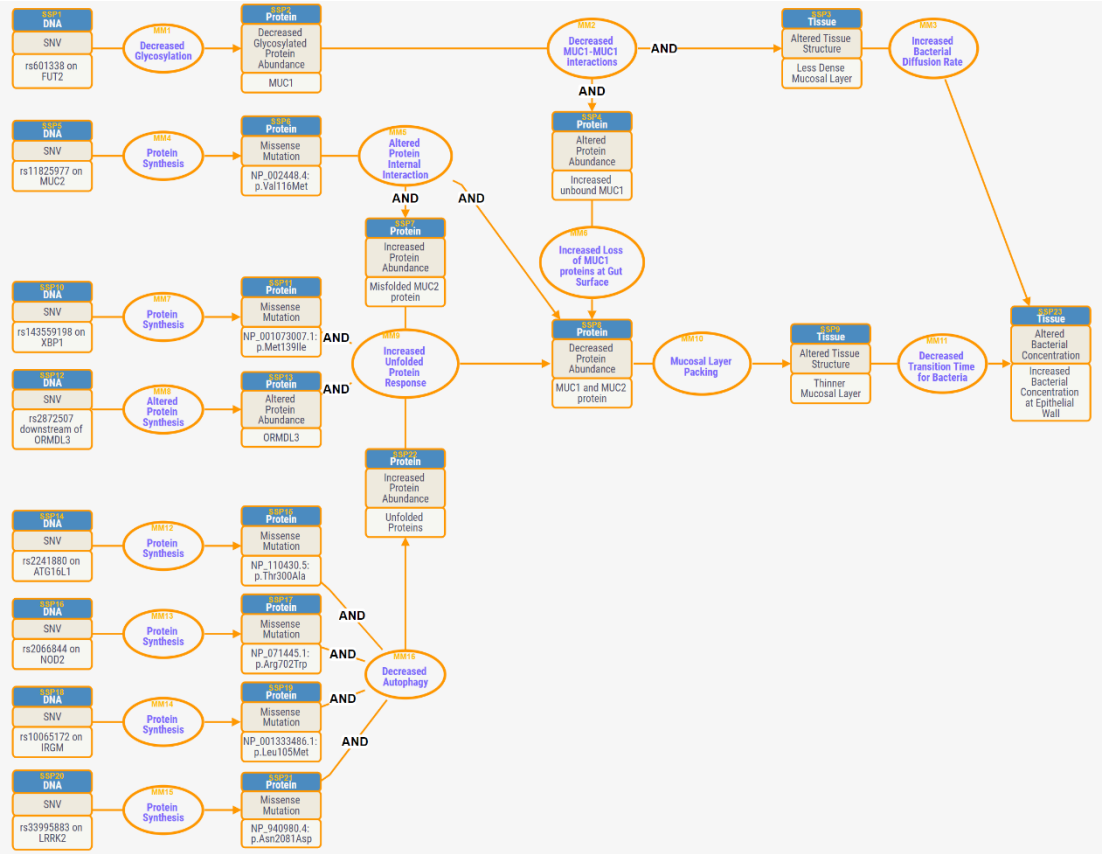


Figure 3-7: Part of the barrier integrity mechanism graph for Crohn's disease, showing the role of risk variants that affect bacterial penetration through the mucosal layer. SSP=Substate Perturbation and MM=Mechanism Module. An SNP (SSP1) in *FUT2* affects glycosylation (MM1) of MUC1 (Kelly et al., 1995; Mcgovern et al., 2010). The hypoglycosylated state of the MUC1 (SSP2) affects its interaction strength with other mucins (MM2) resulting in a less dense mucosal layer (SSP3) (Hall et al., 2017a). Weaker mucin interactions result in more rapid diffusion of bacteria through the mucosal layer (MM3) and faster mucin loss at the gut surface (MM6), one of the three factors contributing to an overall lower abundance of mucin (SSP8). A second factor is a missense SNP in *MUC2 (SSP5)*, the main constituent of gut mucin, resulting in a less stable protein (MM5) (Heazlewood et al., 2008; Moehle

et al., 2006). The third factor is the state of the unfolded protein response (UPR) in the mucin-producing Goblet cells (MM9). The UPR system reduces protein production in response to the accumulation of excessive misfolded or unfolded protein in the ER (X. Ma et al., 2017). Because of the normal rapid loss of mucins at the gut interface, Goblet cells are among the most hard-working protein-producing cell types (Gersemann et al., 2009), and so are particularly susceptible to changes in the UPR (X. Ma et al., 2017). The UPR threshold is influenced by variants in two genes, *XBP1* (Kaser et al., 2008) and *ORMLD3* (Barrett et al., 2008; Moffatt et al., 2007). The extent of misfolded protein is also influenced by the efficiency of autophagy (MM16), involving variants affecting four genes – *NOD2* (Warren Strober et al., 2014), *IRGM* (Chauhan et al., 2016), *LRRK2* (Hui et al., 2018) and *ATG16L1*(Salem et al., 2015). As the overall disease mechanism described in the mechanism graph has not been tested experimentally, all the branches are labeled medium confidence (orange color).

Currently, MecCog schemas are manually constructed, relying on human understanding to extract and infer causal connections between mechanism components from literature. Given the scattered and incomplete nature of mechanistic information in literature, this process is complex and requires a combination of prior biological knowledge together with searching for and assimilating new facts and evidence from the literature. These activities are labor-intensive and work best when the schema builder is an expert on the schema subject. To achieve scale for the resource, we require an expert-crowdsourcing strategy, soliciting inputs from

appropriate domain experts. The resource is structured so that experts can build schemas based on their knowledge and can also edit and comment on existing schemas. The current version of the MecCog platform supports these activities in the following ways: i) acknowledging contributors to a schema as authors and curators, ii) providing a schema specific commenting interface to solicit input, and iii) allowing versioning of schemas to update content. To implement the crowdsourcing model, we will work closely work with disease-specific research communities (such as IBD Genetics, Crohn's & Colitis Foundation, and Alzforum).

An obvious question is whether mechanism schemas can be constructed automatically given the structured and unstructured data available in the biomedical domain. The structure of a mechanism schema shares features with that of knowledge graphs (KG), a knowledge representation system initiated by Google in 2012 (https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html). There nodes (aka subjects) represent entities such as real-world objects, events or concepts, and edges (aka predicates) link the nodes with relationships. Because of a KG's ability to integrate and represent multi-relational databases, many biological KGs (Sosa *et al.*, 2019; Celebi *et al.*, 2019; Chen *et al.*, 2019; Himmelstein *et al.*, 2017; PDBe-KB consortium, 2020; https://digitalinsights.qiagen.com/coronavirus-network-explorer/) are being generated, using a combination of manual and automated mining of subject-predicate-object (SPO) triplets from biomedical literature and from bioinformatics relational databases. The elemental SSP-MM-SSP units of a MecCog schema are a subset of

SPO triplets and so it is in principle possible to construct a schema by extracting appropriate triplets from a comprehensive knowledge graph. However, preliminary tests of this process suggest that current knowledge graphs do not capture a large fraction of the triplets incorporated in the corresponding mechanism schemas. There are multiple reasons for this, including the absence of biological knowledge in knowledge graphs and the absence of causal reasoning components. We envisage that in the future comprehensive and well-structured KGs will be combined with a repository of biological knowledge and reasoning machines to generate a wide variety of biological mechanisms, as well as providing evaluation of evidence strength and identifying current gaps in mechanism knowledge.

3.5 Acknowledgments

This work was supported in part by the National Institute of Health [R01GM104436 to JM]. KK's conference travel related to this research was supported in part by NSF award DGE-1632976. We thank Rappid (https://www.jointjs.com/) for providing their JavaScript library under an academic license. We thank Lipika Ray, Yizhou Yin, Maya Zuhl, Christian Presley, and undergraduate students in the University of Maryland Human Genetic course for many useful suggestions and comments on MecCog software. We thank Mark Tonelli for feedback on the MecCog framework and for reviewing the cystic fibrosis schema.

Chapter 4: A framework to quantitatively represent and analyze mechanisms relating genetic variants to complex trait disease

4.1 Introduction

Genome-wide association studies (GWAS) have identified thousands of associations between genetic variants and complex trait diseases (Dehghan, 2018). A major focus has been to further identify the mechanisms underlying these associations in order to develop full disease models. The construction of disease models is non-trivial for two primary reasons. First, mechanistic information is scattered throughout the literature, is established with varying degrees of confidence, and the results are often ambiguous or contradictory to other findings. Second, biological knowledge of disease mechanisms is qualitative and descriptive, and so not immediately amenable to most quantitative modeling. To address the first obstacle, we have previously developed MecCog (Darden et al., 2018) a framework for integrating and representing disease mechanism knowledge in a structured format. The resulting schemas are able to comprehensively represent mechanisms by which genetic variants cause disease phenotypes. However, they are still purely qualitative. Here, we describe two developments aimed at overcoming the second obstacle. The first is a framework to quantitatively encode MecCog schemas, providing a method for building quantitative models using biological knowledge of disease mechanism. Using simulated data, we show that these Disease Mechanism Circuit (DCM) models are useful for exploring properties of complex trait diseases that are not experimentally accessible. While these models are informative, incomplete knowledge of the quantitative aspects of

mechanism limits their application. To address this, we have developed a hybrid neural network model of the relationship between genetic variants and disease-related phenotypes. We show that these models are able to learn key quantitative aspects of disease mechanism from GWAS data, so opening the way for large scale quantitative modeling of complex trait disease.

We have applied both the disease mechanisms circuit model and the neural network to the analysis of the role of reduced gut wall barrier integrity in Crohn's, a complex trait inflammatory bowel disease (Gorlatova et al., 2011; Pal, Chao, et al., 2017; Pal, Kundu, et al., 2017). More than 200 GWAS loci related to Crohn's disease risk (De Lange et al., 2017) are known. Follow-up experimental work has revealed many of the mechanisms underlying those associations and has led to an overall model of the disease (Ahluwalia et al., 2018; Atreya & Siegmund, 2018; Fischer & Neurath, 2017) involving 11 subprocesses (Jostins et al., 2012). Briefly, the overall mechanism is as follows: The human gut contains a large load of commensal bacteria. Even in a healthy individual, a small number of these will continually pass through the gut wall and enter the underlying tissue (Schroeder, 2019). A concentration of immune system cells immediately below the gut wall normally ensures that penetrating bacteria are quickly removed. Some of the Crohn's risk variants affect the integrity and thickness of the outer mucosal layer of the gut wall, the tightness of the epithelial cell junctions, and the effectiveness of outer antibacterial defenses, resulting in a higher flux of bacteria into the tissues. Risk variants in other genes affect the efficiency with which the innate immune system and the adaptive immune system deal with the penetrating

bacteria. Other risk variants affect the appropriateness of the inflammatory response. Finally, further risk variants affect the extent of gut wall tissue damage occurring as a result of inflammation and the efficiency of repair. In any particular individual, a subset of risk variants will be present - on average approximately half the maximum possible load. Each individual with the disease will likely have a different subset of risk variants from other individuals, and in this sense, the disease mechanism is different in each case.

We demonstrate the utility of the circuit model by investigating two phenomena related to nonlinearity in the relationship between genetic variants and disease phenotypes. GWAS does not capture inter-gene dependencies of phenotypes, and these missing epistatic effects have been proposed as an explanation of why GWAS results only reflect a fraction of disease heritability (Maher, 2008). First, we use the model together with simulated GWAS data to investigate the extent to which differences in the effect size of single risk variants is observed in different individuals. Unavoidably GWAS provides average properties over a study population's diverse genetic backgrounds and cannot show whether a variant has a large or small effect in a particular individual (Stringer et al., 2011). If disease phenotypes were a linear function of an individual's relevant variants, this would not be an issue. But inter-gene interactions will result in varying effect sizes. Knowledge of an individual's variant effect sizes may be key in judging what available treatments will be most effective. Many model organism studies have demonstrated that the effect size of mutations varies as a function of the genetic background (Chow et al.,

2016; Galardini et al., 2019; Vu et al., 2015). We show that the non-linear components of the barrier integrity circuit model do result in a dramatic variation of variant effect sizes across individuals in a GWAS population. Second, we directly address the extent to which interactions between relevant genetic variants are non-linear, which is epistatic, using the circuit model. The extent of epistasis is a much-debated question in studies of complex trait disease (Manolio et al., 2009; Wei et al., 2014). Pair-wise knockout and RNA silencing experiments in a range of model organisms have shown that these effects are very wide-spread (Byrne et al., 2007; Cardinale & Cambray, 2017; Costanzo et al., 2016). Despite this evidence, pair-wise statistical tests using human GWAS data seldom detect epistatic effects (J. Li et al., 2011). As with the variation in effect size for single variants, one reason for this may be that population epistatic effects are masked by averaging over all genetic backgrounds in a sampled population. Consistent with the model organism studies, we find a very large range of non-additive effects between risk variant pairs in a GWAS population.

The circuit model provides an effective means of investigating the effect of genetic variants in each individual in a population, so overcoming major limitations in the GWAS data. But it requires knowledge of not only the disease mechanism (represented by the mechanism graph) but also the functional form of each node's response and the parameter values for those functions. Although functional forms may be estimated, as we do in the circuit model study, obtaining true parameter values is extremely labor-intensive, and because of the properties of this type of

biological circuit, maybe almost impossible to obtain with sufficient accuracy (Chou & Voit, 2009). Machine learning methods, including neural networks, have been used to predict disease phenotypes from genotype inputs for complex trait diseases (Laksshman et al., 2017; Pal, Kundu, et al., 2017; Zeigler et al., 1990). Model parameters are learned from the data and at least for neural networks, there should be good robustness with respect to data noise (Borodinov et al., 2019). But although these methods perform as well as additive models (Pal, Kundu, et al., 2017), they provide no insight into the underlying mechanism. An attractive solution to this issue is the incorporation of prior knowledge of network architecture so that only nodes representing known functional units in the system are included. The approach was first suggested in 1992 for modeling bioreactor function (Psichogios & Ungar, 1992), but has been little explored. Recently, it has been successfully used for building an integrated model for yeast cells (Dcell (J. Ma et al., 2018a)). We have developed a sparse neural network representation of genetic disease mechanisms, a Mechanism Architecture Neural Network (MANN) model. We show that MANN input/output relationships can be learned from GWAS data, that individual nodes reproduce the model's functional form and behavior, and that it is possible to distinguish between alternative graph connectivity, so allowing the evaluation of alternative mechanism hypotheses.

4.2 Results

4.2.1 Mechanism of barrier integrity disruption in Crohn's disease

MecCog mechanism schemas (Darden et al., 2018) are composed of input substate perturbation – mechanism module – output substrate perturbation (SSP-MM-SSP) triplets, where each SSP represents a perturbed entity at some stage of biological organization (e.g. a DNA variant, altered protein abundance, altered cell state) and each MM (e.g. transcription, protein-protein interaction, cell signaling) represents the productive activity by which the input SSP produces the output SSP. Simple schemas follow the progression of substate perturbations from a single genetic variant across stages of biological organization (RNA, protein, macromolecular complexes, organelles, cells, organs) to the disease phenotype, and are suitable for representing the mechanisms underlying monogenic disease and individual cancer-relevant mutations (see www.meccog.org for examples). For complex trait diseases like Crohn's, where multiple genetic variants affect the disease phenotype, contributing individual variant schemas are combined to produce a mechanism graph, including interactions between the variants. Figure 4-1 shows the graph for part of the barrier integrity mechanism, where eight Crohn's risk variants affect the concentration of bacteria at the epithelial wall of the gut, B_e. A higher bacterial concentration at the epithelial wall causes higher bacterial flux into the underlying tissues where they interact with the gut immune cells that lead to inflammation and increased risk for the disease (Okumura & Takeda, 2018). Here, a SNP in *MUC1* (Franke et al., 2010) affects the strength of the corresponding protein's interactions with other mucins (Kadayakkara et al., 2010) (M1), implying a less dense mucosal layer (S1). In turn,

weaker mucin interactions result in more rapid diffusion of bacteria through the mucosal layer (M3) and faster mucin loss at the gut surface (S2), one of three factors contributing to an overall lower abundance of mucin (S3). A second factor is a missense SNP in *MUC2*, the main constituent of gut mucin, resulting in a less stable protein (M5) (Heazlewood et al., 2008; Moehle et al., 2006). The third factor is the state of the unfolded protein response (UPR) in the mucin-producing Goblet cells (M8). The UPR system reduces protein production in response to the accumulation of excessive misfolded or unfolded protein (X. Ma et al., 2017). Because of the normal rapid loss of mucins at the gut interface, Goblet cells are among the most hard-working protein-producing cell types (Gersemann et al., 2009), and so are particularly susceptible to changes in the UPR (X. Ma et al., 2017). The UPR threshold is influenced by variants in two genes, XBP1 (Adolph et al., 2012) and ORMDL3 (M. Li et al., 2017). The extent of overall misfolded protein is also influenced by the efficiency of autophagy (M10), involving variants affecting four genes (NOD2 (Warren Strober et al., 2014), IRGM (Chauhan et al., 2016), LRRK2 (Hui et al., 2018), and ATG16L1 (Salem et al., 2015)). This barrier integrity graph is not complete: Additional components include a SNP in FUT2 affecting mucin-mucin interaction strength via alterations to glycosylation (Hall et al., 2017b). IL10 (Hasnain et al., 2013) also influences the UPR response, and in turn the influence of IL10 variants is modulated by variants in two members of the intracellular signaling pathway, STAT1 and STAT3 (Hasnain et al., 2013). Other significant contributions to barrier integrity are the efficiency of xenophagy (Knodler & Celli, 2011) and antibiotic production (J. Wehkamp et al., 2008) by epithelial wall Paneth cells and the

tightness of the epithelial cell-cell junctions, influenced by a variant affecting C1orf106 (Mohanan et al., 2018). Nevertheless, the graph captures sufficient disease biology to provide a test system for the Disease Mechanism Circuit (DCM) and Mechanism Architecture Neural Network (MANN) models.

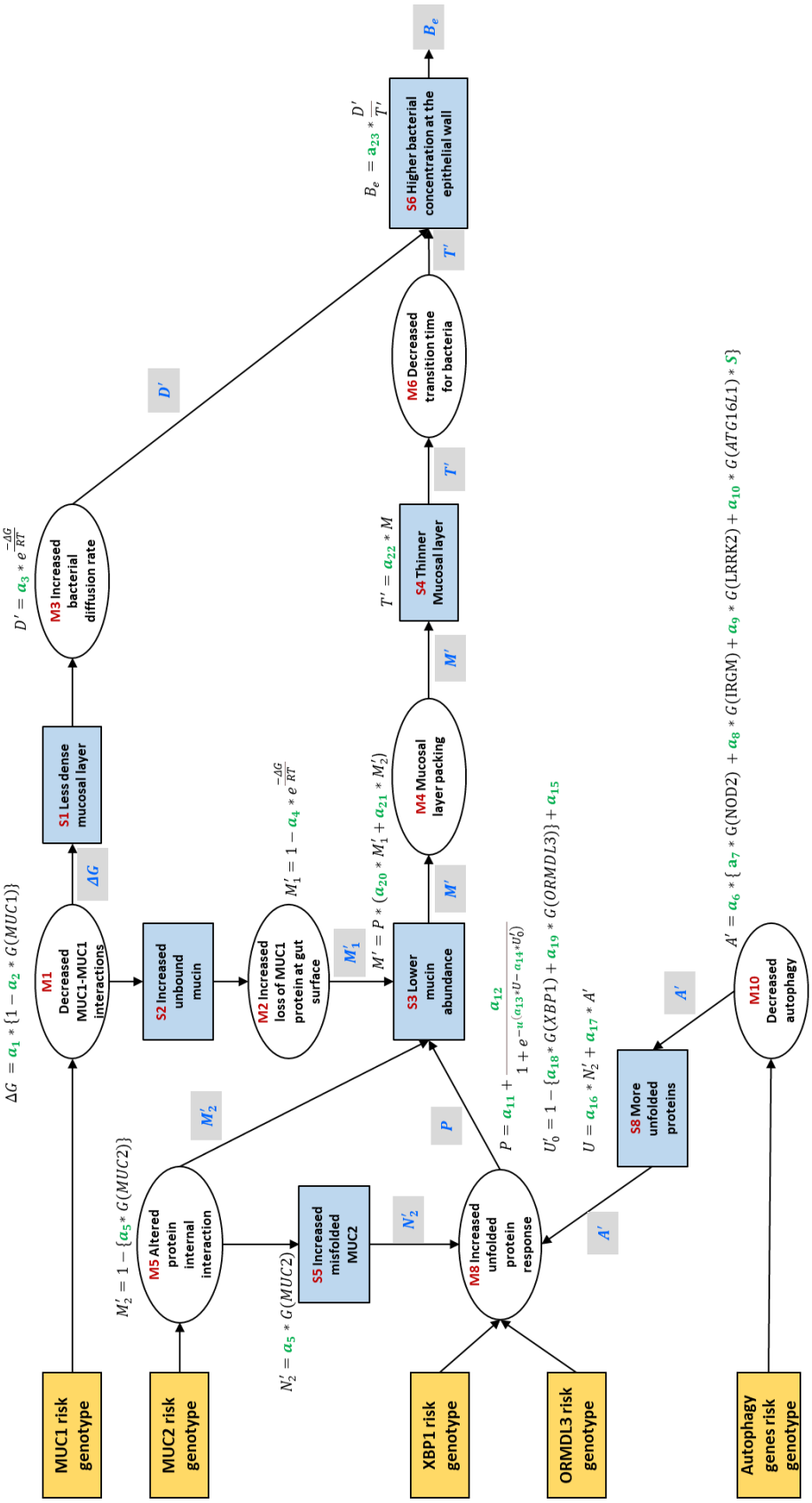

MUC1 risk genotype
MUC2 risk genotype
XBP1 risk genotype
ORMDL3 risk genotype
Autophagy genes risk genotype
M1 Decreased MUC1-MUC1 interactions
M2 Increased loss of MUC1 protein at gut surface
M3 Increased bacterial diffusion rate
M4 Mucosal layer packing
M5 Altered protein internal interaction
M6 Decreased transition time for bacteria
M8 Increased unfolded protein response
M10 Decreased autophagy
S1 Less dense mucosal layer
S2 Increased unbound mucin
S3 Lower mucin abundance
S4 Thinner Mucosal layer
S5 Increased misfolded MUC2
S6 Higher bacterial concentration at the epithelial wall
S8 More unfolded proteins
$\Delta G = a_1 * \{1 - a_2 * G(MUC1)\}$
$D' = a_3 * e^{\frac{-\Delta G}{RT}}$
$M_1' = 1 - a_4 * e^{\frac{-\Delta G}{RT}}$
$M_2' = 1 - \{a_5 * G(MUC2)\}$
$N_2' = a_5 * G(MUC2)$
$M' = P * (a_{20} * M_1' + a_{21} * M_2')$
$T' = a_{22} * M$
$B_e = a_{23} * \frac{D'}{T'}$
$P = a_{11} + \frac{a_{12}}{1 + e^{-u(a_{13} * U - a_{14} * U_0')}}$
$U_0' = 1 - \{a_{18} * G(XBP1) + a_{19} * G(ORMDL3)\} + a_{15}$
$U = a_{16} * N_2' + a_{17} * A'$
$A' = a_6 * \{a_7 * G(NOD2) + a_8 * G(IRGM) + a_9 * G(LRRK2) + a_{10} * G(ATG16L1) * S\}$
ΔG
D'
M_1'
M_2'
N_2'
M'
T'
P
A'
B_e

Figure 4-1. Part of the barrier integrity mechanism graph for Crohn's disease, showing the role of variants in eight genes that affect bacterial penetration through the gut mucosal layer. Genetic variant inputs to the network are shown as yellow rectangles, substate perturbations as blue rectangles, and mechanism modules as clear ovals. Edge labels show variables output from the preceding node and provide input to subsequent nodes. For each mechanism component, the corresponding approximate node function is shown (for details see Methods). Parameters determining quantitative behavior are in green.

4.2.2 Building a disease mechanism circuit (DMC) from a disease mechanism graph

Each substate perturbation (SSP) and mechanism module (MM) in the barrier integrity mechanism graph represents an altered value of a physical quantity: abundance of a protein, strength of a protein-protein interaction, thickness of the epithelial layer, and so on. The magnitude of these quantities can be computed given one or more input values to the SSP or MM nodes. For example, in Figure 4-1 there are two inputs into the node S6 'Higher bacterial concentration at the epithelial wall' – a thinner mucosal layer, T, output from S4; and a faster bacterial diffusion rate D through the layer, output from M3. Quantitatively this node (S6) represents the perturbation of the bacterial concentration at the epithelial cell wall arising from a thinner and less dense mucosal layer. The simplest model of this process is the diffusion of bacteria across a planar mucosal layer, with a rate proportional to the diffusion constant D and inversely proportional to the layer thickness, T. Then, the

flux of bacteria arriving at the epithelial wall increases with faster diffusion (larger *D*) and with a thinner layer, *T*. We assume that at steady state, the bacterial concentration at the epithelial cell wall (B_e) can be expressed as proportional to the incoming flux.

$$B_e = a_{23} * \frac{D}{T}$$

where a_{23} is constant.

Approximate analytical functions are assigned to each node. Table 4-1 (Methods) shows the node functions and the estimated parameters values. Any particular combination of risk genotypes for variants in the eight genes (the input risk genotype vector, where for each variant, 0 represents no risk alleles, 1 a single, heterozygous risk allele, and 2, homozygous risk alleles) will propagate through the circuit to determine B_e. A higher B_e, indicating a higher bacterial concentration at the epithelial wall, will cause higher bacterial flux into the underlying tissues and so higher disease risk (Okumura & Takeda, 2018). Thus, B_e is an intermediate Crohn's disease risk phenotype.

4.2.3 Sources of non-linearity in the DMC of barrier integrity

As noted earlier, nonlinear interactions between variants will result in a nonlinear relationship between input risk genotypes and disease phenotypes. There are two explicit sources of inter-gene non-linearity in the circuit: First, the division function (F12 in Table 4-1) at node S6, representing the relationship between bacterial diffusion rate (D), the thickness of the mucosal layer (T), and the bacterial concentration B_e. Second, the sigmoid function (F7 in Table 4-1) at node M8

representing the unfolded protein response (UPR). In practice, the sigmoid function dominates the non-linear behavior of the circuit.

4.2.4 Functional form of the Unfolded Protein Response (UPR)

The UPR is a cellular response to the accumulation of unfolded protein. There are two main components, both resulting from unfolded proteins binding to and so activating kinases, IRE1, and PERK (Mendez et al., 2015). Activated IRE1 acts as an RNAase that results in abnormal splicing of *XBP1*, leading to increased chaperone production and so an increase in the amount of successfully folded proteins. Activated PERK phosphorylates the α-subunit of eukaryotic translation initiation factor 2 (eIF2α) (Harding et al., 1999) and traps it in the GDP-bound inactive state, so blocking eIF2α recycling. That results in the attenuation of global protein synthesis (Harding et al., 2000; Wek et al., 2006). Three experimental studies (Korennykh et al., 2009; Han Li et al., 2010; Mendez et al., 2015) have shown that IRE1 RNase activity (representing UPR activation) has a sigmoidal response to the concentration IRE1 enzyme. The Hill coefficient found in these studies is between 3.3 and 8 (The Hill coefficient is a dimensionless parameter characterizing the steepness of a sigmoid-like response (Frank, 2013). No data are available for the response of PERK to increase in unfolded protein abundance, relevant to the barrier integrity model. We expect that the Hill coefficient of that response will not be less than for IRE1, and will likely be greater: whereas a relatively gradual increase in chaperone concentration (mediated by IRE1) is expected, protein production shut-down is a serious decision for a cell, and so may be abrupt.

Sigmoid functions, in which at a low concentration of some molecule in the system there is no response, but at some threshold, there is a co-operative transition to a large system response, which then saturates, are very often found in biological system input/output relationships (Andersen et al., 2002; Frank, 2013). The Hill coefficients are often large, for example in MAPK signaling pathways up to 8 (Blüthgen & Herzel, 2003), in the response of the bacterial flagellar motor to the concentration of the chemotaxis response regulator CheY-P, 20 (Yuan & Berg, 2013), and in oocyte maturation in response to progesterone concentration in Xenopus Oocytes, a Hill coefficient of 35 (Ferrell & Machleder, 1998).

In the barrier integrity model, the sigmoid function at node M8 describes a cell's total protein production rate (*P*) in response to the total abundance of unfolded or misfolded protein (*U*) in the cell:

$$P = a_{11} + (\frac{a_{12}}{1 + e^{-u(U-U_0)}})$$

where $a_{11}, a_{12}\ and\ u$ are constants.

The model assumes that below some threshold of unfolded protein abundance, U_0, cells produce new protein with a rate independent of that quantity. Above the threshold, there is a progressive shut-down of protein production, with a rate determined by the parameter '*u*'. We have investigated the behavior of the circuit models for two values of *u (-1* and *-6)* corresponding to Hill coefficients of about 3.0 and 15.

Figure 4-2 illustrates how the sigmoid function results in a variation in the effect size of single variants as a function of genetic background and in variable epistatic interactions between risk alleles affecting gene pairs. At low unfolded protein abundance, represented by the plateau on the left of the plot, cell protein production is unaffected. Then, at some unfolded protein threshold (around 0.75 in the figure) there is a rapid decline in production to the low plateau (0.1 of the maximum in this model) at the right side of the plot. Panel A shows the positions of the three risk genotypes of *NOD2* for a particular genetic background from the other seven genes affecting this node. All three *NOD2* genotypes fall on the initial plateau, so that production and hence B_e are unaffected by the which genotype is present. That is, the effect size is zero. In panel B, a different genetic background from the other genes results in a higher level of unfolded protein and so shifts the *NOD2* genotypes to the right. The no-risk genotype is still on the plateau, but the single risk variant is at the start of the slope, and the homozygous risk variant genotype is near the bottom. The homozygous risk genotype now has a very large effect size. The *NOD2* genotypes in each individual in a population will fall at different positions depending on the individual's genetic background, and so a range of effect sizes is produced.

To see one way in which non-linear, epistatic, interactions are created between pairs of genes, consider two genes with a genetic background such that each gene's genotypes all fall on the initial plateau, as for *NOD2* in panel A. The linear expectation is that the phenotype effect of the combined risk variants for the two genes will be the sum of the individual effects, in this example zero. But it is easy to

see that the combined impact on the unfolded protein level may actually result in a large loss of protein production by producing a point on the slope. As with the single variant case, different genetic backgrounds will produce different degrees of nonlinearity.

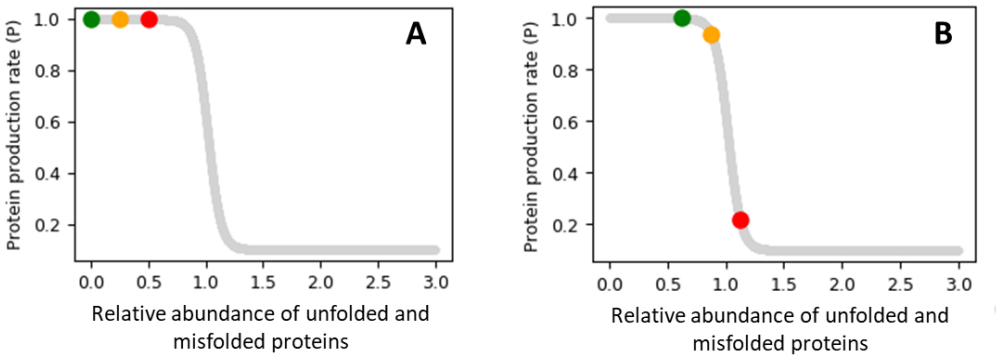


Figure 4-2. Model of the unfolded protein response (UPR). The X-axis is the relative abundance of unfolded protein and the Y-axis is the fraction of full protein production in the cell. At low unfolded protein abundance, protein production is unaltered, but at a threshold of around 0.75 in this model, production begins to be reduced and then falls fairly rapidly to a lower threshold (0.1 of full production). Risk variants in eight genes affect the unfolded protein abundance. The three points on each panel represent the unfolded protein abundance for each *NOD2* genotype (no risk variant, green; heterozygous risk variant, orange; and homozygous risk variant, red) for a particular genetic background of the other seven genes. In panel A, all the genotypes fall on the leading full protein production plateau, so that production is unaffected by the genotype status of *NOD2*. In panel B, a higher risk variant background in the other seven genes results in a higher unfolded protein abundance. The zero risk allele

NOD2 genotype is still on the plateau, the single risk allele is on the start of the downslope, resulting in a small decrease in production, and the homozygous risk allele genotype falls near the bottom of the slope, resulting in a large change. Examples here are for the more strongly non-linear model.

4.2.5 Characteristics of the barrier integrity model

The quantitative behavior of the barrier integrity circuit was simulated by calculating the output relative epithelial layer bacterial concentration (B_e) for each possible vector of input risk genotypes. Genotypes were encoded as 0 for the homozygous non-risk alleles, 1 for heterozygous, and 2 for homozygous risk alleles. There are 6561 (3^8) possible input genotype vectors. Parameters values of the circuit are listed in Table 4-1. Each genotype vector was feedforward to generate the corresponding output B_e value. Two sets of simulation data were generated corresponding to moderately non-linear behavior of the UPR ($u = -1$ in F7, Hill coefficient approximately 3) and a higher value ($u = -6$, Hill coefficient approximately 15). In order to generate a reference linear approximation, a linear regression fit of B_e to the eight genotype input values was performed, using the python scikit-learn package's linear regression module (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) with default parameters.

Figure 4-3A shows the distribution of relative bacterial abundance, B_e, for the set of all possible risk variant genotypes. For the moderately non-linear model ($u=-1$), the

distribution is skewed towards low B_e, with ~79% of genotype vectors generating B_e values less than 10, and only ~4.5% above 20. In contrast, Figure 4-3C shows that B_e values generated by the more strongly non-linear model (u=-6) are skewed more towards large B_e values, with ~24% above 20. Figure 4-3B shows that for the moderately non-linear model there is an approximately parabolic curve relating model B_e values to those generated by linear regression, with the predicted values for low-risk situations negative. That is, a best fit linear model substantially underestimates the high risk B_e values and will be inaccurate for low risk. For the more strongly non-linear model (Figure 4-3D), the regression fit is not monotonic and has little relationship to the true B_e values. Thus, a linear model is not able to represent the circuit's behavior.

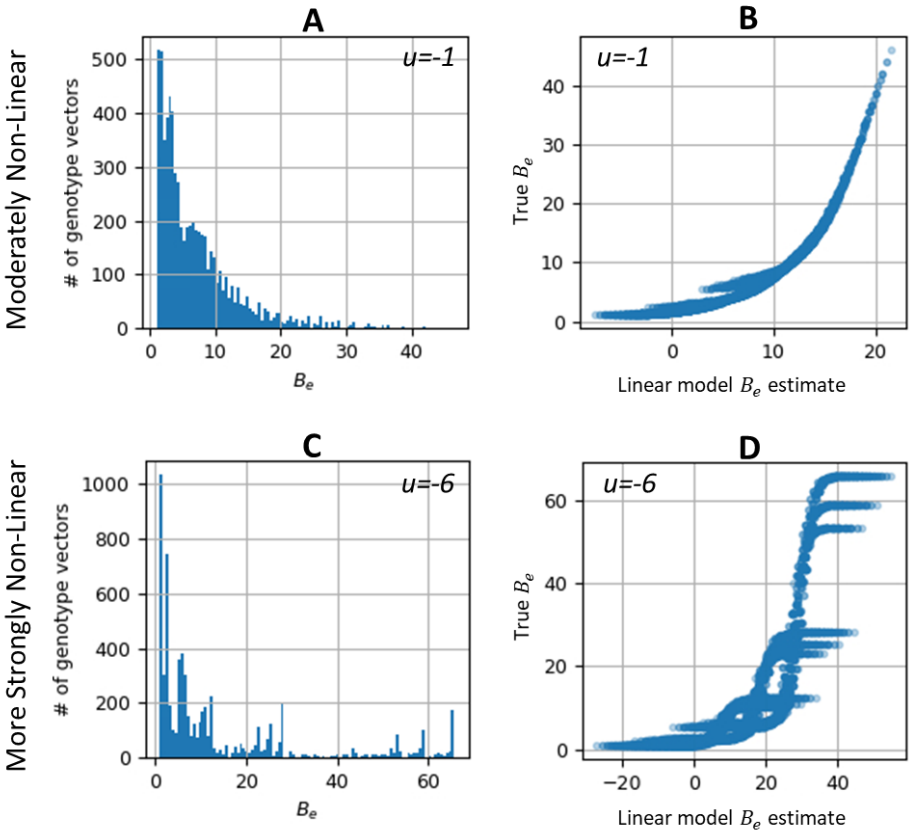


Figure 4-3. Distributions of bacterial concentration **(**B_e) across the 6561 input genotype vectors. Figure 4-3A shows the distribution of B_e values and Figure 4-3B shows the scatter plot between the true B_e values (Y-axis) and a linear regression approximation (X-axis) in the moderately (u=*-1*) non-linear UPR model. Figures 4-3C and D show the corresponding results for a more non-linear model (u=*-6*). The B_e distribution plot is skewed towards low B_e (<10) in the moderately non-linear case but skewed towards larger B_e (>20) in the more strongly non-linear case. The linear regression approximation fit is poor for both cases and is not monotonic for higher non-linearity.

4.2.6 Single variant effect varies across genetic backgrounds

The Barrier Integrity DMC provides a human disease model for investigating the degree to which the population average obtained from a GWAS experiment masks variation across individuals with different genetic backgrounds. To this end, we have examined the variation in effect size for each of the eight risk variants as a function of genetic background.

For every risk variant in the barrier integrity graph, there were 2187 (3^7) genetic backgrounds to be considered. For a specific background, the effect size of the risk variant was calculated as:

$$Variant\ Effect\ Size\ (V_s) = B_{e_{gi}} - B_{e_{g0}}$$

where $B_{e_{g0}}$ is the B_e value when the genotype of the risk variant is *0* (non-risk allele homozygous) in the background *g*, and $B_{e_{gi}}$ is the B_e value when the genotype of the risk variant is *i=1* (heterozygous) *or 2* (risk allele homozygous) in the background *g*. To simplify the analysis, only *i=2* was considered. In addition to considering all possible backgrounds, we also considered just those backgrounds found in a human population, using the WTCCC genotype dataset of 2000 Crohn's patients and 3000 controls (Burton et al., 2007).

Figure 4-4 shows the variation in the effect size of the *ATG16L1* risk variant. In the moderately non-linear model (Figure 4-4A), the effect size in the WTCCC population varies from near zero to ~14. The maximum value for any genotype vector is 42 (Figure 4-3), so that the effect size for this one variant spans 1/3 of the full range. The average is only 2.6. That is, the GWAS result suggests that this risk variant will only make a small contribution to increasing bacterial concentration, but in fact, for particular individuals, fully 1/3 the maximum possible bacterial load is caused by this single variant. The variation is even larger in the more strongly non-linear model (Figure 4-4B), varying between ~zero and ~60, spanning the total spread found across all genotype vectors (Figure 4-3). The average is 10.5, for some individuals only about 1/6 of the actual effect size. Although individual variant effect size tends to increase with the number of other risk variants present, there is a very large scatter. For instance, in the moderately non-linear ATG16L1 model, at a typical level of background risk alleles found in WTCCC (~1100-1200 on the X-axis, Figure 4-4A),

effect size varies from about 1 to 9. The other seven genes show similar patterns (Figure S6).

Thus, to the extent that the model represents the properties of real biology, average effect sizes derived from GWAS data mask a wide range of variation across individuals - in one individual an ATG16L1 risk allele may contribute almost no increase in bacteria penetrating the gut mucosal layer. In another individual, the increase may be very substantial, so contributing greatly to increased disease risk. As described earlier (Figure 4-2), in this model, the variation in effect size primarily arises from the nature of the sigmoid response to the level of unfolded protein.

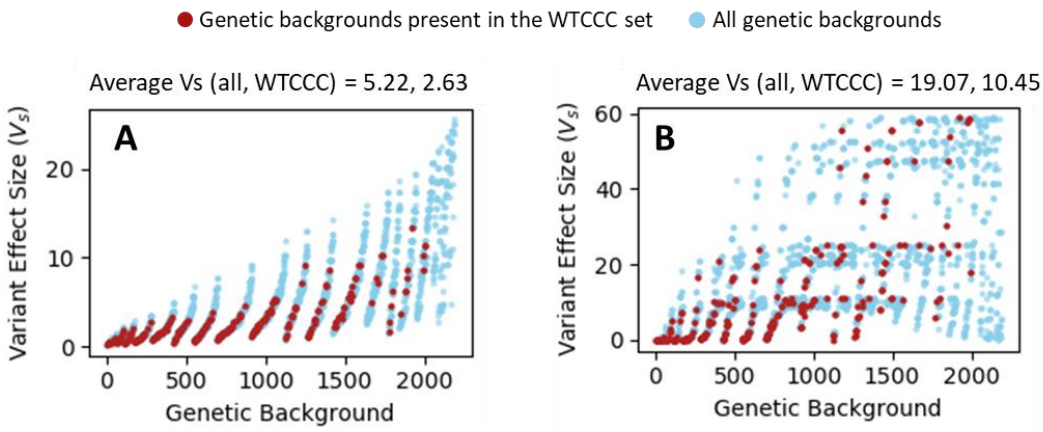


Figure 4-4. Variation in the *ATG16L1* risk variant effect size as a function of the genetic background for the barrier integrity model. Each point represents the effect size for homozygous risk variants in *ATG16L* for one specific genetic background. Figure 4-4A shows the effect size distribution for the moderately non-linear simulation model and Figure 4-4B for the more strongly non-linear model. The X-axis of each plot is sorted by the risk-allele load in the genetic background (ranging

from 0 where no other risk alleles are present to 14 where the seven other genes have homozygous risk alleles). Red points are for the 252 genetic backgrounds found in the WTCCC study population dataset and blue points cover all possible backgrounds (2187). The average effect sizes expected in GWAS are given above each plot. The effect size increases with the background risk variant load, though very non-monotonically. In the moderately non-linear model, effect sizes vary from zero to 14 in the WTCCC study population, about 1/3 of the total variation across all genotype vectors (zero to 42, Figure 4-2). For the more strongly non-linear model, the variation is from zero to 60 in the WTCCC population. In each case, the population average masks the range.

4.2.7 Epistatic effects are masked by population averaging

The size of each of the possible 28 pair-wise epistatic interactions was evaluated as a function of the genetic background of the other six risk variants. For every variant pair in the barrier integrity graph, there are 729 (3^6) possible genetic backgrounds. For a specific background, the non-additive epistatic effect of a variant pair was calculated as:

$$Epistatic\ Effect\ Size\ (E) = (B_{e_{gij}} - B_{e_{g00}}) - \{(B_{e_{g0j}} - B_{e_{g00}}) + (B_{e_{gi0}} - B_{e_{g00}})\}$$

where $B_{e_{g00}}$ is the epithelial layer bacterial concentration when the genotype of both the variants is *0* (homozygous non-risk-alleles) in a specific background g; $B_{e_{gij}}$ is the value when the genotype of both of the variants is $i=j=1$(heterozygous) or $i=j=2$ (risk allele homozygous) in a specific background g; $B_{e_{g0j}}$ is the value when the

genotype of one variant (i) is *0* and the other is *j=1* or *j=2* in the specific background *g;* and $B_{e_{gi0}}$ is the value when the genotype of variant (j) is *0* and the other is *i=1* or *i=2* in a specific background *g*. This formula captures the change in B_e when both the variants carry risk alleles versus the change in B_e assuming the contribution from each of the contributing variants is additive. A positive *E* value indicates a positive epistatic effect (i.e. the effect is higher than the additive effect), and a negative value indicates a negative epistatic effect (i.e. the effect is lower than the additive effect). To simplify the analysis, only *i=2 and j=2* were considered. For each variant pair, the set of 729 possible backgrounds was considered as well as the set of backgrounds found in a human population, using the WTCCC genotype dataset of 2000 Crohn's patients and 3000 controls (Burton et al., 2007).

Figure 4-5A shows the variation in the epistatic effect for an example risk variant pair, *ATG16L1 – NOD2,* using the moderately non-linear UPR model. For the WTCCC population, the *ATG16L1 – NOD2* epistatic effect size varies from 0 to ~13, but the average is only 4.1. Other gene pairs (Figures S7) show a similar pattern of highly variable positive epistatic effects, with population averages much lower than the effect in many individuals. As is the case in the single variant analysis (Figure 4-4), the very largest effects are seen in the highest risk genetic backgrounds, not often found in the WTCCC population. But within the population, epistatic effects are still substantial, with up to a 10-fold higher bacterial concentration at the epithelial layer above that expected from the additive model.

Figure S8 shows that the size of the epistatic effects is substantially larger in the more strongly non-linear model (up to a 50-fold change in bacterial concentration over the that expected from the linear model) and for many gene pairs may be positive or negative depending on the genetic background. The presence of both positive and negative effects leads to a greater damping of population average values. For example, for *ORMDL3 – XBP1* (Figure 4-5B) the size of the epistatic effect in the WTCCC population varies from approximately -10 to +10, resulting in an average of only 0.48. As outlined earlier, the variable nonlinear effects arise primarily from the sigmoid for the sigmoid response to the amount of unfolded protein (Figure 4-2).

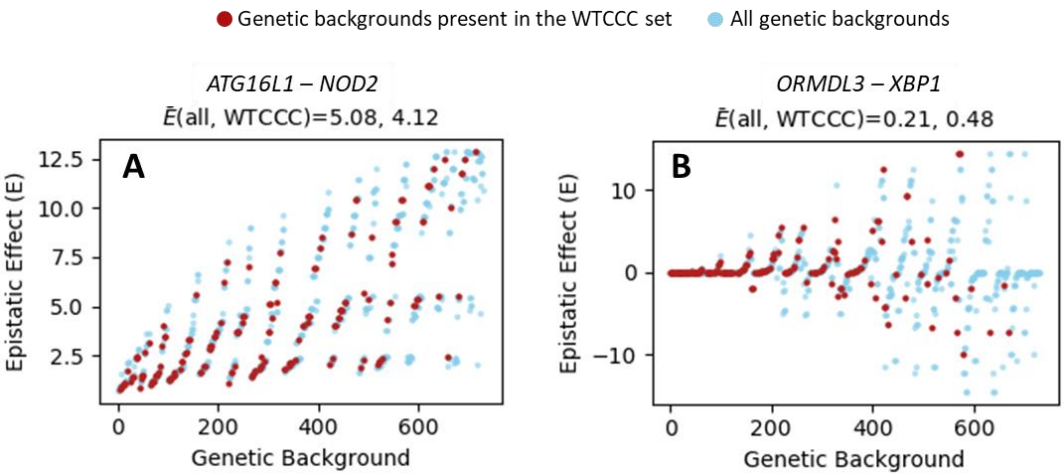


Figure 4-5. Variation in the epistatic effect as a function of the genetic background. Figure 4-5A shows the epistatic effect distribution of the *ATG16L1 – NOD2* risk variant pair in the moderately non-linear UPR model. Figure 4-5B shows the distribution for the ORMDL3 – *XBP1* risk variant pair in the more strongly non-linear model. The X-axis of each plot is sorted by the risk-allele load in the background (ranging from 0 where no other risk alleles are present to 12 where the six other genes

have homozygous risk alleles, in total 729 combinations). Red points are for genetic backgrounds found in the WTCCC study population and blue points cover all possible backgrounds. The average epistatic effect sizes are shown above the plots. The size of the epistatic effect varies dramatically as a function of the genetic background in each individual, resulting in misleadingly low population averages.

4.2.8 Constructing a Mechanism Architecture Neural Network

As outlined earlier, representing a mechanism graph as a sparse neural network should allow node functions to be learned from GWAS data, avoiding a major limitation of the circuit model. We designed and implemented a version of this approach, a Mechanism Architecture Neural Network (MANN). In this representation, a MANN is a sparse neural network where each substate perturbation and mechanism module in a mechanism graph is represented by a node and the internode connectivity is that of the mechanism graph. Figure 4-6 shows the barrier integrity MANN. Each node is represented by one neuron.

The output from each neuron is:

$$Output = BatchNorm(Tanh\left(\sum x_i w_i + b\right))$$

where $\{w_i\}$ are the weights on the 'i' edges connecting inputs to the neuron, $\{x_i\}$ are the values output from the input nodes, and *b* is the bias parameter. *Tanh* is the nonlinear transforming hyperbolic tangent activation function. The *Batch Norm* (Ioffe & Szegedy, 2015) function normalizes input values for each hidden layer in a network leading to faster convergence, decreased importance of initial weights in the network, and reduced overfitting (https://towardsdatascience.com/batch-

normalization-in-neural-networks-1ac91516821c). PyTorch (https://pytorch.org/), an open-source machine learning library, was used to implement and train the MANN. There are 71 parameters to be trained (the weights on each of 25 edges, a bias parameter and two *Batch Norm* parameters for each of the 15 internal nodes, and a bias parameter for the output node).

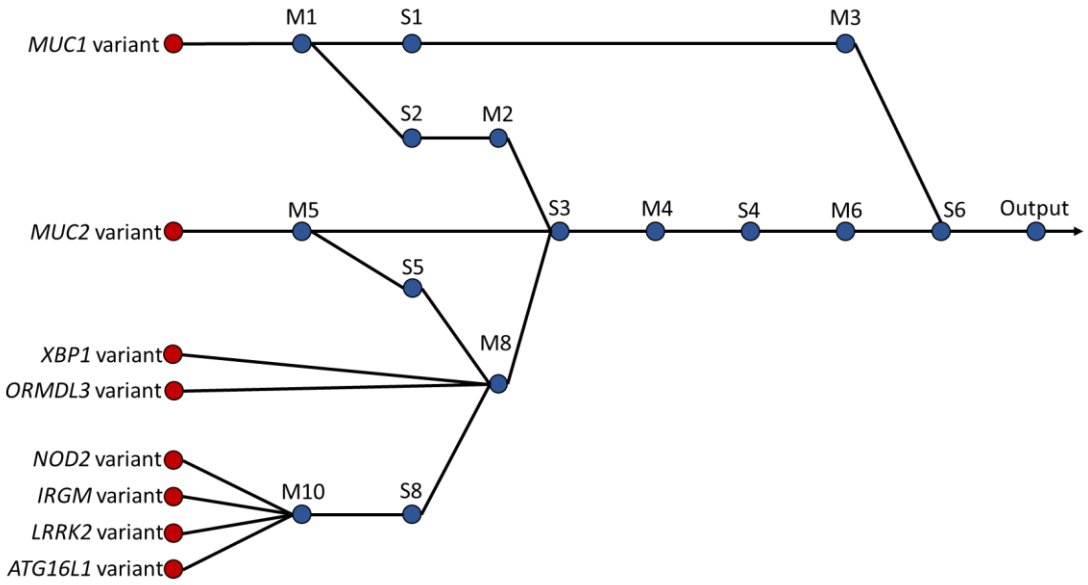
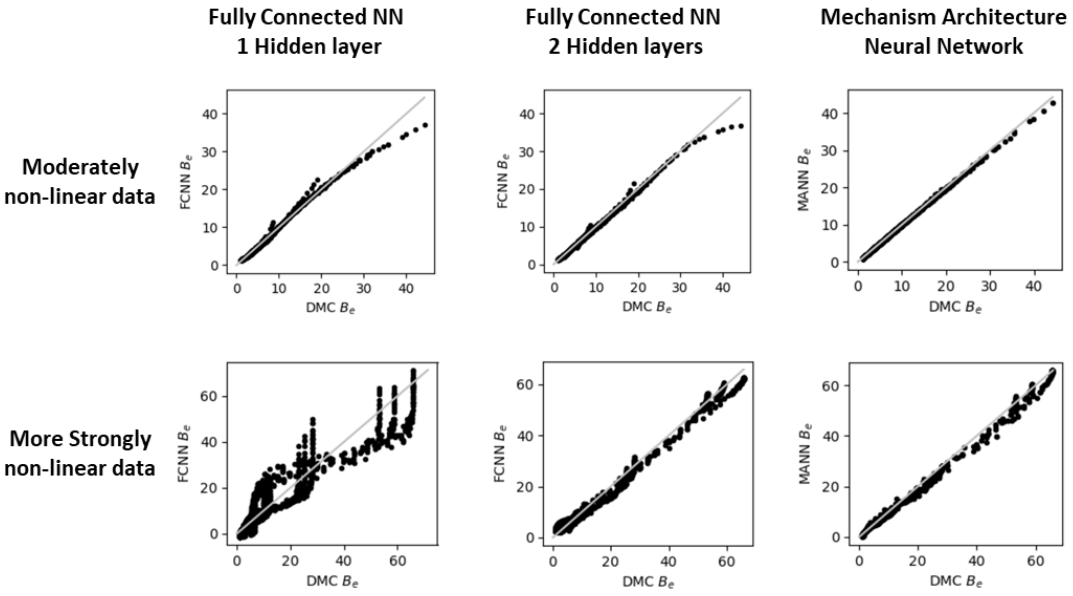


Figure 4-6. Mechanism Architecture Neural Network (MANN) for the Crohn's barrier integrity model. Each blue dot corresponds to one of the nodes in the mechanism graph (Figure 4-1) and is represented by a neuron. Red dots indicate the inputs to the network. The connectivity between the neurons is that of the mechanism graph.

4.2.9 Training and testing the Barrier Integrity MANN

The Disease Mechanism Circuit (DMC) generated B_e values corresponding to each of the input 6561 genotype vectors were used to train and test the MANN. Using the moderately non-linear UPR model data the average MSE (Mean Square Error) of the MANN is 0.13. The two equivalent fully connected FCNN reference networks have a slightly large MSE of 0.22 for the two hidden layer network and 0.38 with the single hidden layer. This pattern is the same for the more strongly non-linear data, with an MSE for the MANN of 2.17, FCNN with two layers 3.20, and FCNN with single layer, 32.22. Based on these results, the MANN is able to learn the relationship between input genotype vectors and B_e with high accuracy and its overall performance is slightly better than fully connected networks with an approximately equal number of parameters. Figure 4-7 shows a more detailed performance comparison. The moderately non-linear data plots show that the FCNNs fail to accurately predict the high B_e values (>30) possibly due to the lack of sufficient training data in this range (as shown in Figure 4-3A). But the MANN is able to fit this portion of the data as well as the rest. For the more strongly non-linear model, the plots show that a single hidden layer network is unable to properly represent the data, while the double layer network and the MANN are more successful. The fully connected double layer network likely does better here than with the moderately non-linear data because there are more training points at high B_e values. Overall, the MANN delivers the lowest MSEs, but these data are generally easy to train.

We also trained and tested the networks using the experimental WTCCC dataset (Burton et al., 2007) and disease status (1 for disease, 0 for control) instead of simulated bacterial concentration as the output. The MANN delivers an average ROC (Receiver Operating Characteristics) - AUC (Area Under The Curve) of 0.63, slightly better than the FCNN with two layers (0.59). Thus, the MANN is able to learn the relationship between an individual's genotype data and their disease status slightly better than a conventional network. Apparently, with the noisy experimental data, explicitly incorporating knowledge of the interactions between the genetic variants results in an improved fit between the genotype-phenotype data. Interestingly, including 86 risk variants with a fully connected neural network (one hidden layer with three neurons, 265 parameters) yields a ROC-AUC of only 0.72, and attempts by a number of groups to obtain better results with a range of machine learning methods and up to 160 risk variants did not succeed in obtaining a larger AUC (Daneshjou et al., 2017). In this context, the AUC of 0.63 for just the eight barrier integrity variants



is remarkable.

Figure 4-7. Comparison of MANN performance in reproducing the output bacterial concentration B_e with that of fully connected neural networks, for a sample cross-validation set. Each point corresponds to a specific input genotype vector. A perfect model would have all points on the diagonal. For both the moderate and more strongly non-linear data, the MANN performs better than the fully connected networks. FCNN: Fully connected Neural Network, DMC: Disease mechanism circuit, MSE: Mean square error, MANN: Mechanism Architecture Neural Network.

4.2.10 The Barrier Integrity MANN is able to learn node functions

While it is useful that a MANN representation can capture the relationship between genotype inputs and the bacterial concentration, B_e, and does so more robustly than a conventional neural network (Figure 4-7), the real advantage should be in also capturing aspects of a biological mechanism circuit that may not be known. To examine this, we next ask to what extent individual nodes in the trained Barrier Integrity MANN exhibit the same response to genotype inputs as the corresponding nodes in the DMC. That is, can the network effectively learn DMC node functions from the data? For this purpose, the output from each neuron representing a node is compared with the calculated output from the corresponding DMC node, for each input genotype vector. Figure 4-8 shows the results for the 10 nodes where the output is a physical quantity, for example, total protein production, P (F7), and the thickness of the mucosal layer, T (F11) (Table 4-1). For the five nodes in the top row of the plots, there are only three input values corresponding to the three possible risk

genotypes of *MUC1* (nodes M1, M2, and M3) or *MUC2* (M5 and S5), and so only three points of comparison. Nevertheless, the node functions of two of these, M2 and M3, are exponentials so that capturing their behavior is non-trivial. For the five nodes in the second row of the plots, the consequences of combinations of input genotype vectors from multiple genes must be captured, so reproduction by the neural network is generally demanding. For example, the output from node M8, the total protein production 'P', is determined by the sigmoid function representing the unfolded protein response (the UPR), and input genotype combinations from six genes. The output behavior of this node is well reproduced, with correlations coefficients of 0.99 and 0.96. The relatively poor correlation coefficients for nodes S3 and S4 (down to 0.92) are a consequence of the single output from S3 providing the single input to S4. Although these two nodes represent distinct physical quantities (total mucosal protein abundance 'M'' for S3 and mucosal layer thickness 'T'' for S4) the simple coupling between them allows multiple combinations of node performance to satisfy the final output requirement. With this exception, the output node functions do faithfully capture the disease circuit behavior.

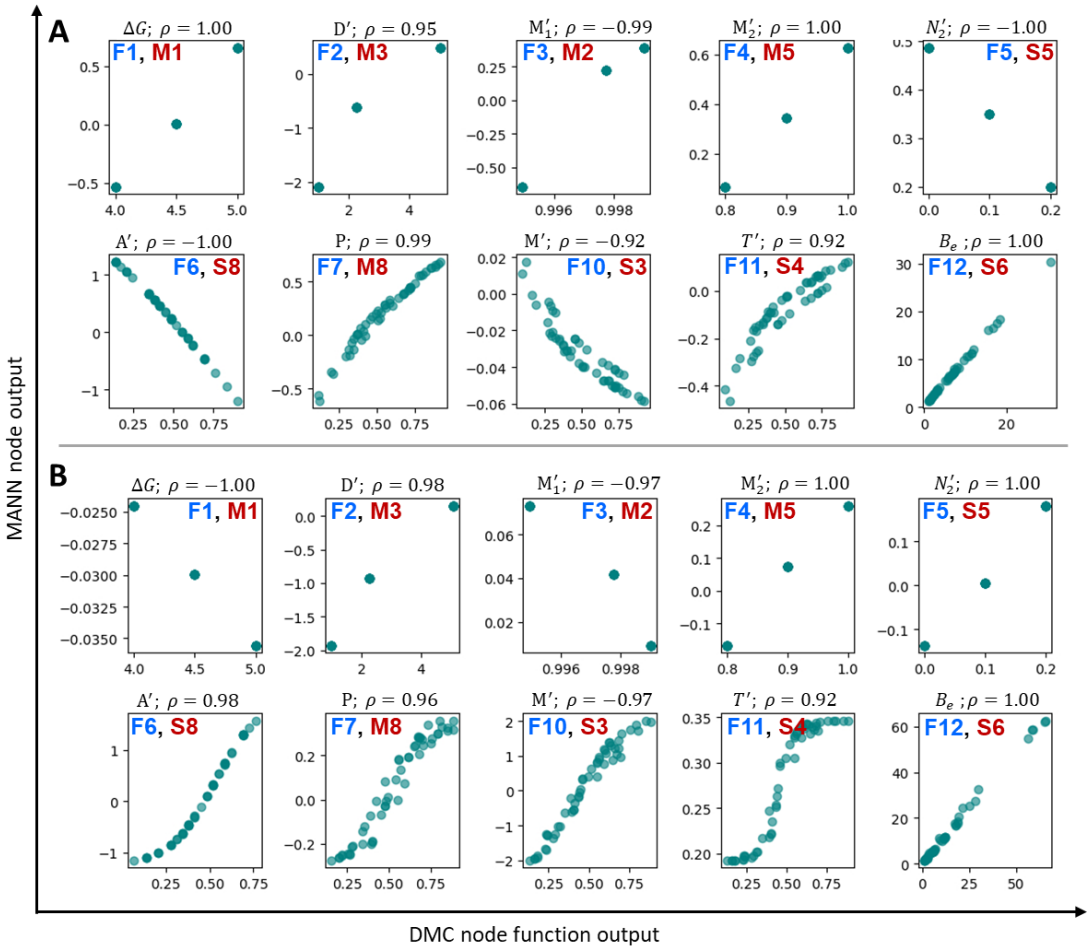


Figure 4-8. Relationship between output from nodes in the Barrier Integrity MANN and the model on which it was trained (model node functions are listed in Table 4-1). Each point represents a node output generated by a specific input genotype vector. ρ is the Pearson's correlation coefficient. F numbers refer to the functions in Table 4-1 and M and S to the nodes in the Disease Mechanism Circuit (figure 4-1) and the MANN (figure 4-6). The upper panel shows the results for the moderately non-linear model and the lower for the more strongly non-linear one. Results for a single bootstrap are shown, those for other bootstraps are similar. A MANN that perfectly

captures a disease circuit's node properties would result in all points on the diagonal and a correlation coefficient of 1.0. The worse correlation coefficient is 0.92.

4.2.11 The Barrier Integrity MANN is able to distinguish between alternative mechanism graph topologies

In practice, the connectivity of a disease mechanism graph is often uncertain - there may be alternative hypotheses for particular steps in a mechanism and some aspects of the mechanism may be speculative or missing. As a result, mechanism graphs may have incorrect connections and missing connections or components. To evaluate whether a MANN can be used to distinguish between possible alternative topologies of this type, 300 alternate versions of the barrier integrity MANN were constructed. Each alternate topology was set up such that the input genotype vector to the network had the genotypes arranged in a different order than the one used for the original barrier integrity MANN. Changing the order feeds the input to the wrong nodes in the network. This is equivalent to creating and dropping an equal number of edges in the network. 300 randomly ordered input vectors were created to represent the alternate topologies. Each of the randomly ordered vectors was used to train the MANN using the protocol described above, on both the moderately and more strongly non-linear models. The average mean square error (MSE) for the test sets was used as a metric for choosing between topologies: In principle, the network with the correct topology should have the lowest MSE between the model and MANN generated bacterial concentrations, $\{B_e\}$.

Figure 4-9A shows that the topologies trained on the moderately non-linear model have a very narrow range of MSE relative to the range of B_e values (1 to 43, Figure 4-2). The correct barrier integrity topology is ranked 20th among other topologies. Figure 4-9B shows a wider range of the MSE when the topologies are trained on the more strongly non-linear data. The correct barrier integrity topology model is ranked 1st in this set with a MSE of 0.73. This result suggests that as non-linearity in the system increases, a MANN's power to distinguish the correct topology improves.

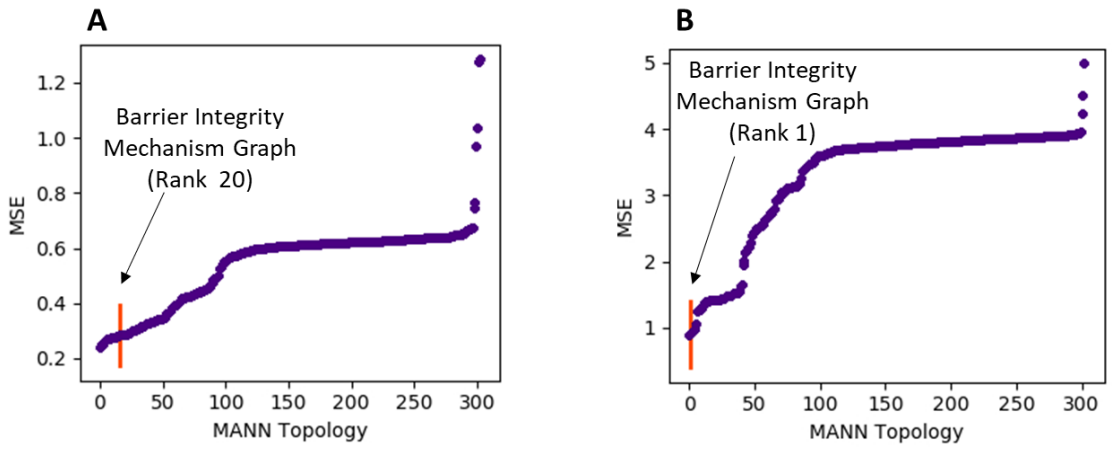


Figure 4-9. Ranking of alternative MANN topologies by performance. MSE: Mean Square Error. MANN: Mechanism Architecture Neural Network. Each point represents the MSE for a specific topology. The red bar shows the position of the correct barrier integrity mechanism graph topology. Figure 4-9A shows the performance of MANN topologies trained on the moderately non-linear model. Figure 4-9B shows the performance on the more strongly non-linear simulation data. Topologies are sorted by MSE. The figures show that the MANN is usually able to distinguish the correct barrier integrity topology model from alternatives.

4.3 Discussion

4.3.1 Summary of the results

This chapter presented a framework to build a quantitative model of mechanism for complex trait disease. The model was generated from a MecCog mechanism graph that qualitatively represents the mechanisms by which genetic variants cause disease phenotypes as a result of perturbation propagation across stages of biological organization. For the model, an explicit analytical function and parameters were added to each node in the mechanism graph, representing its physical properties. The node functions, together with the mechanism graph connectivity generate a circuit where output quantitative phenotype values can be computed from a given set of input disease-associated genetic variants. We used this Disease Mechanism Circuit to investigate the role of gut barrier integrity in Crohn's disease. The circuit allowed us to address key questions concerning the interpretation of Genome-Wide Association Study (GWAS) data, specifically, the extent to which the effect size of a variant contributing to disease risk varies as a function of the genetic background in an individual and the extent to which averaging over a GWAS population masks epistatic effects between pairs of variants. We also implemented a hybrid neural network approach, including prior knowledge in a neural network. For this, we constructed a mechanism architecture neural network (MANN) by integrating the topology of the mechanism graph into the architecture of the neural network. We showed that a MANN can reproduce the relationship between input genotype vectors and the output barrier integrity phenotype as effectively as a fully connected neural network using simulated data and that is more effective than a fully connected

network using experimental genotype case-control data obtained from a large scale GWAS for Crohn's disease. We also showed that the MANN can learn individual complex node behavior from the data, and that it can also be used to address uncertainties in graph topologies. These results are for model systems but suggest a number of practical applications. Two of these are outlined below.

4.3.2 Evaluating the effectiveness of drug targets

Genetic background has been shown to play a role in the effectiveness of existing drugs (Madian et al., 2012; Mallah et al., 2020). A drug can be regarded as altering the function of a node in the mechanism graph for a disease, for example, by inhibiting the action of a protein. Our results suggest the MANN framework can be used to identify whether a drug will be effective for a particular individual, given their genetic variants. One possible case where this may be applied is for Mongersen (Sands et al., 2020), an antisense SMAD7 inhibitor. A clinical study found that only 65% of Crohn's patients respond to the drug (Monteleone et al., 2015), and given the large variation among risk variants in Crohn's patients, at least some of that response variation is likely genetic. Mongersen acts to increase the anti-inflammatory TGB-beta signaling by decreasing the abundance of SMAD7 protein, found to be elevated in Crohn's patients (Monteleone et al., 2012). However, there are many other genes with Crohn's risk variants (such as αvβ8 Integrin, *STAT3*, *TLR4*, *SMAD2*, *SMAD3*) that are part of the anti-inflammatory TGB-beta signaling pathway. A MANN representing the mechanism graph that links the risk variants of Mongersen relevant genes to the disease phenotype can be trained using the existing case/control data.

Then in each individual, the activity of the SMAD7 node is reduced, mimicking the action of the drug. The effect of this change on predicted disease risk should then provide an estimate of drug effectiveness.

4.3.3 Coarse grain MANN

Complex trait disease mechanism graphs can often be decomposed into semi-autonomous subprocesses. For example, eleven subprocesses have been proposed for Crohn's disease (Jostins et al., 2012). In each individual, the extent to which a subprocess is affected will vary. For example, the contribution of the barrier integrity subprocess circuit to overall disease risk is represented by the concentration of bacteria at the gut epithelial cell wall, B_e. Our analysis shows this likely varies substantially from individual to individual, as a function of the eight associated risk variants. In choosing appropriate treatments it can be valuable to know which subprocesses are most involved in a patient. To this end, we propose the design of a *coarse-grain MANN* to investigate the contribution of subprocesses to complex trait disease risk. Figure 4-10 illustrates this concept.

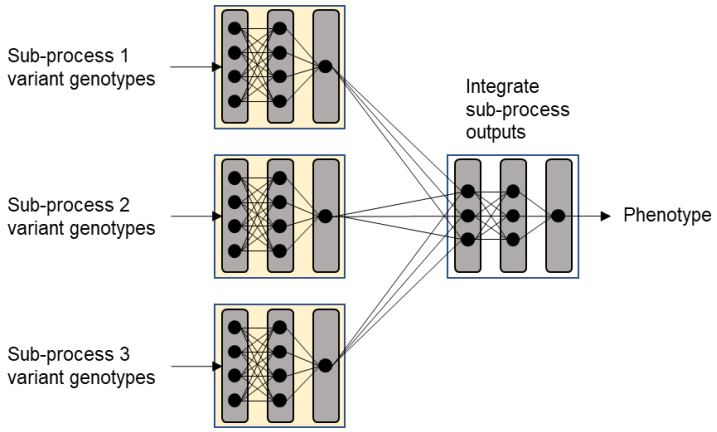


Figure 4-10: Example of a coarse grain MANN architecture with three subprocesses. Each dot represents a neuron. Each subprocess is represented by a local neural network, with input genotype vectors. Outputs from these networks are feed into a neural network that produces a final phenotype value.

The full network has two stages. The first stage has separate networks for each subprocess considered, three in this example. Each of these subnetworks has a conventional architecture, with one input node for each risk variant affecting that subprocess. An assumption here is that these subnetworks will be able to capture the non-linear interactions between variants within a subprocess. Outputs from the subnetworks are then fed into the second stage integrating network. This second stage network models the interactions between the subprocesses to output disease risk or another disease-related phenotype. Measuring the output from each subprocess network for an individual's input genotype vector will provide an estimate of the contribution of each subprocess to the disease phenotype.

4.4 Methods

4.4.1 Node functions and parameter values

Quantitative encoding of the barrier integrity mechanism graph was done by adding analytical functions for each substate perturbation and mechanism module in the graph. Functions are intended to capture the physics and chemistry represented by each node. Table 4-1 shows the node functions and parameter values of the barrier integrity model.

The node functions were defined as follows:

M1: Decreased MUC1-MUC1 interactions

The presence of *MUC1* risk allele results in reduced glycosylation of the MUC1 protein. In turn, that results in weaker interactions with other mucins and with the other MUC1 molecules (Hall et al., 2017b). The interaction strength is expressed in terms of the free energy change of binding, ΔG, a negative number in units kcal/mol. The change in free energy resulting from the presence of risk variants is $\Delta\Delta G$, a positive quantity. With this assumption, $\Delta G$is computed as:

$$\Delta G = a_1 * \{1 - a_2 * G(MUC1)\}$$

where G(MUC1) is the number of *MUC1* risk variants (0, 1, or 2) and a_1 and a_2 are constants. The unperturbed value of ΔG is assumed to be 5 kcal/mol, so a_1= 5. a_2 is set to 0.10, so that each risk allele reduces the protein-protein interaction free energy by 10%.

M3: Increased bacterial diffusion rate

Weaker interactions between MUC1 protein molecules results in a higher fraction of the molecules dissociated from each other. K_d, the dissociation constant is related to the free energy of association between two MUC1 molecules by $K_d \propto e^{-\Delta G/RT}$. Assuming the rate of bacterial diffusion coefficient, *D*, is proportional to the fraction of unassociated MUC1 molecules, normalized D' can be expressed as:

$$D' \propto K_d \propto e^{-\Delta G/RT}$$

$$D' = a_3 * e^{\frac{-\Delta G}{RT}}$$

where R is gas constant (1.985 X 10^{-3} kcal K^{-1} mol^{-1}), T is body temperature (310.50 K), and a_3 is constant. a_3 is chosen so that $D' = 1$ when $\Delta G = 5$.

M2: Increased Loss of MUC1 protein at gut surface

MUC1 protein molecules normally become detached from the mucosal layer as a result of peristaltic motion in the gut (Paone & Cani, 2020). Weaker interactions between the mucin molecules will result in a greater rate of detachment. k_{off} represents the rate at which mucin molecules become detached. We assume that k_{off} depends on the change in the free energy of association between two MUC1 molecules carrying the risk variants:

$$k_{off} \propto e^{-\Delta G/RT}$$

Thus, at steady state, the normalized abundance of MUC1 can be expressed as –

$$M_1' = 1 - a_4 * e^{\frac{-\Delta G}{RT}}$$

where a_4 is constant. For a non-risk allele status of MUC1, $M_1' = 0.9$ and $\Delta G = 5$.

M5: Altered MUC2 protein internal interactions and S5: Increased misfolded MUC2

The MUC2 risk alleles are associate with decreased stability of the MUC2 protein (Heazlewood et al., 2008; Moehle et al., 2006), assumed to result in lower protein abundance. The normalized abundance of misfolded MUC2, N_2' is assumed to be proportional to the number of risk alleles (0, 1, or 2):

$$N_2' = a_5 * G(MUC2)$$

where a_5 is a constant. The normalized abundance of folded MUC2 protein is

$$M_2' = 1 - a_5 * G(MUC2)$$

M10: Decreased autophagy

As described earlier, risk alleles associated with four genes (*ATG16L1*, *NOD2*, *IRGM*, *LRRK2*) are expected to affect the efficiency of autophagy in removing unfolded/misfolded proteins. The functional dependency of autophagy and the interdependence between risk alleles for different genes is not known. A simple linear sum over the risk alleles is used, with separate weights for each gene. For ATG16L1 the effect of risk variants is accentuated by a positive feedback loop - the presence of risk alleles increases the rate at which the protein is subject to proteolytic cleavage (Kaser & Blumberg, 2014). The rate of cleavage also depends on the extent of cell stress, and the lower the abundance of ATG16L1, the greater cell stress because of decreased autophagy. The impact of this loop is modeled by a parameter *S* that increases the co-efficient for ATG16L1.

$$A' = a_6 * \{a_7 * G(NOD2) + a_8 * G(IRGM) + a_9 * G(LRRK2) + a_{10} * G(ATG16L1) * S)\}$$

where A' is the relative abundance of unfolded protein due to impaired autophagy (zero for no risk allele in these genes), a_7, a_8, a_9, a_{10} are the weight coefficients of the for each gene. a_6 is such that when all risk alleles are present A$= 0.9$.

M8: Increased unfolded protein response

As described earlier, the unfolded protein response (UPR) is a mechanism by which a cell shuts down protein production if too much unfolded/misfolded protein accumulates. The URP is a threshold response – below some level of the

unfolded/misfolded protein there is no UPR, above that threshold, complete shut-down of protein synthesis is triggered until conditions improve. The total protein production rate of the cell is modeled using a sigmoid function:

$$P = a_{11} + \frac{a_{12}}{1 + e^{-u(a_{13} * U - a_{14} * U_0')}}$$

where U is the accumulated unfolded/misfolded protein, U_0' is an offset of the UPR so there is no reduction in protein production at low levels of U, u determines the steepness of the transition, and $a_{11}, a_{12}, a_{13}, a_{14}$ are constants.

The threshold U_0' is affected by risk variants in two UPR related genes, XBP1 and ORMDL3. We assume that these reduce the threshold for the onset of the UPR as a linear function of the risk variant genotypes:

$$U_0' = 1 - \{a_{18} * G(XBP1) + a_{19} * G(ORMDL3)\} + a_{15}$$

where $a_{15,} a_{18}, a_{19}$ are constants.

The value of *U* is calculated as the sum of the abundance of the MUC2 misfolded protein (N_2') and the abundance of the unfolded proteins arising from impaired autophagy (A').

S3: Decreased mucin abundance and S4: Thinner mucosal layer

The normalized total mucin abundance (M') is determined by the overall relative rate of protein production, *P*, and the abundance of the principal mucin proteins, MUC1 (M_1') and MUC2 (M_2'). The normalized thickness (T') of the mucosal layer is assumed to be directly proportional to the total mucin abundance:

$$M' = P * (a_{20} * M_1' + a_{21} * M_2')$$

$$T' = \boldsymbol{a_{22}} * M$$

where $a_{20}, a_{21} and\ a_{22}$ are constants. As MUC2 is the secreted form of mucin and dominates the composition of the mucosal layer (Johansson et al., 2011), a higher weight ($a_{21} = 0.9$) is assigned to it compared to MUC1 ($a_{20} = 0.1$).

S6: Higher bacterial concentration at the epithelial wall

This node describes the perturbation of the concentration of bacteria at the epithelial cell wall arising from a thinner and less dense mucosal layer. The simplest model of this process is the diffusion of bacteria across a planer mucosal layer. According to this model, the influx of bacteria arriving at the epithelial cell wall is proportional to the diffusion constant D' (output from node M3) and inversely proportional to the mucosal layer thickness T' (output from node S4). Thus, we assume that at steady state, the bacterial concentration at the epithelial cell wall (B_e) can be expressed as:

$$B_e = a_{23}\frac{D'}{T'}$$

where a_{23} is a constant.

Table 4-1. Analytical functions and parameters in the barrier integrity mechanism circuit.

Node	Physical Property	Function Number	Analytical Function	Parameter Values
M1: Decreased MUC1-MUC1 interactions	Change in free energy of binding ($\Delta G'$)	F1	$\Delta G = a_1 * \{1 - a_2 * G(MUC1)\}$	a_1 =5 a_2 =0.10
M3: Increased bacterial diffusion rate	Diffusion coefficient (D')	F2	$D' = a_3 * e^{\frac{-\Delta G}{RT}}$	R = 0.00198 T = 310.50

				a_3 =3323.55
M2: Increased Loss of MUC1 protein at gut surface	MUC1 protein abundance (M_1')	F3	$M_1' = 1 - a_4 * e^{\frac{-\Delta G}{RT}}$	a_4 =332.35
M5: Altered protein internal interaction	MUC2 protein abundance (M_2')	F4	$M_2' = 1 - a_5 * G(MUC2)$	a_5 =0.10
S5: Increased misfolded MUC2	MUC2 misfolded protein abundance (N_2)	F5	$N_2' = a_5 * G(MUC2)$	a_5 =0.10
M10: Decreased autophagy	Total unfolded protein abundance (A)	F6	$A' = a_6 * \{a_7 * G(NOD2) + a_8 * G(IRGM) + a_9 * G(LRRK2) + a_{10} * G(ATG16L1) * S)\}$	a_6 = 0.28 a_7 = 0.50 a_8 = 0.25 a_9 = 0.25 a_{10} = 0.25 S = 2.50
M8: Increased unfolded protein response	Total Protein production rate (P)	F7	$P = a_{11} + \frac{a_{12}}{1 + e^{-u(a_{13} * U - a_{14} * U_0')}}$	a_{11} = 0.10 a_{12} = 0.90 a_{13} = 3.00 a_{14} = 3.00 u = (-1, -6)
	Total abundance of unfolded and misfolded protein (U)	F8	$U = a_{16} * N_2' + a_{17} * A'$	a_{16} = 1.00 a_{17} = 1.80
	Unfolded protein response threshold (U_0')	F9	$U_0' = 1 - \{a_{18} * G(XBP1) + a_{19} * G(ORMDL3)\} + a_{15}$	a_{18} = 0.05 a_{19} = 0.05 a_{15} = 0.017
S3: Decreased	Total mucin abundance (M')	F10	$M' = P * (a_{20} * M_1' + a_{21} * M_2')$	a_{20} = 0.10 a_{21} = 0.90

mucin abundance				
S4: Thinner mucosal layer	Mucosal layer thickness (T')	F11	$T' = a_{22} * M'$	a_{22} = 1.00
S6: Higher bacterial concentration at the epithelial wall	Bacterial concentration at epithelial cell wall (B_e)	F12	$B_e = a_{23} * \frac{D'}{T'}$	a_{23} = 1.00

$G(Y)$ represents the genotype of the risk variant affecting gene 'Y'. For example, $G(MUC1)$ represents the genotype of the risk variant associated with *MUC1*.

4.4.2 Training and testing the Barrier Integrity MANN

Two types of datasets were used to train and test the MANN: First, the Disease Mechanism Circuit (DMC) generated B_e values corresponding to each of the input 6561 genotype vectors. Second, the experimental WTCCC genotype dataset (Burton et al., 2007) of 2000 Crohn's cases and 2000 controls. For each of these datasets, the train-test divisions were generated by randomly dividing the dataset into a 75%-25% ratio using python's scikit-learn *train_test_split* module (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html). For cross-validation, 30 random train-test sets were generated. PyTorch modules were used to train the network using the Mean Square Error (MSE) loss function, the ADAM stochastic gradient descent algorithm as the optimizer, and a mini-batch size of 500. To evaluate the performance, the average of the mean square error (MSE) for the DMC dataset and ROC-AUC for the WTCCC dataset over the test portions of 30 cross-validations was used. As controls for the DMC dataset, two versions of a fully

connected neural network (FCNN) were also trained. One is a single hidden layer (5 neurons) network (61 parameters), and the other a two hidden layer network (4 neurons in each layer) network (77 parameters). As a control for the WTCCC dataset, a two hidden layer network (4 neurons in each layer) network (77 parameters) was trained. The node configuration, training, and evaluation procedures for the fully connected networks were the same as used for the MANN.

4.5 Acknowledgments

This work was supported in part by the National Institute of Health [R01GM104436 to JM]. We thank Yizhou Yin, Lipika Ray, Lindley Darden for many useful suggestions and comments.

Chapter 5: Conclusion

In this dissertation, I developed computational methods to identify disease-causing variants from sequencing data, qualitatively represent the mechanism by which these variants cause the disease phenotype, and quantitatively encode the mechanism representations to analyze emergent properties in complex trait diseases. In the last chapter, I briefly summarize the conclusions of each project and discuss future directions in each area.

5.1 Interpreting causative variants in DNA sequencing tests

In the first part of my dissertation, I developed a variant prioritization pipeline *VarP* to address an incompletely solved problem on building accurate computational methods for variant interpretation that can be used in DNA sequencing tests to provide the clinical diagnosis. We objectively assessed *VarP* on data from the John Hopkins DNA Diagnostic Laboratory as part of a CAGI gene panel challenge. The challenge assessment revealed that *VarP* performed the best among 17 other submissions, and correctly matched 36 out of 106 patients to one of the 14 monogenic disease classes, using sequence data for the 83 panel genes. The correctly matched set included 10 cases where the Hopkins pipeline could not find causative variants, but VarP was able to. We then investigated the incorrectly matched cases and found 17 where *VarP* did find a causative variant and 53 undiagnosed cases where neither *VarP* nor the Hopkins pipeline could find any causative variant relevant to the tested disease class. Investigating the missed diagnosed cases revealed several sub-optimal features of *VarP* but one of the major contributing factors was placing too much trust

on variant pathogenicity annotations in the Human Gene Mutation Database (HGMD). In post-analysis, we showed that omitting the HGMD annotations increases the success rate of correctly matching disease class from 36 to 40. We then investigated sequencing artifacts in the data to check if these contributed to the high number of undiagnosed cases. We found that generally, the data was of high quality with rare artifacts such as zero coverage for exon-60 of the *HYDIN* gene for 78 samples and an abnormally high homozygous/heterozygous ratio in one sample. These did not appear to make a large contribution to the missing diagnostic variants. Re-analyzing the incorrectly matched cases revealed seven patients where *VarP* found high confidence pathogenic variants in genes associated with a different disease class from the one referred for testing at the Hopkins laboratory. We speculate that these may be cases of incorrect clinical diagnosis. Lastly, we integrated protein structure data to the *VarP* pipeline to investigate if mechanistic insights could be derived for the putative causative variants. We found that ~50% of the missense variants with unknown clinical significance had structure coverage and analyzing the structures revealed potential molecular mechanisms of the variants.

5.1.1 Improving genetic disease diagnosis

The Hopkins laboratory dataset revealed that the clinical diagnosis for ~50% of the cases could not be confirmed because neither the Hopkins pipeline nor any of the CAGI methods found causative variants in the tested genes. Others (Clark et al., 2018) have also reported that rare disease pipelines in general have a success rate

below 50%. There are several possible explanations for the low yield of diagnostic variants:

A) The disease-causing gene list is incomplete as evidenced by the fact that many new genes are continually being discovered (Posey et al., 2019) and cataloged in crowdsourced resources such as the Genomic England PanelApp (Martin et al., 2019). As a result, if causative variants are in genes that are not part of the panel sequencing gene list, they will be missed. For such undiagnosed cases, follow-up whole-exome sequencing can be performed to identify a larger set of genes hosting possible causative variants and investigating if any of these genes are related to the disease phenotype. Whole exome sequencing (WES) has helped to resolve cases of newborns with a severe combined immunodeficiency disease phenotype that were not solved using panel sequencing (Chan, Punwani, Kadlecek, Cowan, Olson, Mathes, Sunderam, Fu, et al., 2016; Mallott et al., 2013; Patel et al., 2015b). However, expanding the gene list for variant analysis can also be challenging in terms of time to analyze the data and chances of increasing false positives.

B) The impact of coding variants might not be accurately estimated by the general *in silico* predictors (such as PolyPhen (Adzhubei et al., 2013), SIFT (Kumar et al., 2009), or CADD (Smedley et al., 2016)). For example, in chapter 2 we showed a case of a missense variant (C603S in NR3C2 protein) where two out of four general predictors estimated it to be deleterious indicating a 50% chance of it being pathogenic. But analysis of protein structure revealed it to be a strong case of a protein destabilizing mutation. Similarly, variants disrupting existing splicing regulatory sequences, creating new ones, or activating the cryptic ones can be

overlooked by the general predictors. To this end, building advanced gene-specific predictors or mechanism-specific predictors that make use of the specialized features can be beneficial. In the recent CAGI, a gene-specific ensemble variant interpretation method (Yin et al., 2017a) and a sequence feature based-splicing impact prediction method (Cheng et al., 2019) have shown promising results.

C) Causative variants can be in the non-coding region of a patient's genome. But these variants are often not screened and if they are, they are ignored in the analysis because accurately interpreting their impact on the disease phenotype is difficult. As GWAS data shows that most disease-associated variants are mapped to the non-coding region of the genome, it suggests that regulator elements can have a prominent role in genetic diseases (Farh et al., 2015; Hindorff et al., 2009). Given the recent availability of the regulatory data in GTEx (Lonsdale et al., 2013) and ENCODE (Dunham et al., 2012), steps can be taken to compile accurate and comprehensive annotations for the regions of the genome relevant for disease phenotypes. These high-quality annotations can then be integrated into the variant interpretation pipelines as well being used to develop impact prediction in silico tools for non-coding variants.

D) It may be possible that the DNA sequencing approach is adequate for diagnosing specific types of genetic diseases such as inborn errors of metabolism (IEM), and that biomarker-based approaches can yield better diagnostic power. A recent study (Adhikari et al., 2020) showed that WES is less effective in newborn screening for inborn metabolic disorders compared to the established tandem mass spectrometry

approach. Therefore, a strategy where the two approaches are used in conjugation may help to reduce the false positive and negative rates in diagnostic tests.

5.1.2 Standardizing evidence of pathogenicity for causative variants

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have proposed different types of evidence that can be used to assess the impact of the clinically relevant variants. These include variant segregation data, functional data, population data, and computational predictive data (Richards et al., 2015b). Despite this remarkable effort to standardize evidence assessment, several issues with its use have been recognized: A) Older clinically relevant databases like HGMD are not using these guidelines; instead it uses its evidence-based classification system that has resulted in erroneous labeling of disease-causing variants as shown in chapter 2 and by others (Cassa et al., 2013); B) The ClinVar database uses the guidelines, but a recent study (Shah et al., 2018) shows instability in the classification system, where a large fraction of variants previously labeled pathogenic become reclassified to uncertain significance and conflicting interpretation; C) Others (Nykamp et al., 2017) have been developing new guidelines based on the ACMG-AMP guidelines. This lack of convergence for evidence of pathogenicity is leading to an unclear definition of "known pathogenic" variants (van Rooij et al., 2020). We propose that a possible solution for evidence assessment can be implemented in two steps: 1) Developing an expert-sourced *evidence of pathogenicity ontology* that will contain a comprehensive collection of all the relevant evidence types that can help to interpret a variant's impact. Variants can

be tagged with appropriate ontology terms based on the available evidence. Such an idea has been implemented in the Gene Ontology community, where the Evidence and Conclusion Ontology (ECO) (Chibucos et al., 2017) has been developed to describe the evidence supporting gene annotations. 2) Developing an expert-sourced variant impact classification scheme that makes use of the evidence types described in the above ontology. Software for this can be distributed as executable for easy integration into existing clinically relevant resources.

5.2 Systems-level representation of mechanisms by which genetic variants cause disease phenotypes

In the second part of my dissertation, I co-developed the theory of a graphical representation framework for genetic disease mechanisms called MecCog, based on concepts in the philosophy of biology and computational biology. Next, I implemented it as a web-based platform to manually build integrated systems-level representations (a.k.a., mechanism schemas) of mechanisms by which a genetic variant causes a disease phenotype. The MecCog platform facilitates the integration of mechanism information in terms of perturbation propagation across stages of biological organization, evaluation of the evidence related to that information, and identification of the uncertainties, ambiguities, and ignorance in that information. The platform provides functionalities to create, store, browse, and search schemas. I have designed graphical notations and curated ontology-informed class terms so as to consistently and intuitively represent the types of mechanism components found in schemas. The schema visualizer in the platform is interactive and tightly integrates

the graphics, text, and hyperlinks to evidence sources. We have built MecCog schemas to describe mechanisms in monogenic disease (cystic fibrosis), cancer (Lynch syndrome), and complex trait disease (Crohn's disease). We have also tested and shown that the MecCog formalism can be used to create mechanism graphs by combining multiple schemas to describe interactions between genetic variants. Lastly, I have equipped the MecCog platform with features to support expert-sourcing for schema construction.

5.2.1 Scaling mechanism schemas in MecCog

The manual construction of the MecCog schemas relies on human understanding to extract and infer causal connections between mechanism components from the literature. The scattered and incomplete nature of the mechanistic information in the literature makes this process complex and requires a combination of prior biological knowledge together with searching for, assimilating, and assessing new facts and evidence from the literature. This makes schema construction labor-intensive. To achieve scale, some degree of automation in the construction process is needed. A combination of the following strategies can help in this regard:

A) A lot of biological information is already represented in a structured format such as pathways (KEGG (M Kanehisa & Goto, 2000), Reactome (Fabregat et al., 2017)), networks (STRING (Szklarczyk et al., 2018)), relational databases (UniProt (Bateman et al., 2017), PDB (Burley et al., 2017)), and knowledge graphs (Hetionet (Daniel Scott Himmelstein et al., 2017)). It is possible to extract the information from these resources in the form of triplets of subject-predicate-object (SPO). As the elementary

SSP-MM-SSP units of a schema are a subset of SPO triplets, parts of the schema can be automatically generated this way. Nevertheless, an assessment needs to be done as to what extent these extracted triplets represent the disease mechanism space, and devise strategies to shortlist high confidence and relevant triplets for a schema. B) To mine biological information from literature, BioCreative (Critical Assessment of Information Extraction systems in Biology) has identified a series of Natural Language Processing (NLP)-based tools (http://biocreative.sourceforge.net/bionlp_tools_links.html) for the bio-entity recognition and relationship extraction tasks. Using these tools, databases of SPO triplets from PubMed such as SemMedDB (https://skr3.nlm.nih.gov/SemMedDB) (Kilicoglu et al., 2012) are being generated. Such NLP-based tools and databases can also be explored to retrieve connections between mechanism components in a schema. However, a potential caveat of these automatically generated triplets can be that they will miss out on the biological knowledge whose information is spread out across paragraphs in literature and so, would need a reasoning strategy to piece them together for creating the triplets. C) One of the critical activities while building a schema is to find and analyze all the evidence of a particular component to assess its relevance for the disease mechanism description. A first and important step for this is to find all the papers that discuss the component in the context of the disease phenotype. Unsupervised machine learning techniques like Topic modeling (http://machinelearningtext.pbworks.com/w/file/fetch/47924743/BleiLafferty2009.pdf) can be helpful in this case, inferring the latent topical structure of a collection of documents and identifying word-groups or expressions that best characterize the

document set. This technique can be used to automatically identify the set of PubMed articles that are characterized by the words relating directly or indirectly to the mechanism component.

5.3 Quantitative representation of mechanisms relating genetic variants and complex trait disease

In the third part of my dissertation, I developed a framework to quantitatively represent and analyze the mechanism described in the MecCog mechanism graph format. Using a subsystem on the role of gut barrier integrity in Crohn's disease, we demonstrated that a mechanism graph can be transformed into a computable circuit by assigning appropriate node functions and parameters that represent the behavior of the graph components. We also showed that the degree of non-linearity in the circuit can be altered by the choice of parameter values. We then investigated the use of the circuit to get insights for interpreting GWAS data which otherwise would require a demanding experimental setup. The circuit showed: A) The effect size of a GWAS risk variant can change drastically as the function of the specific risk allele-load in the genetic background; B) The size of the epistatic effect between pairs of GWAS risk variants can get diluted by averaging the effects across genetic backgrounds. Next, we demonstrated that the node functions and parameters can be learned in a data-driven manner using a Mechanism Architecture Neural Network (MANN) – a sparse neural network wired based on the topology of the mechanism graph. We also showed MANNs can be used to address uncertainties in mechanism graph topologies.

5.3.1 Complex trait disease risk assessment using MANN

MANN provides a novel way to quantitatively encode a systems-level representation of mechanism by which genetic variants cause disease phenotype. The inclusion of prior biological knowledge in the form of the neural network topology provides a number of practical advantages such as a decreased number of training parameters compared to a fully connected neural network, a need for fewer training data, more interpretable, and easy testing of multiple topologies. In addition, as these sparse neural networks can also facilitate analysis of how a genotypic profile can influence disease phenotype, they can be a useful system to study many diagnostic and therapeutic aspects of complex trait diseases. One of these areas is risk prediction. Most often, the genetic risk of an individual is assessed through the polygenic risk score (PRS), a weighted sum of the number of risk alleles an individual carries (C. M. Lewis & Vassos, 2020). This summing across variants strategy in PRS assumes an additive genetic architecture of complex trait diseases, with independence of risk variants. It does not consider possible nonlinear interactions between variants that can influence the disease phenotype. A few studies have compared PRS with machine learning (ML) algorithms that can model non-linearity (such as support vector machines, random forests, gradient boost) and interestingly found that PRS performs better than the ML algorithms (Gola et al., 2020; Vivian-Griffiths et al., 2019). Going ahead, a subject for investigation is performance comparison between PRS and MANN.

Appendix

Varant

Varant is an open-source genetic variant annotation tool (written in Python, http://compbio.berkeley.edu/proj/varant). Varant provides five categories of annotation based on 17 data sources: variant identity and frequency, experimentally-defined genomic features, predicted genomic features, variant/gene phenotypes, and prediction of mutation impact. It has been used in a number of clinical research studies (Chan, Punwani, Kadlecek, Cowan, Olson, Mathes, Sunderam, Man Fu, et al., 2016; Patel et al., 2015a; Punwani et al., 2016).

QC Analysis Results

Supp. Fig. S2A shows that the Ts/Tv ratios for all the samples are clustered between 2.2 and 3.2. For the human genome based on 1000Genomes data, Wang et al. 2015 have shown that the Ts/Tv ratio is ~3 for SNVs inside exons and ~ 2 elsewhere. Since the capture regions cover more than just exons (1350 exonic and 39 intronic regions for 83 genes), the Ts/Tv ratio for SNVs is expected to lie between 2 and 3, consistent with the plot and the pattern is very similar for the 1000 Genomes samples. The Het/Hom ratios for all samples except one (P8) are clustered between 1.1 and 2.6. P8 is an outlier and carries more homozygous SNPs. It is established that on a genome-scale, the Het/Hom ratio is close to 1.5 (McKernan et al., 2009; Schuster et al., 2010) but it also depends on whether a population incorporates recent admixture (skewing towards heterozygosity) or inbreeding (skewing towards homozygosity). The

Het/Hom ratio range here is similar to that of the 1000Genomes sample. Thus, by these measures, the data, with the exception of P8, are of good quality.

Supp. Fig. S2B shows that for the capture v01 samples have 1200 to 2000 low quality and 200 to 940 no call sites. The captures v02 samples have 230 to 330 low quality and 90 to 180 no call sites. We expect that if any causative variant falls on one of these no call or low-quality sites it will be completely missed.

Supp. Fig. S2C shows that the number of common SNVs per sample cluster between 232 and 312 for the 96 samples sequenced using Capture v01 and between 132 and 175 for the 10 samples sequenced using Capture v02, consistent with Capture v02 covering 19 fewer genes than Capture v01. Non-African samples have a lower rare variant load (ranging from 8 to 33) than the African samples (ranging from 19 to 80). This pattern is also seen in the samples from 1000Genomes samples and has been previously reported in the literature (Durbin et al., 2010; Zawistowski et al., 2014). The novel (i.e. not found in 1000 Genomes and ExAC) SNV count is in the range of 0 to 8 with a median of 1 per sample and is similar to the count observed in the 1000 Genomes dataset where the range is 0 to 12 with a median of 2 per sample. These novel variants be present in the patient's family or be de novo in the patient but it is not possible to distinguish these two situations given only the patient's variant data.

Supp. Fig. S2D shows that for the 96 Capture v01 samples, the common Indel count is between 5 and 16 whereas for the 10 Capture v02 samples the count is 3 to 6 per sample. We observe that the distributions of the rare Indel is between 0 and 12 and novel Indel is between 0 and 4. Two African samples (P2 and P83) are identified as outliers carrying more rare Indels compared to rest of the Hopkins samples and 1000

Genomes dataset. However, the Indel counts are similar to the 1000 Genomes dataset. By all these measures, the Hopkins data appears to be of high quality.

Supp. Fig. S3 shows the distribution of average read depth for 83 genes across all samples. Average read depth varies more substantially across the 106 samples (horizontal variation) than across the 83 genes (vertical). Genes not included (blue boxes) in the 10 capture v02 samples are evident. Though there is variation in the gene coverage across samples, from 107X to 983X, even the lowest coverage should be adequate for diagnostic analysis and confirmatory testing as shown by Strom et al. 2014.

We identified nine capture regions (occurring in eight genes) with anomalous read depth in more than 90 of the 106 samples (Supp. Table S1). One of these has very high coverage and the others have low coverage. Two of these regions lie in the major isoform of one gene, HYDIN, one high (Exon 53) and one low (Exon 60). Exon 53 has greater than 600X coverage in 70 samples (Supp. Fig. S4). Exon 60 has no coverage in 78 samples and less than 20X coverage in three more samples. The other anomalous regions are unlikely to affect downstream analysis because they are either deep intron, present in a minor isoform of the gene or the actual coverage of the region is at least 100X in most of the samples.

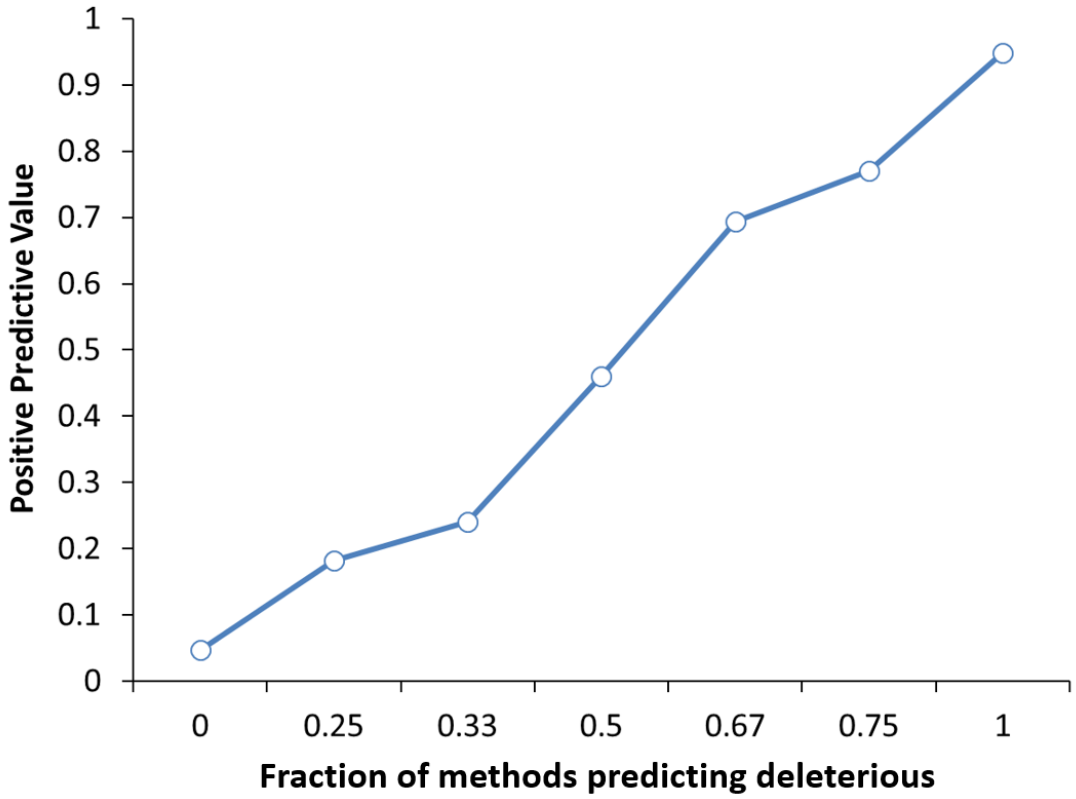


Figure S1. Relationship between the fraction of methods that agree on a deleterious assignment for variants and the positive predictive value, PPV (fraction of predicted deleterious variants that are deleterious), for 10695 HGMD missense mutations and 10240 interspecies variants with available predictions for at least two out of the four methods (SNPs3D Profile, SIFT, Polyphen2 and CADD). By this measure, 77% of variants for which at least 3 of 4 methods predict deleterious are in fact deleterious.

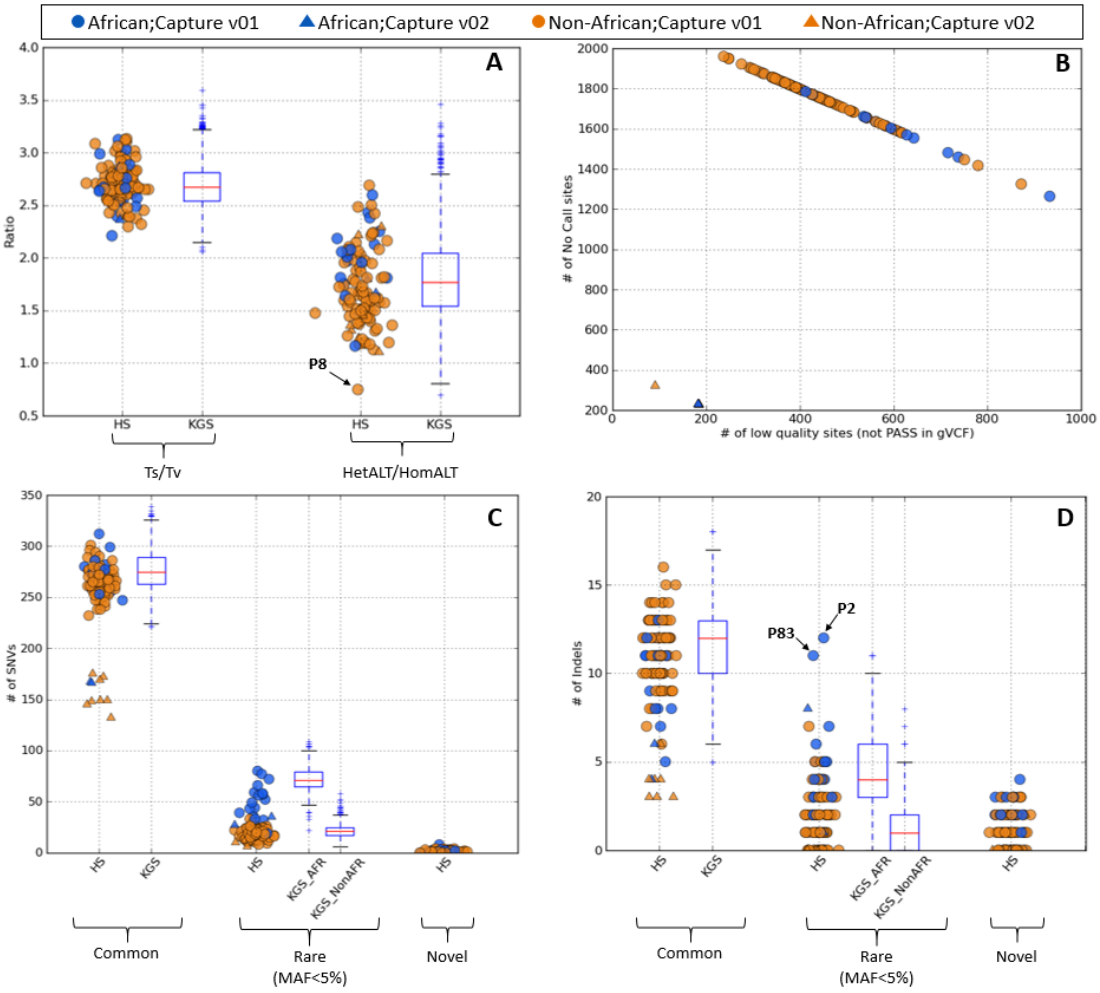


Figure S2. Comparison of variant calling quality for 106 Hopkins samples versus 2,504 1000Genomes samples across the 83 genes in the panel. Only high-quality calls are included. HS: Hopkins Samples, KGS: 1000 Genomes samples, KGS_AFR: African samples in 1000Genomes, KGS_NonAFR: Non-African samples in 1000 Genomes. Circles represent HS sequenced using Capture v01 and triangles represent the HS sequenced using Capture v02. African samples are blue, Non-African are brown. Figure S2A shows the distribution of Transition vs. Transversion (Ts/Tv) and Heterozygous SNVs vs. Homozygous SNVs (HetALT/HomALT). By both measures, HS and KGS data are similar, except for the for HetALT/HomALT ratio of sample

P8, an outlier with an excess of homozygous SNVs. Figure S2B shows the distribution of no call sites versus low-quality sites (not PASS in the gVCF file). Causative variants falling on any these sites will probably be missed. Figure S2C shows the distribution for common, rare and novel SNV types. For the common and rare SNV types, the HS and KGS distributions are similar. The lower counts of HS common SNVs for capture 2 reflects the fact that only 64 genes are included. There is a similar effect for rare variants, obscured by the crowding of points (medium 20 counts for capture v01 and 15 for v02). The rare variant distribution for both HS and KGS reflects the fact that Non-African samples have a lower rare variant load than the African samples. The novel variants load (between 0 and 8 variant with a median of 1 per sample) in HS is much lower than the rare variants. Figure S2D shows the distribution for common, rare and novel Indels. The distribution of common and rare Indels in HS is similar to the KGS distribution. Like SNVs, the lower counts of HS common indels for capture 2 is evident, in the plot. Two African samples (P2 and P83) are seen to carry more rare Indels and are outliers compared to the 1000Genomes dataset. For six samples there is a slightly higher number of rare than novel Indel in HS.

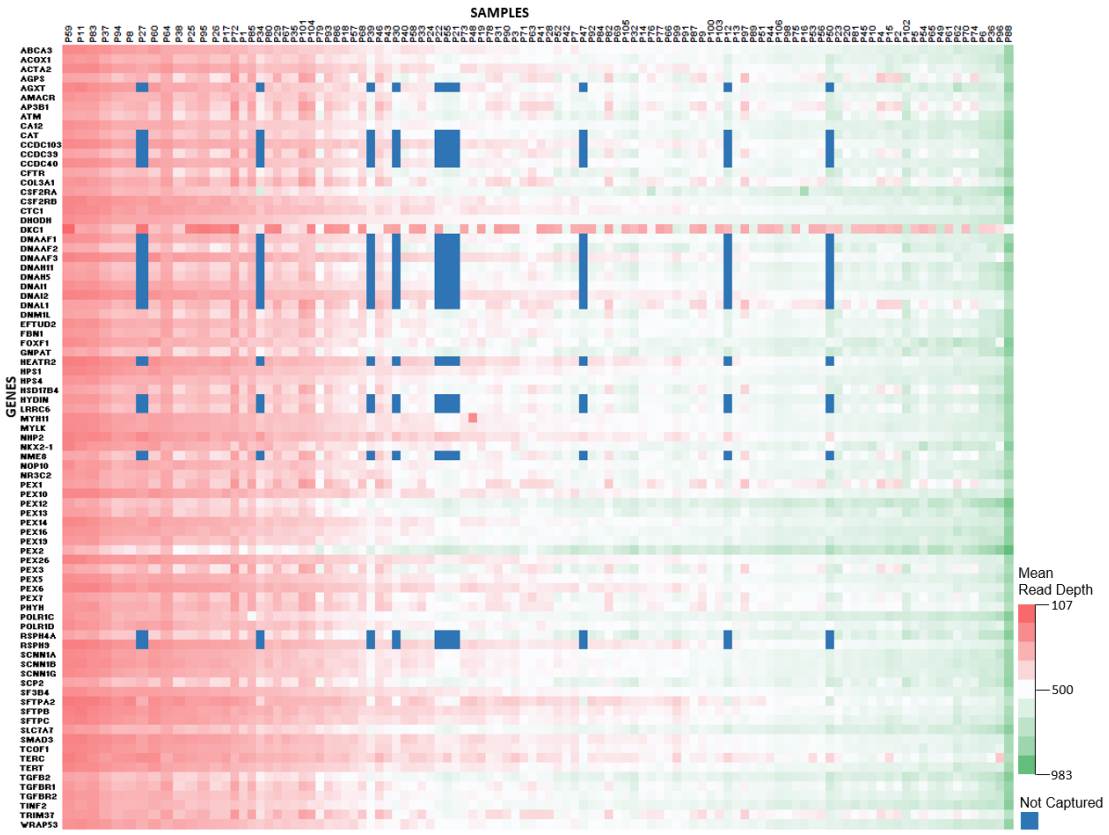


Figure S3. Heat-map of average read depth for 83 genes across the 106 samples, color-coded red (low depth: ~100) to green (high: ~950). Blue indicates the corresponding gene was not captured. Each column is for a different sample, and there are 83 rows, one for each gene. It is evident that coverage varies substantially across samples, from a low of 107X to a high of 983X.

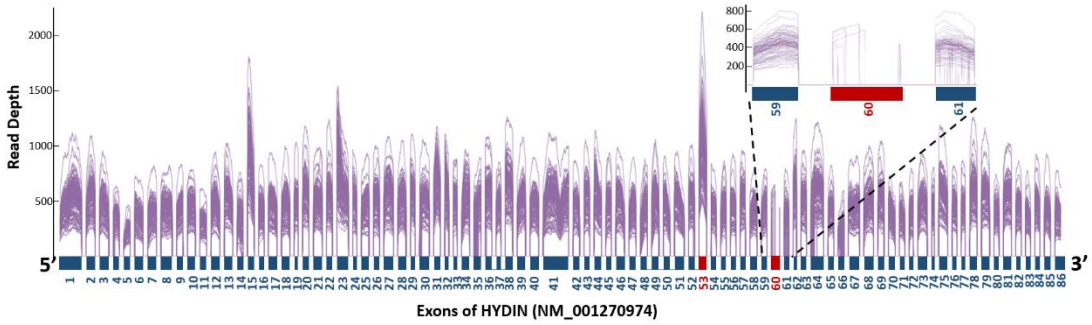


Figure S4. Exon-wise read depth for the HYDIN gene. Each purple line represents one sample and each rectangle represents one exon. The red rectangle indicates exons with anomalous coverage. The plot shows that Exon 53 has very high coverage and Exon-60 has very low coverage or no coverage for many samples compared to other exons in the gene. The inset shows a zoomed-in view of the read depth for Exon 59, Exon 60 and Exon 61.

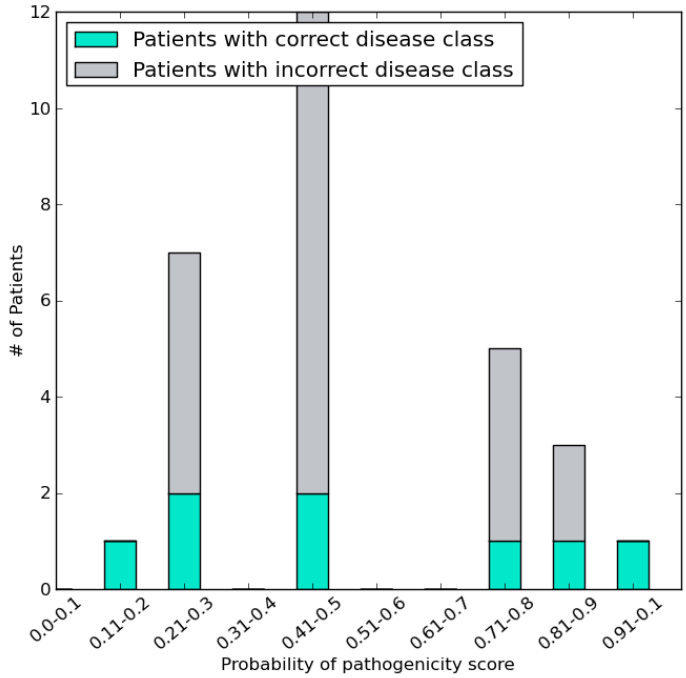


Figure S5. Distribution of correct and incorrect assignments of pathogenicity for patients based on missense mutations, as a function of the assigned probability of pathogenicity.

Table S1. The nine regions with anomalous average read depth observed in more than 90 of the 106 samples. The eight "LOW COVERAGE" capture regions have low average read depth compared to other regions in the same gene. The one "HIGH COVERAGE" capture region has high read depth compared to other regions in the same gene. The "# of Samples" column is subdivided into coverage bins from no coverage to high coverage for "LOW COVERAGE" regions and from high to very high for "HIGH COVERAGE" regions. The HYDIN gene has two anomalous capture regions – 1. Low coverage, no reads in 78 samples and 2. A very high coverage of > 600X in 70 samples. The other anomalous regions are either deep intron, or present only in a minor isoform or actual coverage of the region is at least greater than or equal to 50X in all the samples.

	Gene	Capture Region	Genomic Region	Min. Mean Coverage in Samples	Max. Mean Coverage in Samples	# of Samples				
						No Coverage	1X – 20X	20X – 50X	50X – 100X	>100X
LOW COVERAGE	CCDC39	chr3:180369900-180370104	Exon (CDS)	60.27	381.93	-	-	-	8	88
	CSF2RA	chrX:1422103-1422305	Exon (CDS) in minor isoform	15.0	149.45	-	8	78	12	8
	DNM1L	chr12:32832247-32832449	First Exon of CDS	63.37	363.59	-	-	-	2	104
	CTC1	chr17:8151271-8151404	First Exon of CDS	68.18	314.30	-	-	-	7	99
	HYDIN	chr16:71021776-71022086	Exon (CDS)	0.0	262.83	78	3	5	6	4
	HSD17B4	chr5:118831376-118831558	Intron	0.0	172.08	9	1	3	64	29
	TERT	chr5:1294835-1295154	First Exon of CDS	55.67	270.49	-	-	-	8	98
	FBN1	chr15:48787269-48787507	Exon (CDS)	114.01	411.15	-	-	-	-	106
HIGH COVERAGE	Gene	Capture Region	Genomic Region	Min. Mean Coverage in *Samples	Max. Mean Coverage in *Samples	# of Samples				
						100X-300X	300X – 600X	600X – 900X	900X – 1300X	>1300X
	HYDIN	chr16:71007681-71007973	Exon (CDS)	357.29	1594.32	-	26	48	20	2

Table S2. Five cases where two pairs of Indels in the CCDC40 gene were selected to satisfy a compound heterozygous model and leading to incorrect disease assignments. Each pair of Indels is very close to each other suggesting possible false variants arising from realignment errors or errors near repeat regions in the genome.

Patient ID	Variant Position (chrom: position)	Reference Allele	Alternate Allele	Variant Type
P31	chr17: 78064120	ACGCGCG	A	Deletion
	chr17: 78064128	AGGCACGTGCACGAACAAGGGACG	A	Deletion
P58	chr17: 78064144	A	ACC	Insertion
	chr17: 78064145	A	AC	Insertion
P61	chr17: 78064002	GC	G	Deletion
	chr17: 78064004	ACGTGCACGAAGAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCGGGC	A	Deletion
P91 & P66	chr17: 78063996	ACGCAGGCACGTGCACGAAGAACACGGGACGCG	A	Deletion
	chr17: 78064052:	CGGGACGCGCGCGGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCACGAACAA	C	Deletion

Table S3. Number of distinct variants that led to disease class prediction in 106 patients. 105 distinct potentially causative variants occurred only once in 78 patients. 14 potentially causative variants occurred twice or more in the remaining 28 patients. AD: Autosomal Dominant, HR: Homozygous Recessive, and CH: Compound Heterozygous.

# of distinct variant that led to classification	# of times seen in patients	Occurred as AD or HR		Occurred as part of CH pair	
		# of patients with correct disease class	# of patients with incorrect disease class	# of patients with correct disease class	# of patients with incorrect disease class
105	1	17	29	14	18
11	2	2	12		6
1	3		3		
1	4	3		1	
1	6		4		2

Table S4. Percentage of correct disease assignments in each of the three variant selection categories after removing HGMD from the method. Accuracy increases compared to the pipeline with HGMD. Overall trends remain the same - as expected, accuracy is highest in Category-1, then Category-2, then Category-3., and novel variant assignments are more accurate than for rare variants.

Category	Variant Considered	Minor Allele Frequency			% Correct Assignment
		Novel	<= 0.005	<=0.01	
Category-1	In ClinVar with Pathogenic or Likely pathogenic tag	1/1	4/7	0/1	5/9: 55%
Category-2	Missense (Predicted damaging either by SNPs3D, SIFT, PolyPhen2 or CADD) Frameshift / Non-Frameshift Indel NonSense Direct Splicing Any variant predicted damaging by dbscSNVs	11/16	13/39	3/7	27/62: 43%
Category-3	All other missense, UTR, and Intronic	5/18	2/13	1/2	8/33: 24%
		17/35: 49%	19/59: 32%	4/10: 40%	

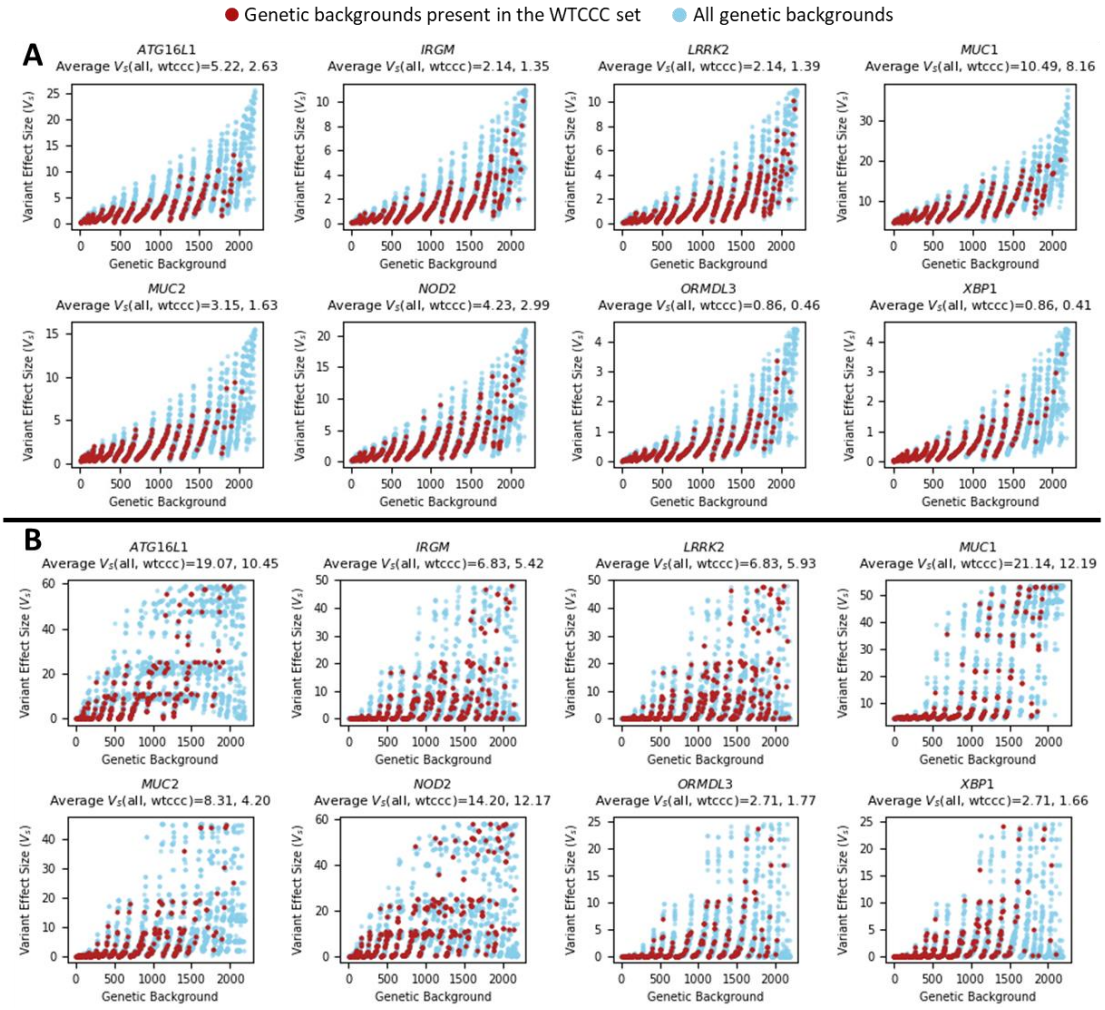


Figure S6. Variation in risk variant effect size as a function of the genetic background for the barrier integrity model for all genes. Each panel shows the effect size distribution of a specific risk variant against genetic backgrounds. Figure 4A shows the effect size distributions for the moderately non-linear simulation model and Figure 4B for the more strongly non-linear model. The X-axis of each plot is sorted by the risk-allele load in the genetic background (ranging from 0 where no other risk alleles are present to 14 where the seven other genes have homozygous risk alleles, Red points are for the genetic backgrounds found in the WTCCC study population dataset and blue points cover all possible backgrounds (2187).

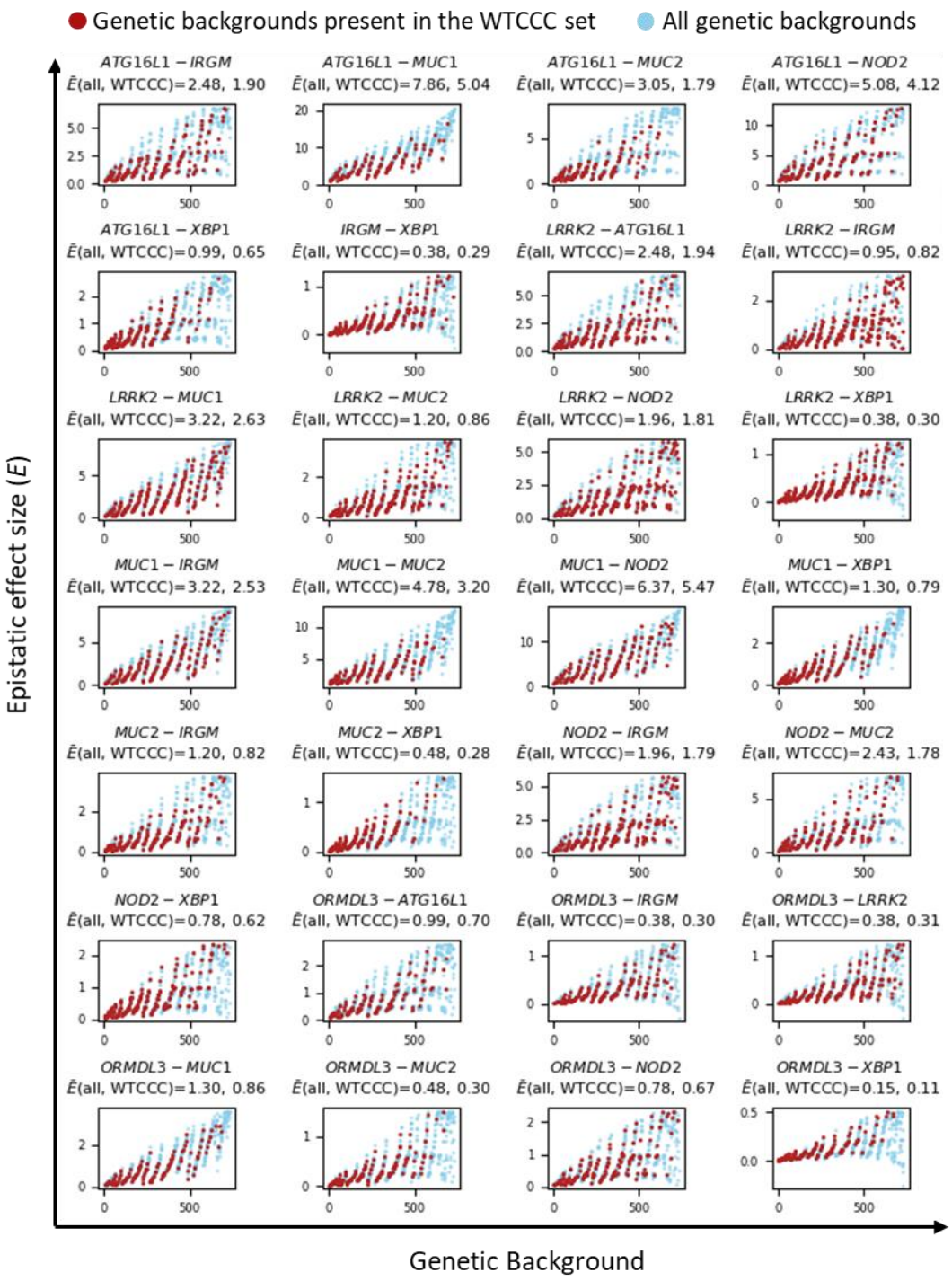


Figure S7. Variation in the size of the epistatic effect size as a function of the genetic background in the moderately non-linear barrier integrity model for all gene pairs.

Each panel shows the epistatic effect size distribution for a variant pair against genetic backgrounds. The X-axis of each plot is sorted by the risk-allele load in the background (ranging from 0 where no other risk alleles are present to 12 where the six other genes have homozygous risk alleles, in a total of 729 combinations). The load was calculated as the linear sum of the genotype values in the background. Red points are for genetic backgrounds found in the WTCCC study population dataset and blue points cover all possible backgrounds. The average epistatic effect size over all backgrounds and the WTCCC backgrounds is shown above the plots.

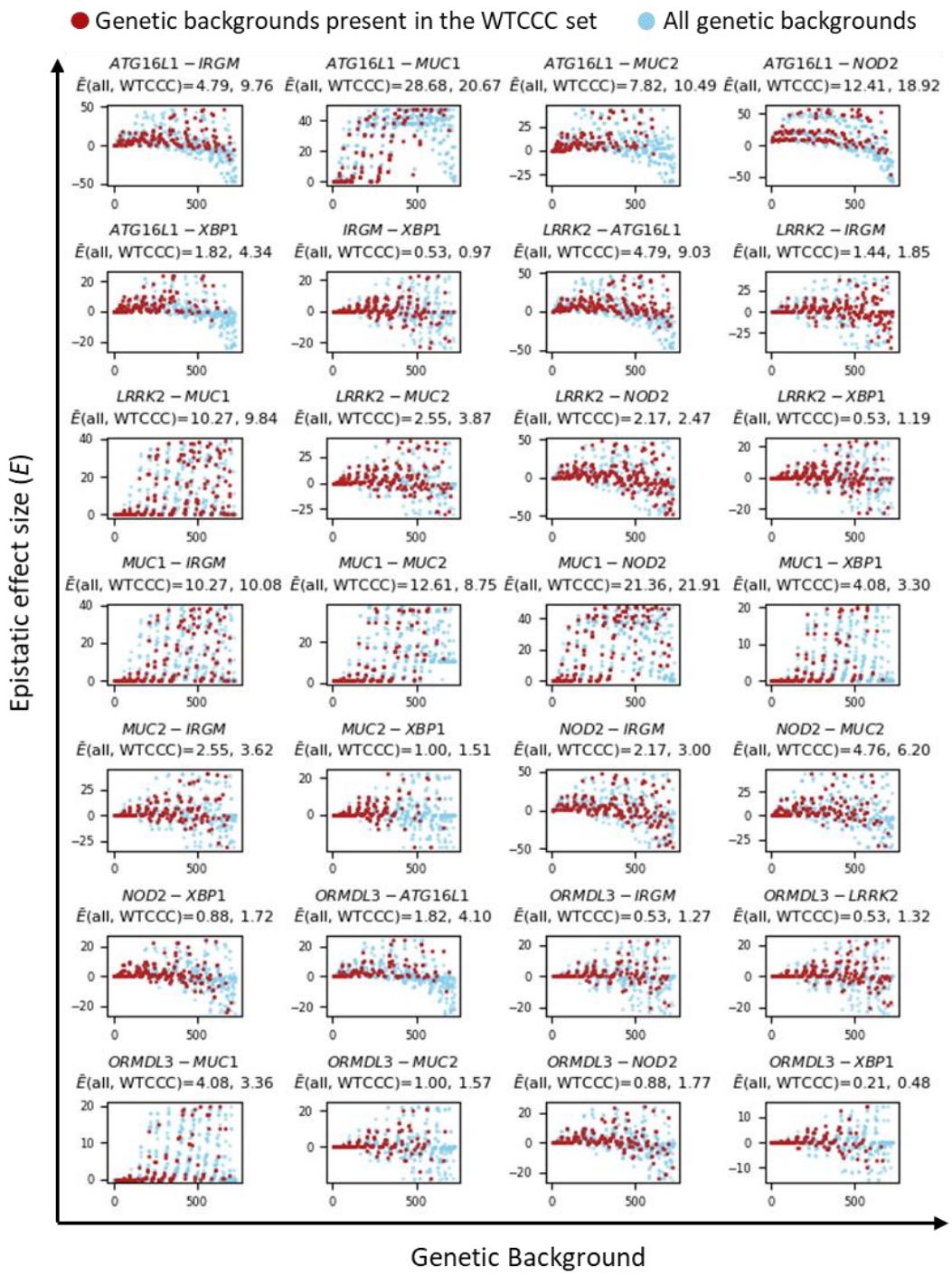


Figure S8. Variation in the size of the epistatic effect size as a function of the genetic background in the more strongly non-linear barrier integrity model for all gene pairs.

Each panel shows the epistatic effect size distribution for a variant pair against genetic backgrounds. The X-axis of each plot is sorted by the risk-allele load in the background (ranging from 0 where no other risk alleles are present to 12 where the six other genes have homozygous risk alleles, in a total of 729 combinations). The load was calculated as the linear sum of the genotype values in the background. Red points are for genetic backgrounds found in the WTCCC study population dataset and blue points cover all possible backgrounds. The average epistatic effect size over all backgrounds and the WTCCC backgrounds is shown above the plots.

Bibliography

Acharya, S., Wilson, T., Gradia, S., Kane, M. F., Guerrette, S., Marsischky, G. T., Kolodner, R., & Fishel, R. (1996). hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13629–13634. https://doi.org/10.1073/pnas.93.24.13629

Adhikari, A. N., Gallagher, R. C., Wang, Y., Currier, R. J., Amatuni, G., Bassaganyas, L., Chen, F., Kundu, K., Kvale, M., Mooney, S. D., Nussbaum, R. L., Randi, S. S., Sanford, J., Shieh, J. T., Srinivasan, R., Sunderam, U., Tang, H., Vaka, D., Zou, Y., … Brenner, S. E. (2020). The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nature Medicine*, *26*(9), 1392–1397. https://doi.org/10.1038/s41591-020-0966-5

Adolph, T. E., Niederreiter, L., Blumberg, R. S., & Kaser, A. (2012). Endoplasmic reticulum stress and inflammation. *Digestive Diseases*, *30*(4), 341–346. https://doi.org/10.1159/000338121

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In *Current Protocols in Human Genetics* (pp. 7.20.1-7.20.41). John Wiley & Sons, Inc. https://doi.org/10.1002/0471142905.hg0720s76

Ahluwalia, B., Moraes, L., Magnusson, M. K., & Öhman, L. (2018). Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies. In *Scandinavian Journal of Gastroenterology* (Vol. 53, Issue 4, pp. 379–389). Taylor and Francis Ltd.

https://doi.org/10.1080/00365521.2018.1447597

Al Nabhani, Z., Dietrich, G., Hugot, J. P., & Barreau, F. (2017). Nod2: The intestinal gate keeper. In *PLoS Pathogens* (Vol. 13, Issue 3). Public Library of Science. https://doi.org/10.1371/journal.ppat.1006177

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J. H., Yang, J., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. https://doi.org/10.1038/nature09410

Allot, A., Peng, Y., Wei, C. H., Lee, K., Phan, L., & Lu, Z. (2018). LitVar: A semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Research*, *46*(W1), W530–W536. https://doi.org/10.1093/nar/gky355

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151

Andersen, M. E., Yang, R. S. H., French, C. T., Chubb, L. S., & Dennison, J. E. (2002). Molecular circuits, biological switches, and nonlinear dose-response relationships. In *Environmental Health Perspectives* (Vol. 110, Issue SUPPL. 6, pp. 971–978). Public Health Services, US Dept of Health and Human Services. https://doi.org/10.1289/ehp.02110s6971

Asgari, Y., Khosravi, P., Zabihinpour, Z., & Habibi, M. (2018). Exploring candidate

biomarkers for lung and prostate cancers using gene expression and flux variability analysis. *Integrative Biology (United Kingdom)*, *10*(2), 113–120. https://doi.org/10.1039/c7ib00135e

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics* (Vol. 25, Issue 1, pp. 25–29). Nat Genet. https://doi.org/10.1038/75556

Atreya, R., & Siegmund, B. (2018). Development of therapy for and prediction of IBD - Getting personal. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 15, Issue 2). Nature Publishing Group. https://doi.org/10.1038/nrgastro.2017.166

Auer, P. L., Reiner, A. P., Wang, G., Kang, H. M., Abecasis, G. R., Altshuler, D., Bamshad, M. J., Nickerson, D. A., Tracy, R. P., Rich, S. S., & Leal, S. M. (2016). Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *American Journal of Human Genetics*, *99*(4), 791–801. https://doi.org/10.1016/j.ajhg.2016.08.012

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … Abecasis, G. R. (2015). A global

reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. B. (1988). Genetic disorders in children and young adults: A population study. *American Journal of Human Genetics*, *42*(5), 677–693. /pmc/articles/PMC1715177/?report=abstract

Barnich, N., Aguirre, J. E., Reinecker, H. C., Xavier, R., & Podolsky, D. K. (2005). Membrane recruitment of NOD2 in intestinal epithelial cells is essential for nuclear factor-KB activation in muramyl dipeptide recognition. *Journal of Cell Biology*, *170*(1), 21–26. https://doi.org/10.1083/jcb.200502153

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., … Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, *40*(8), 955–962. https://doi.org/10.1038/ng.175

Bartosova, Z., Fridrichova, I., Bujalkova, M., Wolf, B., Ilencikova, D., Krizan, P., Hlavcak, P., Palaj, J., Lukac, L., Lukacova, M., Böör, A., Haider, R., Jiricny, J., Nyström-Lahti, M., & Marra, G. (2003). Novel *MLH1* and *MSH2* germline mutations in the first HNPCC families identified in Slovakia. *Human Mutation*, *21*(4), 449–449. https://doi.org/10.1002/humu.9127

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., & Boone, C. (2013). Genetic Interaction Networks: Toward an Understanding of Heritability. *Annual Review of Genomics and Human Genetics*, *14*(1), 111–133.

https://doi.org/10.1146/annurev-genom-082509-141730

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Bear, C. E. (2020). A Therapy for Most with Cystic Fibrosis. *Cell*, *180*(2), 211. https://doi.org/10.1016/j.cell.2019.12.032

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., Adam, S., Bulman, D. E., Jones, S. J. M., Avard, D., Nguyen, M. T., Rousseau, F., Marshall, C., Wintle, R. F., Shen, Y., … Boycott, K. M. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, *94*(6), 809–817. https://doi.org/10.1016/j.ajhg.2014.05.003

Beltrame, L., Calura, E., Popovici, R. R., Rizzetto, L., Rivero Guedez, D., Donato, M., Romualdi, C., Draghici, S., & Cavalieri, D. (2011). *The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways*. *27*(15), 2127–2133. https://doi.org/10.1093/bioinformatics/btr339

Blüthgen, N., & Herzel, H. (2003). How robust are switches in intracellular signaling cascades? *Journal of Theoretical Biology*, *225*(3), 293–300.

https://doi.org/10.1016/S0022-5193(03)00247-9

Bordbar, A., McCloskey, D., Zielinski, D. C., Sonnenschein, N., Jamshidi, N., & Palsson, B. O. (2015). Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *Cell Systems*, *1*(4), 283–292. https://doi.org/10.1016/j.cels.2015.10.003

Borm, M. E. A., van Bodegraven, A. A., Mulder, C. J. J., Kraal, G., & Bouma, G. (2008). The effect of NOD2 activation on TLR2-mediated cytokine responses is dependent on activation dose and NOD2 genotype. *Genes and Immunity*, *9*(3), 274–278. https://doi.org/10.1038/gene.2008.9

Borodinov, N., Neumayer, S., Kalinin, S. V., Ovchinnikova, O. S., Vasudevan, R. K., & Jesse, S. (2019). Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy. *Npj Computational Materials*, *5*(1), 1–8. https://doi.org/10.1038/s41524-019-0148-5

Boué, S., Talikka, M., Westra, J. W., Hayes, W., Di Fabio, A., Park, J., Schlage, W. K., Sewer, A., Fields, B., Ansari, S., Martin, F., Veljkovic, E., Kenney, R., Peitsch, M. C., & Hoeng, J. (2015). Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database : The Journal of Biological Databases and Curation*, *2015*, bav030. https://doi.org/10.1093/database/bav030

Bourke, S. J. (2006). Interstitial lung disease: progress and problems. *Postgraduate Medical Journal*, *82*(970), 494–499. https://doi.org/10.1136/pgmj.2006.046417

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., Prlic, A., Sastry, A.,

Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M. T., Nielsen, J., … Palsson, B. O. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272–281. https://doi.org/10.1038/nbt.4072

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The single global macromolecular structure archive. In *Methods in Molecular Biology* (Vol. 1607, pp. 627–641). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7000-1_26

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., … Worthington, J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. https://doi.org/10.1038/nature05911

Byrne, A. B., Weirauch, M. T., Wong, V., Koeva, M., Dixon, S. J., Stuart, J. M., & Roy, P. J. (2007). A global analysis of genetic interactions in Caenorhabditis

elegans. *Journal of Biology*, *6*(3). https://doi.org/10.1186/jbiol58

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. In *Frontiers in Genetics* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fgene.2020.00424

Cardinale, S., & Cambray, G. (2017). Genome-wide analysis of E. coli cell-gene interactions. *BMC Systems Biology*, *11*(1). https://doi.org/10.1186/s12918-017-0494-1

Carter, C. O. (1977). Monogenic disorders. *Journal of Medical Genetics*, *14*(5), 316–320. https://doi.org/10.1136/jmg.14.5.316

Caruso, R., Warner, N., Inohara, N., & Núñez, G. (2014). NOD1 and NOD2: Signaling, host defense, and inflammatory disease. In *Immunity* (Vol. 41, Issue 6, pp. 898–908). Cell Press. https://doi.org/10.1016/j.immuni.2014.12.010

Cassa, C. A., Tong, M. Y., & Jordan, D. M. (2013). Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Human Mutation*, *34*(9), 1216–1220. https://doi.org/10.1002/humu.22375

Celebi, R., Uyar, H., Yasar, E., Gumus, O., Dikenelli, O., & Dumontier, M. (2019). Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3284-5

Chan, A. Y., Punwani, D., Kadlecek, T. A., Cowan, M. J., Olson, J. L., Mathes, E. F., Sunderam, U., Fu, S. M., Srinivasan, R., Kuriyan, J., Brenner, S. E., Weiss, A., & Puck, J. M. (2016). A novel human autoimmune syndrome caused by

combined hypomorphic and activating mutations in ZAP-70. *Journal of Experimental Medicine*, *213*(2), 155–165. https://doi.org/10.1084/jem.20150888

Chan, A. Y., Punwani, D., Kadlecek, T. A., Cowan, M. J., Olson, J. L., Mathes, E. F., Sunderam, U., Man Fu, S., Srinivasan, R., Kuriyan, J., Brenner, S. E., Weiss, A., & Puck, J. M. (2016). A novel human autoimmune syndrome caused by combined hypomorphic and activating mutations in ZAP-70. *Journal of Experimental Medicine*, *213*(2), 155–165.

Chandonia, J., Adhikari, A., Carraro, M., Chhibber, A., Cutting, G. R., Fu, Y., Gasparini, A., Jones, D. T., Kramer, A., Kundu, K., Lam, H. Y. K., Leonardi, E., Moult, J., Pal, L. R., Searls, D. B., Shah, S., Sunyaev, S., Tosatto, S. C. E., Yin, Y., & Buckley, B. A. (2017). Lessons from the CAGI-4 Hopkins clinical panel challenge. *Human Mutation*, *38*(9), 1155–1168. https://doi.org/10.1002/humu.23225

Chao, K. L., Gorlatova, N. V., Eisenstein, E., & Herzberg, O. (2014). Structural Basis for the Binding Specificity of Human Recepteur d'Origine Nantais (RON) Receptor Tyrosine Kinase to Macrophage-stimulating Protein. *Journal of Biological Chemistry*, *289*(43), 29948–29960. https://doi.org/10.1074/jbc.M114.594341

Chauhan, S., Mandell, M. A., & Deretic, V. (2016). Mechanism of action of the tuberculosis and Crohn disease risk factor IRGM in autophagy. *Autophagy*, *12*(2), 429–431. https://doi.org/10.1080/15548627.2015.1084457

Cheadle, J. P., Meredith, A. L., & al-Jader, L. N. (1992). A new missense mutation (R1283M) in exon 20 of the cystic fibrosis transmembrane conductance

regulator gene. *Human Molecular Genetics*, *1*(2), 123–125. http://www.ncbi.nlm.nih.gov/pubmed/1284468

Chen, I. Y., Agrawal, M., Horng, S., & Sontag, D. (2019). Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. *Biocomputing 2020*, 19–30. https://doi.org/10.1142/9789811215636_0003

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1653-z

Chibucos, M. C., Siegele, D. A., Hu, J. C., & Giglio, M. (2017). The evidence and conclusion ontology (ECO): Supporting GO annotations. In *Methods in Molecular Biology* (Vol. 1446, pp. 245–259). Humana Press Inc. https://doi.org/10.1007/978-1-4939-3743-1_18

Chou, I. C., & Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. In *Mathematical Biosciences* (Vol. 219, Issue 2, pp. 57–83). Math Biosci. https://doi.org/10.1016/j.mbs.2009.03.002

Chow, C. Y., Kelsey, K. J. P., Wolfner, M. F., & Clark, A. G. (2016). Candidate genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in Drosophila. *Human Molecular Genetics*, *25*(4), 651–659. https://doi.org/10.1093/HMG/DDV502

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J.,

Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., & Kingsmore, S. F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *Npj Genomic Medicine*, *3*(1). https://doi.org/10.1038/s41525-018-0053-8

Cleynen, I., & Halfvarsson, J. (2019). How to approach understanding complex trait genetics - inflammatory bowel disease as a model complex trait. In *United European gastroenterology journal* (Vol. 7, Issue 10, pp. 1426–1430). NLM (Medline). https://doi.org/10.1177/2050640619891120

Cohen, P. R. (2015). DARPA's Big Mechanism program. *Physical Biology*, *12*(4), 045008. https://doi.org/10.1088/1478-3975/12/4/045008

Come, J. H., Fraser, P. E., & Lansbury, P. T. (1993). A kinetic model for amyloid formation in the prion diseases: Importance of seeding. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(13), 5959–5963. https://doi.org/10.1073/pnas.90.13.5959

Condren, M. E., & Bradshaw, M. D. (2013). Ivacaftor: A Novel Gene-Based Therapeutic Approach for Cystic Fibrosis. *The Journal of Pediatric Pharmacology and Therapeutics*, *18*(1), 8–13. https://doi.org/10.5863/1551-6776-18.1.8

Cornish, A., & Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, *2015*, 456479. https://doi.org/10.1155/2015/456479

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Van Leeuwen, J., Van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353*(6306). https://doi.org/10.1126/science.aaf1420

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press.

Cutting, G. R. (2015). Cystic fibrosis genetics: From molecular understanding to clinical application. In *Nature Reviews Genetics* (Vol. 16, Issue 1, pp. 45–56). Nature Publishing Group. https://doi.org/10.1038/nrg3849

Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P. L., Babbi, G., Lena, P. Di, Casadio, R., Edwards, M., Gifford, D., Jones, D. T., Sundaram, L., Bhat, R. R., Li, X., Pal, L. R., Kundu, K., Yin, Y., Moult, J., Jiang, Y., … Morgan, A. A. (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*, *38*(9), 1182–1192. https://doi.org/10.1002/humu.23280

Danilkovitch, A., Donley, S., Skeel, A., & Leonard, E. J. (2000). Two Independent Signaling Pathways Mediate the Antiapoptotic Action of Macrophage-Stimulating Protein on Epithelial Cells. *Molecular and Cellular Biology*, *20*(6),

2218–2227. https://doi.org/10.1128/mcb.20.6.2218-2227.2000

Darden, L., Kundu, K., Pal, L. R., & Moult, J. (2018). Harnessing formal concepts of biological mechanism to analyze human disease. *PLoS Computational Biology*, *14*(12). https://doi.org/10.1371/journal.pcbi.1006540

De Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S. G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/ng.3760

Dehghan, A. (2018). Genome-wide association studies. In *Methods in Molecular Biology* (Vol. 1793, pp. 37–49). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7868-7_4

Devuyst, O. (2015). The 1000 genomes project: Welcome to a new world. In *Peritoneal Dialysis International* (Vol. 35, Issue 7, pp. 676–677). Multimed Inc. https://doi.org/10.3747/pdi.2015.00261

Drummond, J. T., Li, G. M., Longley, M. J., & Modrich, P. (1995). Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science (New York, N.Y.)*, *268*(5219), 1909–1912. https://doi.org/10.1126/science.7604264

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., …

Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Economou, M., Trikalinos, T. A., Loizou, K. T., Tsianos, E. V., & Ioannidis, J. P. A. (2004). Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: A metaanalysis. *American Journal of Gastroenterology*, *99*(12), 2393–2404. https://doi.org/10.1111/j.1572-0241.2004.40304.x

Edwards, S. L., Beesley, J., French, J. D., & Dunning, M. (2013). Beyond GWASs: Illuminating the dark road from association to function. In *American Journal of Human Genetics* (Vol. 93, Issue 5, pp. 779–797). Cell Press. https://doi.org/10.1016/j.ajhg.2013.10.012

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, *6*(5). https://doi.org/10.1186/gb-2005-6-5-r44

Ellner, S. P., & Guckenheimer, J. (2006). *Dynamic Models in Biology | Princeton University Press*.

https://press.princeton.edu/books/paperback/9780691125893/dynamic-models-in-biology

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Duenas Roca, C., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., … D'eustachio, P. (2017). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*, 649–655. https://doi.org/10.1093/nar/gkx1132

Fajac, I., Viel, M., Sublemontier, S., Hubert, D., Bienvenu, T., Pasteur, M., Helliwell, S., Houghton, S., Webb, S., Foreraker, J., Coulden, R., Flower, C., Bilton, D., Keogan, M., Eskandari, S., Snyder, P., Kreman, M., Zampighi, G., Welsh, M., … Warnock, D. (2008). Could a defective epithelial sodium channel lead to bronchiectasis. *Respiratory Research*, *9*(1), 46. https://doi.org/10.1186/1465-9921-9-46

Fang, H., Wu, Y., Narzisi, G., ORawe, J. A., Barrón, L. T. J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., Lyon, G. J., Gudmundsson, J., Sulem, P., Gudbjartsson, D., Masson, G., Agnarsson, B., Benediktsdottir, K., Sigurdsson, A., Magnusson, O., Gudjonsson, S., … Rothberg, J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, *6*(10), 89. https://doi.org/10.1186/s13073-014-0089-z

Farh, K. K. H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic

and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Femminella, G. D., Thayanandan, T., Calsolaro, V., Komici, K., Rengo, G., Corbi, G., & Ferrara, N. (2018). Imaging and molecular mechanisms of Alzheimer's disease: A review. In *International Journal of Molecular Sciences* (Vol. 19, Issue 12). MDPI AG. https://doi.org/10.3390/ijms19123702

Ferrell, J. E., & Machleder, E. M. (1998). The biochemical basis of an all-or-none cell fate switch in xenopus oocytes. *Science*, *280*(5365), 895–898. https://doi.org/10.1126/science.280.5365.895

Fischer, S., & Neurath, M. F. (2017). Precision Medicine in Inflammatory Bowel Diseases. *Clinical Pharmacology and Therapeutics*, *102*(4), 623–632. https://doi.org/10.1002/cpt.793

Frank, S. A. (2013). Input-output relations in biological systems: Measurement, information and the Hill equation. In *Biology Direct* (Vol. 8, Issue 1). Biol Direct. https://doi.org/10.1186/1745-6150-8-31

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125. https://doi.org/10.1038/ng.717

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., & Morris, Q. (2018). GeneMANIA update 2018. *Web Server Issue Published Online*, *46*.

https://doi.org/10.1093/nar/gky311

Galardini, M., Busby, B. P., Vieitez, C., Dunham, A. S., Typas, A., & Beltrao, P. (2019). The impact of the genetic background on gene deletion phenotypes in Saccharomyces cerevisiae . *Molecular Systems Biology*, *15*(12). https://doi.org/10.15252/msb.20198831

Gámez-Pozo, A., Trilla-Fuertes, L., Berges-Soria, J., Selevsek, N., López-Vacas, R., Díaz-Almirón, M., Nanni, P., Arevalillo, J. M., Navarro, H., Grossmann, J., Gayá Moreno, F., Gómez Rioja, R., Prado-Vázquez, G., Zapater-Moros, A., Main, P., Feliú, J., Martínez Del Prado, P., Zamora, P., Ciruelos, E., … Fresno Vara, J. Á. (2017). Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-10493-w

Gersemann, M., Becker, S., Kübler, I., Koslowski, M., Wang, G., Herrlinger, K. R., Griger, J., Fritz, P., Fellermann, K., Schwab, M., Wehkamp, J., & Stange, E. F. (2009). Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis. *Differentiation*, *77*(1), 84–94. https://doi.org/10.1016/j.diff.2008.09.008

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, a, Schenck, a, Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B., Kleefstra, T., Brunner, H. G., … Veltman, J. a. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. https://doi.org/10.1038/nature13394

Giral, H., Landmesser, U., & Kratzer, A. (2018). Into the Wild: GWAS Exploration of Non-coding RNAs. In *Frontiers in Cardiovascular Medicine* (Vol. 5). Frontiers Media S.A. https://doi.org/10.3389/fcvm.2018.00181

Girardin, S. E., Boneca, I. G., Viala, J., Chamaillard, M., Labigne, A., Thomas, G., Philpott, D. J., & Sansonetti, P. J. (2003). Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *Journal of Biological Chemistry*, *278*(11), 8869–8872. https://doi.org/10.1074/jbc.C200651200

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., & König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*, *44*(2), 125–138. https://doi.org/10.1002/gepi.22279

Gorlatova, N., Chao, K., Pal, L. R., Araj, R. H., Galkin, A., Turko, I., Moult, J., & Herzberg, O. (2011). Protein characterization of a candidate mechanism SNP for Crohn's disease: The macrophage stimulating protein R689C substitution. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027269

Greenberg, S. A., & Amato, A. A. (2004). Uncertainties in the pathogenesis of adult dermatomyositis. In *Current Opinion in Neurology* (Vol. 17, Issue 3, pp. 359–364). Curr Opin Neurol. https://doi.org/10.1097/00019052-200406000-00018

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I., Fitzgerald, G. A., Dolinski, K., Grosser, T., & Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific

networks. *Nature Genetics*, *47*(6), 569–576. https://doi.org/10.1038/ng.3259

Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., Ainscough, B. J., Ramirez, C. A., Rieke, D. T., Kujan, L., Barnell, E. K., Wagner, A. H., Skidmore, Z. L., Wollam, A., Liu, C. J., Jones, M. R., Bilski, R. L., Lesurf, R., Feng, Y. Y., … Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. In *Nature Genetics* (Vol. 49, Issue 2, pp. 170–174). Nature Publishing Group. https://doi.org/10.1038/ng.3774

Grimes, C. L., Ariyananda, L. D. Z., Melnyk, J. E., & O'Shea, E. K. (2012). The innate immune protein Nod2 binds directly to MDP, a bacterial cell wall fragment. *Journal of the American Chemical Society*, *134*(33), 13535–13537. https://doi.org/10.1021/ja303883c

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. In *Genome Biology* (Vol. 20, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1730-3

Gupta, S., Gellert, M., & Yang, W. (2012). Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. *Nature Structural and Molecular Biology*, *19*(1), 72–79. https://doi.org/10.1038/nsmb.2175

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, *3*(10), 1871–1878. https://doi.org/10.1371/journal.pcbi.0030189

Ha, N. T., Freytag, S., & Bickeboeller, H. (2014). Coverage and efficiency in current

SNP chips. *European Journal of Human Genetics*, *22*(9), 1124–1130. https://doi.org/10.1038/ejhg.2013.304

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017a). Human genetic variation and the gut microbiome in disease. In *Nature Reviews Genetics* (Vol. 18, Issue 11, pp. 690–699). Nature Publishing Group. https://doi.org/10.1038/nrg.2017.63

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017b). Human genetic variation and the gut microbiome in disease. In *Nature Reviews Genetics* (Vol. 18, Issue 11, pp. 690–699). Nature Publishing Group. https://doi.org/10.1038/nrg.2017.63

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., & Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nature Biotechnology*, *35*(5), 463–474. https://doi.org/10.1038/nbt.3834

Hanrahan, J. W., Matthes, E., Carlile, G., & Thomas, D. Y. (2017). Corrector combination therapies for F508del-CFTR. In *Current Opinion in Pharmacology* (Vol. 34, pp. 105–111). Elsevier Ltd. https://doi.org/10.1016/j.coph.2017.09.016

Harding, H. P., Novoa, I., Zhang, Y., Zeng, H., Wek, R., Schapira, M., & Ron, D. (2000). Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Molecular Cell*, *6*(5), 1099–1108. https://doi.org/10.1016/S1097-2765(00)00108-8

Harding, H. P., Zhang, Y., & Ron, D. (1999). Protein translation and folding are coupled by an endoplasmic- reticulum-resident kinase. *Nature*, *397*(6716), 271–274. https://doi.org/10.1038/16729

Harley, J. B., Alarcón-Riquelme, M. E., Criswell, L. A., Jacob, C. O., Kimberly, R.

P., Moser, K. L., Tsao, B. P., Vyse, T. J., Langefeld, C. D., Nath, S. K., Guthridge, J. M., Cobb, B. L., Mirel, D. B., Marion, M. C., Williams, A. H., Divers, J., Wang, W., Frank, S. G., Namjou, B., … Kelly, J. A. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nature Genetics*, *40*(2), 204–210. https://doi.org/10.1038/ng.81

Hasnain, S. Z., Tauro, S., Das, I., Tong, H., Chen, A. H., Jeffery, P. L., McDonald, V., Florin, T. H., & McGuckin, M. A. (2013). IL-10 promotes production of intestinal mucus by suppressing protein misfolding and endoplasmic reticulum stress in goblet cells. *Gastroenterology*, *144*(2). https://doi.org/10.1053/j.gastro.2012.10.043

Häuser, F., Deyle, C., Berard, D., Neukirch, C., Glowacki, C., Bickmann, J. K., Wenzel, J. J., Lackner, K. J., & Rossmann, H. (2012). Macrophage-stimulating protein polymorphism rs3197999 is associated with a gain of function: Implications for inflammatory bowel disease. *Genes and Immunity*, *13*(4), 321–327. https://doi.org/10.1038/gene.2011.88

Hayashi, R., Tsuchiya, K., Fukushima, K., Horita, N., Hibiya, S., Kitagaki, K., Negi, M., Itoh, E., Akashi, T., Eishi, Y., Okada, E., Araki, A., Ohtsuka, K., Fukuda, S., Ohno, H., Okamoto, R., Nakamura, T., Tanaka, S., Chayama, K., & Watanabe, M. (2016). Reduced human α-defensin 6 in noninflamed jejunal tissue of patients with Crohn's disease. *Inflammatory Bowel Diseases*, *22*(5), 1119–1128. https://doi.org/10.1097/MIB.0000000000000707

Hayes, B. (2013). Overview of Statistical Methods for Genome-Wide Association

Studies (GWAS). In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1019, pp. 149–169). Methods Mol Biol. https://doi.org/10.1007/978-1-62703-447-0_6

He, C., Kraft, P., Chen, C., Buring, J. E., Paré, G., Hankinson, S. E., Chanock, S. J., Ridker, P. M., Hunter, D. J., & Chasman, D. I. (2009). Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nature Genetics*, *41*(6), 724–728. https://doi.org/10.1038/ng.385

Heazlewood, C. K., Cook, M. C., Eri, R., Price, G. R., Tauro, S. B., Taupin, D., Thornton, D. J., Chin, W. P., Crockford, T. L., Cornall, R. J., Adams, R., Kato, M., Nelms, K. A., Hong, N. A., Florin, T. H. J., Goodnow, C. C., & McGuckin, M. A. (2008). Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS Medicine*, *5*(3), 0440–0460. https://doi.org/10.1371/journal.pmed.0050054

Henry, V. J., Goelzer, A., Ferré, A., Fischer, S., Dinh, M., Loux, V., Froidevaux, C., & Fromion, V. (2017). The bacterial interlocked process ONtology (BiPON): A systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of Biomedical Semantics*, *8*(1). https://doi.org/10.1186/s13326-017-0165-6

Hillmer, R. A. (2015). Systems Biology for Biologists. In *PLoS Pathogens* (Vol. 11, Issue 5). Public Library of Science. https://doi.org/10.1371/journal.ppat.1004786

Himmelstein, Daniel S., & Baranzini, S. E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Computational Biology*, *11*(7). https://doi.org/10.1371/journal.pcbi.1004259

Himmelstein, Daniel Scott, Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*. https://doi.org/10.7554/eLife.26726

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Hol, W. G., Halie, L. M., & Sander, C. (1981). Dipoles of the alpha-helix and beta-sheet: their role in protein folding. *Nature*, *294*(5841), 532–536. http://www.ncbi.nlm.nih.gov/pubmed/7312043

Hu, Z., Yu, C., Furutsuki, M., Andreoletti, G., Ly, M., Hoskins, R., Adhikari, A. N., & Brenner, S. E. (2019). VIPdb, a genetic Variant Impact Predictor Database. *Human Mutation*, *40*(9), 1202–1214. https://doi.org/10.1002/humu.23858

Hucka, M., Bergmann, F. T., Dräger, A., Hoops, S., Keating, S. M., Le Novère, N., Myers, C. J., Olivier, B. G., Sahle, S., Schaff, J. C., Smith, L. P., Waltemath, D., & Wilkinson, D. J. (2018). The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. *Journal of Integrative Bioinformatics*, *15*(1). https://doi.org/10.1515/jib-2017-0081

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel,

J. F., Sahbatou, M., & Thomas, G. (2001a). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, *411*(6837), 599–603. https://doi.org/10.1038/35079107

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., & Thomas, G. (2001b). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, *411*(6837), 599–603. https://doi.org/10.1038/35079107

Hugot, J. P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., Lee, J. C., Beaugerie, L., Naom, I., Dupas, J. L., Van Gossum, A., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C. G., Lennard-Jones, J. E., Cortot, A., Colombel, J. F., & Thomas, G. (1996). Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature*, *379*(6568), 821–823. https://doi.org/10.1038/379821a0

Hui, K. Y., Fernandez-Hernandez, H., Hu, J., Schaffner, A., Pankratz, N., Hsu, N. Y., Chuang, L. S., Carmi, S., Villaverde, N., Li, X., Rivas, M., Levine, A. P., Bao, X., Labrias, P. R., Haritunians, T., Ruane, D., Gettler, K., Chen, E., Li, D., … Peter, I. (2018). Functional variants in the LRRK2 gene confer shared effects on risk for Crohn's disease and Parkinson's disease. *Science Translational Medicine*, *10*(423). https://doi.org/10.1126/scitranslmed.aai7795

Hwang, S., Kim, E., Lee, I., Marcotte, E. M., Church, G. M., Lunshof, J. E., Bamshad, M. J., Do, R., Kathiresan, S., Abecasis, G. R., Pereira, P. C. B., Yang,

Y., Cirulli, E. T., Goldstein, D. B., Tennessen, J. A., Renkema, K. Y., Stokman, M. F., Giles, R. H., Knoers, N. V., … Karolchik, D. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*, 17875. https://doi.org/10.1038/srep17875

Ideker, T., & Krogan, N. J. (2012). Differential network biology. In *Molecular Systems Biology* (Vol. 8). Mol Syst Biol. https://doi.org/10.1038/msb.2011.99

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456. https://arxiv.org/abs/1502.03167v3

Ison, J., ˇ Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., & Kelso, J. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, *42*(22), 13534–13544. https://doi.org/10.1093/nar/gku1206

Johansson, M. E. V., Ambort, D., Pelaseyed, T., Schütte, A., Gustafsson, J. K., Ermund, A., Subramani, D. B., Holmén-Larsson, J. M., Thomsson, K. A., Bergström, J. H., Van Der Post, S., Rodriguez-Piñeiro, A. M., Sjövall, H., Bäckström, M., & Hansson, G. C. (2011). Composition and functional role of the mucus layers in the intestine. In *Cellular and Molecular Life Sciences* (Vol. 68, Issue 22, pp. 3635–3641). Cell Mol Life Sci. https://doi.org/10.1007/s00018-

011-0822-3

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Philip Schumm, L., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., … Whittaker, P. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, *491*(7422), 119–124. https://doi.org/10.1038/nature11582

Kadayakkara, D. K., Beatty, P. L., Turner, M. S., Janjic, J. M., Ahrens, E. T., & Finn, O. J. (2010). Inflammation driven by overexpression of the hypoglycosylated abnormal Mucin 1 (MUC1) links inflammatory bowel disease and pancreatitis. *Pancreas*, *39*(4), 510–515. https://doi.org/10.1097/MPA.0b013e3181bd6501

Kambouris, M., Maroun, R. C., Ben-Omran, T., Al-Sarraj, Y., Errafii, K., Ali, R., Boulos, H., Curmi, P. A., El-Shanti, H., Seelow, D., Schuelke, M., Hildebrandt, F., Nürnberg, P., Fiser, A., Modeller, S., Shen, M., Sali, A., Laskowski, R., MacArthur, M., … Fay, J. (2014). Mutations in zinc finger 407 [ZNF407] cause a unique autosomal recessive cognitive impairment syndrome. *Orphanet Journal of Rare Diseases*, *9*(1), 80. https://doi.org/10.1186/1750-1172-9-80

Kametani, F., & Hasegawa, M. (2018). Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease. In *Frontiers in Neuroscience* (Vol. 12, Issue JAN). Frontiers Media S.A. https://doi.org/10.3389/fnins.2018.00025

Kammermeier, J., Drury, S., James, C. T., Dziubak, R., Ocaka, L., Elawad, M., Beales, P., Lench, N., Uhlig, H. H., Bacchelli, C., & Shah, N. (2014). Targeted gene panel sequencing in children with very early onset inflammatory bowel

disease--evaluation and prospective analysis. *Journal of Medical Genetics*, *51*(11), 748–755. https://doi.org/10.1136/jmedgenet-2014-102624

Kanehisa, M, & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. http://www.ncbi.nlm.nih.gov/pubmed/10592173

Kanehisa, Minoru, Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*, 353–361. https://doi.org/10.1093/nar/gkw1092

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kaser, A., & Blumberg, R. S. (2014). ATG16L1 Crohn's disease risk stresses the endoplasmic reticulum of Paneth cells. In *Gut* (Vol. 63, Issue 7, pp. 1038–1039). BMJ Publishing Group. https://doi.org/10.1136/gutjnl-2013-306103

Kaser, A., Lee, A. H., Franke, A., Glickman, J. N., Zeissig, S., Tilg, H., Nieuwenhuis, E. E. S., Higgins, D. E., Schreiber, S., Glimcher, L. H., & Blumberg, R. S. (2008). XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease. *Cell*, *134*(5), 743–756. https://doi.org/10.1016/j.cell.2008.07.021

Keeling, M. J. (2005). Models of foot-and-mouth disease. In *Proceedings of the*

Royal Society B: Biological Sciences (Vol. 272, Issue 1569, pp. 1195–1202). Royal Society. https://doi.org/10.1098/rspb.2004.3046

Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G., & Lowe, J. B. (1995). Sequence and expression of a candidate for the human Secretor blood group α(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *Journal of Biological Chemistry*, *270*(9), 4640–4649. https://doi.org/10.1074/jbc.270.9.4640

Kendler, K. S., & Neale, M. C. (2009). "Familiality" or Heritability. *Archives of General Psychiatry*, *66*(4), 452. https://doi.org/10.1001/archgenpsychiatry.2009.14

Kikuchi, M., Ogishima, S., Mizuno, S., Miyashita, A., Kuwano, R., Nakaya, J., & Tanaka, H. (2015). Network-based analysis for uncovering mechanisms underlying Alzheimer's disease. In *Systems Biology of Alzheimer's Disease* (Vol. 1303, pp. 479–491). Springer New York. https://doi.org/10.1007/978-1-4939-2627-5_29

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, *28*(23), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591

Kinoshita, J., & Clark, T. (2007). Alzforum. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 401, pp. 365–381). Methods Mol Biol. https://doi.org/10.1007/978-1-59745-520-6_19

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Knodler, L. A., & Celli, J. (2011). Eating the strangers within: Host control of intracellular bacteria via xenophagy. In *Cellular Microbiology* (Vol. 13, Issue 9, pp. 1319–1327). Cell Microbiol. https://doi.org/10.1111/j.1462-5822.2011.01632.x

Kohl, P., & Noble, D. (2009). Systems biology and the virtual physiological human. In *Molecular Systems Biology* (Vol. 5). Mol Syst Biol. https://doi.org/10.1038/msb.2009.51

Konopka, T., & Smedley, D. (2020). Incremental data integration for tracking genotype-disease associations. *PLoS Computational Biology*, *16*(1). https://doi.org/10.1371/journal.pcbi.1007586

Korennykh, A. V., Egea, P. F., Korostelev, A. A., Finer-Moore, J., Zhang, C., Shokat, K. M., Stroud, R. M., & Walter, P. (2009). The unfolded protein response signals through high-order assembly of Ire1. *Nature*, *457*(7230), 687–693. https://doi.org/10.1038/nature07661

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(8), 1073–1081. https://doi.org/10.1038/nprot.2009.86

L. Kretschmann, K., Eyob, H., S. Buys, S., & L. Welm, A. (2010). The Macrophage Stimulating Protein/Ron Pathway as a Potential Therapeutic Target to Impede

Multiple Mechanisms Involved in Breast Cancer Progression. *Current Drug Targets, 11*(9), 1157–1168. https://doi.org/10.2174/138945010792006825

Laksshman, S., Bhat, R. R., Viswanath, V., & Li, X. (2017). DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. *Human Mutation, 38*(9), 1217–1224. https://doi.org/10.1002/humu.23272

Lancaster, M. A., & Huch, M. (2019). Disease modelling in human organoids. *DMM Disease Models and Mechanisms, 12*(7). https://doi.org/10.1242/dmm.039347

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research, 44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., Mcdaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research, 46*. https://doi.org/10.1093/nar/gkx1153

Lang, W. H., Coats, J. E., Majka, J., Hura, G. L., Lin, Y., Rasnik, I., & McMurray, C. T. (2011). Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. *Proceedings of the National Academy of Sciences of the United States of America, 108*(42).

https://doi.org/10.1073/pnas.1105461108

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-559

Lécine, P., Esmiol, S., Métais, J. Y., Nicoletti, C., Nourry, C., McDonald, C., Nunez, G., Hugot, J. P., Borg, J. P., & Ollendorff, V. (2007). The NOD2-RICK complex signals from the plasma membrane. *Journal of Biological Chemistry*, *282*(20), 15197–15207. https://doi.org/10.1074/jbc.M606242200

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. In *Genome Medicine* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00742-5

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. In *Nature Reviews Microbiology* (Vol. 10, Issue 4, pp. 291–305). Nature Publishing Group. https://doi.org/10.1038/nrmicro2737

Li, Han, Korennykh, A. V., Behrman, S. L., & Walter, P. (2010). Mammalian endoplasmic reticulum stress sensor IRE1 signals by dynamic clustering.

Proceedings of the National Academy of Sciences of the United States of America, *107*(37), 16113–16118. https://doi.org/10.1073/pnas.1010580107

Li, Heng. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)*, *30*(20), 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, J., Horstman, B., & Chen, Y. (2011). Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics*, *27*(13). https://doi.org/10.1093/bioinformatics/btr227

Li, M., Zhang, S., Qiu, Y., He, Y., Chen, B., Mao, R., Cui, Y., Zeng, Z., & Chen, M. (2017). Upregulation of miR-665 promotes apoptosis and colitis in inflammatory bowel disease by repressing the endoplasmic reticulum stress components XBP1 and ORMDL3. *Cell Death and Disease*, *8*(3). https://doi.org/10.1038/cddis.2017.76

Li, Y., Cho, H., Wang, F., Canela-Xandri, O., Luo, C., Rawlik, K., Archacki, S., Xu, C., Tenesa, A., Chen, Q., & Wang, Q. K. (2020). Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *Journal of the American Heart Association*, *9*(7), e014146. https://doi.org/10.1161/JAHA.119.014146

Lilyquist, J., Ruddy, K. J., Vachon, C. M., & Couch, F. J. (2018). Common genetic variation and breast cancer Risk—Past, present, and future. In *Cancer Epidemiology Biomarkers and Prevention* (Vol. 27, Issue 4, pp. 380–394). American Association for Cancer Research Inc. https://doi.org/10.1158/1055-9965.EPI-17-1144

Lin, Z., Wang, Z., Hegarty, J. P., Lin, T. R., Wang, Y., Deiling, S., Wu, R., Thomas, N. J., & Floros, J. (2017). Genetic association and epistatic interaction of the interleukin-10 signaling pathway in pediatric inflammatory bowel disease. *World Journal of Gastroenterology*, *23*(27), 4897–4909. https://doi.org/10.3748/wjg.v23.i27.4897

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W. W. L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M. T., Cordeiro, D., Cytrynbaum, C., Dell, S. D., … Marshall, C. R. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine*, *20*(4), 435–443. https://doi.org/10.1038/gim.2017.119

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). Nat Genet. https://doi.org/10.1038/ng.2653

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018a). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298. https://doi.org/10.1038/nmeth.4627

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018b). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298. https://doi.org/10.1038/nmeth.4627

Ma, X., Dai, Z., Sun, K., Zhang, Y., Chen, J., Yang, Y., Tso, P., Wu, G., & Wu, Z. (2017). Intestinal epithelial cell endoplasmic reticulum stress and inflammatory bowel disease pathogenesis: An update review. *Frontiers in Immunology*, *8*(OCT), 1271. https://doi.org/10.3389/fimmu.2017.01271

Madian, A. G., Wheeler, H. E., Jones, R. B., & Dolan, M. E. (2012). Relating human genetic variation to variation in drug responses. In *Trends in Genetics* (Vol. 28, Issue 10, pp. 487–495). Trends Genet. https://doi.org/10.1016/j.tig.2012.06.008

Maher, B. (2008). Personal genomes: The case of the missing heritability. In *Nature* (Vol. 456, Issue 7218, pp. 18–21). Nature Publishing Group. https://doi.org/10.1038/456018a

Malhotra, A., Younesi, E., Gündel, M., Müller, B., Heneka, M. T., & Hofmann-Apitius, M. (2014). ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & Dementia*, *10*(2), 238–246. https://doi.org/10.1016/j.jalz.2013.02.009

Mallah, N., Zapata-Cachafeiro, M., Aguirre, C., Ibarra-García, E., Palacios–Zabalza, I., Macías-García, F., Domínguez-Muñoz, J. E., Piñeiro-Lamas, M., Ibáñez, L., Vidal, X., Vendrell, L., Martin-Arias, L., Sáinz-Gil, M., Velasco-González, V., & Figueiras, A. (2020). Polymorphisms Involved in Platelet Activation and Inflammatory Response on Aspirin-Related Upper Gastrointestinal Bleeding: A Case-Control Study. *Frontiers in Pharmacology*, *11*. https://doi.org/10.3389/fphar.2020.00860

Mallott, J., Kwan, A., Church, J., Gonzalez-Espinosa, D., Lorey, F., Tang, L. F., Sunderam, U., Rana, S., Srinivasan, R., Brenner, S. E., & Puck, J. (2013).

Newborn screening for SCID identifies patients with ataxia telangiectasia. *Journal of Clinical Immunology*, *33*(3), 540–549. https://doi.org/10.1007/s10875-012-9846-1

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. In *Nature* (Vol. 461, Issue 7265, pp. 747–753). Nature. https://doi.org/10.1038/nature08494

Marín de Mas, I., Aguilar, E., Zodda, E., Balcells, C., Marin, S., Dallmann, G., Thomson, T. M., Papp, B., & Cascante, M. (2018). Model-driven discovery of long-chain fatty acid metabolic reprogramming in heterogeneous prostate cancer cells. *PLoS Computational Biology*, *14*(1). https://doi.org/10.1371/journal.pcbi.1005914

Martín-López, J. V., & Fishel, R. (2013). The mechanism of mismatch repair and the functional analysis of mismatch repair defects in Lynch syndrome. *Familial Cancer*, *12*(2), 159–168. https://doi.org/10.1007/s10689-013-9635-x

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U. S., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., Thomas, E., Scott, R. H., Baple, E., Tucci, A., Brittain, H., de Burca, A., Ibañez, K., Kasperaviciute, D., Smedley, D., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. In *Nature Genetics* (Vol. 51, Issue 11, pp. 1560–1565). Nature Publishing

Group. https://doi.org/10.1038/s41588-019-0528-2

Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., Nogueira, M. M., He, F., Parton, A., Lemonnier, N., Gawron, P., Gebel, S., Hainaut, P., Ollert, M., Dogrusoz, U., … Auffray, C. (2018). Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *Npj Systems Biology and Applications*, *4*(1). https://doi.org/10.1038/s41540-018-0059-y

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., Donnelly, P., Green, E., Guyer, M., Schrijver, I., Aziz, N., Farkas, D., Furtado, M., Gonzalez, A., Greiner, T., Grody, W., Hambuch, T., Kalman, L., Kant, J., … Trajanoski, Z. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 26. https://doi.org/10.1186/gm543

Mcgovern, D. P. B., Jones, M. R., Taylor, K. D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., Dutridge, D., Fleshner, P., Shih, D. Q., Melmed, G., Mengesha, E., King, L., Pressman, S., Haritunians, T., Guo, X., … Rotter, J. I. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Human Molecular Genetics*, *19*(17), 3468–3476. https://doi.org/10.1093/hmg/ddq248

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., … Blanchard, A. P. (2009). Sequence and structural

variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527–1541. https://doi.org/10.1101/gr.091868.109

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., Cunningham, F., Eisenstein, M., Weil, M., Chen, A., Visscher, P., Brown, M., McCarthy, M., Yang, J., Pierre, A. Saint, Génin, E., Zuk, O., Schaffner, S., … Liu, X. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Mendez, A. S., Alfaro, J., Morales-Soto, M. A., Dar, A. C., McCullagh, E., Gotthardt, K., Li, H., Acosta-Alvear, D., Sidrauski, C., Korennykh, A. V., Bernales, S., Shokat, K. M., & Walter, P. (2015). Endoplasmic reticulum stress-independent activation of unfolded protein response kinases by a small molecule ATP-mimic. *ELife*, *4*(MAY). https://doi.org/10.7554/eLife.05434

Meryhew, N. L., Kimberly, R. P., Messner, R. P., & Runquist, O. A. (1986). Mononuclear phagocyte system in SLE. II. A kinetic model of immune complex handling in systemic lupus erythematosus. *The Journal of Immunology*, *137*(1).

Mina, E., Thompson, M., Kaliyaperumal, R., Zhao, J., van der Horst, E., Tatum, Z., Hettne, K. M., Schultes, E. A., Mons, B., & Roos, M. (2015). Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study. *Journal of Biomedical Semantics*, *6*(1), 5. https://doi.org/10.1186/2041-1480-6-5

Miskovic, L., Tokic, M., Fengos, G., & Hatzimanikatis, V. (2015). Rites of passage: Requirements and standards for building kinetic models of metabolic

phenotypes. In *Current Opinion in Biotechnology* (Vol. 36, pp. 146–153). Elsevier Ltd. https://doi.org/10.1016/j.copbio.2015.08.019

Moehle, C., Ackermann, N., Langmann, T., Aslanidis, C., Kel, A., Kel-Margoulis, O., Schmitz-Madry, A., Zahn, A., Stremmel, W., & Schmitz, G. (2006). Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *Journal of Molecular Medicine, 84*(12), 1055–1066. https://doi.org/10.1007/s00109-006-0100-2

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., Von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S. A. G., Wong, K. C. C., Illig, T., Vogelberg, C., Weiland, S. K., Von Mutius, E., … Cookson, W. O. C. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, *448*(7152), 470–473. https://doi.org/10.1038/nature06014

Mohanan, V., Nakata, T., Desch, A. N., Lévesque, C., Boroughs, A., Guzman, G., Cao, Z., Creasey, E., Yao, J., Boucher, G., Charron, G., Bhan, A. K., Schenone, M., Carr, S. A., Reinecker, H. C., Daly, M. J., Rioux, J. D., Lassen, K. G., & Xavier, R. J. (2018). C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science, 359*(6380), 1161–1166. https://doi.org/10.1126/science.aan0814

Monteleone, G., Caruso, R., & Pallone, F. (2012). Role of Smad7 in inflammatory bowel diseases. *World Journal of Gastroenterology, 18*(40), 5664–5668. https://doi.org/10.3748/wjg.v18.i40.5664

Monteleone, G., Neurath, M. F., Ardizzone, S., Di Sabatino, A., Fantini, M. C.,

Castiglione, F., Scribano, M. L., Armuzzi, A., Caprioli, F., Sturniolo, G. C., Rogai, F., Vecchi, M., Atreya, R., Bossa, F., Onali, S., Fichera, M., Corazza, G. R., Biancone, L., Savarino, V., … Pallone, F. (2015). Mongersen, an oral SMAD7 antisense oligonucleotide, and crohn's disease. *New England Journal of Medicine*, *372*(12), 1104–1113. https://doi.org/10.1056/NEJMoa1407250

Morosky, S. A., Zhu, J., Mukherjee, A., Sarkar, S. N., & Coyne, C. B. (2011). Retinoic acid-induced gene-I (RIG-I) associates with nucleotide-binding oligomerization domain-2 (NOD2) to negatively regulate inflammatory signaling. *Journal of Biological Chemistry*, *286*(32), 28574–28583. https://doi.org/10.1074/jbc.M111.227942

Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N., Listgarten, J., Younger, S. T., Bernstein, B. E., Root, D. E., & Doench, J. G. (2018). Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. *Nature Biotechnology*, *36*(2), 179–189. https://doi.org/10.1038/nbt.4048

Negroni, A., Pierdomenico, M., Cucchiara, S., & Stronati, L. (2018). NOD2 and inflammation: Current insights. In *Journal of Inflammation Research* (Vol. 11, pp. 49–60). Dove Medical Press Ltd. https://doi.org/10.2147/JIR.S137606

Neurath, M. F. (2014). New targets for mucosal healing and therapy in inflammatory bowel diseases. In *Mucosal Immunology* (Vol. 7, Issue 1, pp. 6–19). Mucosal Immunol. https://doi.org/10.1038/mi.2013.73

Novère, N. Le, Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R.,

Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., … Kitano, H. (2009a). The Systems Biology Graphical Notation. In *Nature Biotechnology* (Vol. 27, Issue 8, pp. 735–741). Nat Biotechnol. https://doi.org/10.1038/nbt.1558

Novère, N. Le, Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., … Kitano, H. (2009b). The Systems Biology Graphical Notation. *Nature Biotechnology*, *27*(8), 735–741. https://doi.org/10.1038/nbt.1558

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y. Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., & Topper, S. (2017). Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine*, *19*(10), 1105–1117. https://doi.org/10.1038/gim.2017.37

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nũez, G., & Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, *411*(6837), 603–606. https://doi.org/10.1038/35079114

Okazaki, T., Murata, M., Kai, M., Adachi, K., Nakagawa, N., Kasagi, ‖noriko, Matsumura, W., Maegaki, Y., & Nanba, E. (2016). Clinical Diagnosis of Mendelian Disorders Using a Comprehensive Gene-Targeted Panel Test for Next-Generation Sequencing. *Yonago Acta Medica*, *59*, 118–125.

Okumura, R., & Takeda, K. (2018). Maintenance of intestinal homeostasis by mucosal barriers. In *Inflammation and Regeneration* (Vol. 38, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s41232-018-0063-z

Onge, R. P. S., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R. W., Nislow, C., Roth, F. P., & Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics*, *39*(2), 199–206. https://doi.org/10.1038/ng1948

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. https://doi.org/10.1093/bib/bbs086

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., & Baralle, F. E. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Human Molecular Genetics*, *12*(10), 1111–1120. https://doi.org/10.1093/hmg/ddg131

Pal, L. R., Chao, K. L., Moult, J., & Herzberg, O. (2017). On the interpretation of gasdermin-B expression quantitative trait loci data. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 114, Issue 38, pp. E7863–E7864). National Academy of Sciences. https://doi.org/10.1073/pnas.1712734114

Pal, L. R., Kundu, K., Yin, Y., & Moult, J. (2017). CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease.

Human Mutation, *38*(9), 1225–1234. https://doi.org/10.1002/humu.23256

Pal, L. R., Yu, C. H., Mount, S. M., & Moult, J. (2015). Insights from GWAS: Emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics*, *16*(8). https://doi.org/10.1186/1471-2164-16-S8-S4

Pansarasa, O., Bordoni, M., Drufuca, L., Diamanti, L., Sproviero, D., Trotti, R., Bernuzzi, S., Salvia, S. La, Gagliardi, S., Ceroni, M., & Cereda, C. (2018). Lymphoblastoid cell lines as a model to understand amyotrophic lateral sclerosis disease mechanisms. In *DMM Disease Models and Mechanisms* (Vol. 11, Issue 3). Company of Biologists Ltd. https://doi.org/10.1242/dmm.031625

Paone, P., & Cani, P. D. (2020). Mucus barrier, mucins and gut microbiota: the expected slimy partners? *Gut*, gutjnl-2020-322260. https://doi.org/10.1136/gutjnl-2020-322260

Park, J.-H., Kim, Y.-G., McDonald, C., Kanneganti, T.-D., Hasegawa, M., Body-Malapel, M., Inohara, N., & Núñez, G. (2007). RICK/RIP2 Mediates Innate Immune Responses Induced through Nod1 and Nod2 but Not TLRs. *The Journal of Immunology*, *178*(4), 2380–2386. https://doi.org/10.4049/jimmunol.178.4.2380

Patel, J. P., Puck, J. M., Srinivasan, R., Brown, C., Sunderam, U., Kundu, K., Brenner, S. E., Gatti, R. A., & Church, J. A. (2015a). Nijmegen Breakage Syndrome Detected by Newborn Screening for T Cell Receptor Excision Circles (TRECs). *Journal of Clinical Immunology*, *35*(2), 227. https://doi.org/10.1007/s10875-015-0136-6

Patel, J. P., Puck, J. M., Srinivasan, R., Brown, C., Sunderam, U., Kundu, K.,

Brenner, S. E., Gatti, R. A., & Church, J. A. (2015b). Nijmegen breakage syndrome detected by newborn screening for T cell receptor excision circles (TRECs). *Journal of Clinical Immunology*, *35*(2), 227–233. https://doi.org/10.1007/s10875-015-0136-6

PDBe-KB consortium. (2020). PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Research*, *48*(Database issue). https://doi.org/10.1093/nar/gkz853

Pegoraro, G., & Misteli, T. (2017). High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. In *Trends in Genetics* (Vol. 33, Issue 9, pp. 604–615). Elsevier Ltd. https://doi.org/10.1016/j.tig.2017.06.005

Peter, I., Mitchell, A. A., Ozelius, L., Erazo, M., Hu, J., Doheny, D., Abreu, M. T., Present, D. H., Ullman, T., Benkov, K., Korelitz, B. I., Mayer, L., & Desnick, R. J. (2011). Evaluation of 22 genetic variants with Crohn's Disease risk in the Ashkenazi Jewish population: a case-control study. *BMC Medical Genetics*, *12*. https://doi.org/10.1186/1471-2350-12-63

Phillips, P. C. (2008). Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. In *Nature Reviews Genetics* (Vol. 9, Issue 11, pp. 855–867). Nat Rev Genet. https://doi.org/10.1038/nrg2452

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., Zandi, P. P., Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M., Smith, S., Middle, C., Rodesch, M., Albert, T., Hannon, G., McCombie, W., Henson, J., … Sham, P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14.

https://doi.org/10.1186/1479-7364-8-14

Pon, A., Jewison, T., Su, Y., Liang, Y., Knox, C., Maciejewski, A., Wilson, M., & Wishart, D. S. (2015). Pathways with PathWhiz. *Nucleic Acids Research*, *43*. https://doi.org/10.1093/nar/gkv399

Posey, J. E. (2019). Genome sequencing and implications for rare disorders. In *Orphanet Journal of Rare Diseases* (Vol. 14, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13023-019-1127-0

Posey, J. E., O'Donnell-Luria, A. H., Chong, J. X., Harel, T., Jhangiani, S. N., Coban Akdemir, Z. H., Buyske, S., Pehlivan, D., Carvalho, C. M. B., Baxter, S., Sobreira, N., Liu, P., Wu, N., Rosenfeld, J. A., Kumar, S., Avramopoulos, D., White, J. J., Doheny, K. F., Witmer, P. D., … Lupski, J. R. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. In *Genetics in Medicine* (Vol. 21, Issue 4, pp. 798–812). Nature Publishing Group. https://doi.org/10.1038/s41436-018-0408-7

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756-63. https://doi.org/10.1093/nar/gkt1114

Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal*, *38*(10), 1499–1511. https://doi.org/10.1002/aic.690381003

Punwani, D., Zhang, Y., Yu, J., Cowan, M. J., Rana, S., Kwan, A., Adhikari, A. N., Lizama, C. O., Mendelsohn, B. A., Fahl, S. P., Chellappan, A., Srinivasan, R., Brenner, S. E., Wiest, D. L., & Puck, J. M. (2016). Multisystem Anomalies in Severe Combined Immunodeficiency with Mutant BCL11B. *New England Journal of Medicine*, *375*(22), 2165–2176. https://doi.org/10.1056/NEJMoa1509164

Ramos, M. P. M., Ribeiro, C., & Soares, A. J. (2019). A kinetic model of T cell autoreactivity in autoimmune diseases. *Journal of Mathematical Biology*, *79*(6–7), 2005–2031. https://doi.org/10.1007/s00285-019-01418-4

Resat, H., Petzold, L., & Pettigrew, M. F. (2009). Kinetic modeling of biological systems. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 541, pp. 311–335). Methods Mol Biol. https://doi.org/10.1007/978-1-59745-243-4_14

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015a). *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. https://doi.org/10.1038/gim.2015.30

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015b). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*,

17(5), 405–424. https://doi.org/10.1038/gim.2015.30

Ridley, K., & Condren, M. (2020). Elexacaftor-tezacaftor-ivacaftor: The first triple-combination cystic fibrosis transmembrane conductance regulator modulating therapy. *Journal of Pediatric Pharmacology and Therapeutics*, *25*(3), 192–197. https://doi.org/10.5863/1551-6776-25.3.192

Robinson, P. N., Kohler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., & Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, *24*(2), 340–348. https://doi.org/10.1101/gr.160325.113

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., & Tress, M. L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, *41*(D1), D110–D117. https://doi.org/10.1093/nar/gks1058

Sadedin, S. P., Dashnow, H., James, P. A., Bahlo, M., Bauer, D. C., Lonie, A., Lunke, S., Macciocca, I., Ross, J. P., Siemering, K. R., Stark, Z., White, S. M., Taylor, G., Gaff, C., Oshlack, A., & Thorne, N. P. (2015). Cpipe: A shared variant detection pipeline designed for diagnostic settings. *Genome Medicine*, *7*(1). https://doi.org/10.1186/s13073-015-0191-x

Salem, M., Ammitzboell, M., Nys, K., Seidelin, J. B., & Nielsen, O. H. (2015). ATG16L1: A multifunctional susceptibility factor in crohn disease. In *Autophagy* (Vol. 11, Issue 4, pp. 585–594). Taylor and Francis Inc. https://doi.org/10.1080/15548627.2015.1017187

Salomon, M. P., Li, W. L. S., Edlund, C. K., Morrison, J., Fortini, B. K., Win, A. K., Conti, D. V., Thomas, D. C., Duggan, D., Buchanan, D. D., Jenkins, M. A., Hopper, J. L., Gallinger, S., Le Marchand, L., Newcomb, P. A., Casey, G., & Marjoram, P. (2016). GWASeq: Targeted re-sequencing follow up to GWAS. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-2459-y

Sands, B. E., Feagan, B. G., Sandborn, W. J., Schreiber, S., Peyrin-Biroulet, L., Colombel, J. F., Rossiter, G., Usiskin, K., Ather, S., Zhan, X., & D'Haens, G. (2020). Mongersen (GED-0301) for active Crohn's disease: Results of a phase 3 study. *American Journal of Gastroenterology*, *115*(5), 738–745. https://doi.org/10.14309/ajg.0000000000000493

Santoso, J. W., & McCain, M. L. (2020). Neuromuscular disease modeling on a chip. In *DMM Disease Models and Mechanisms* (Vol. 13, Issue 7). Company of Biologists Ltd. https://doi.org/10.1242/dmm.044867

Sarkar, A., Duncan, M., Hart, J., Hertlein, E., Guttridge, D. C., & Wewers, M. D. (2006). ASC Directs NF-κB Activation by Regulating Receptor Interacting Protein-2 (RIP2) Caspase-1 Interactions. *The Journal of Immunology*, *176*(8), 4979–4986. https://doi.org/10.4049/jimmunol.176.8.4979

Sarkar, J., Dwivedi, G., Chen, Q., Sheu, I. E., Paich, M., Chelini, C. M., D'Alessandro, P. M., & Burns, S. P. (2019). A long-Term mechanistic computational model of physiological factors driving the onset of type 2 diabetes in an individual. *PLoS ONE*, *13*(2). https://doi.org/10.1371/journal.pone.0192472

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to

candidate causal variants by statistical fine-mapping. In *Nature Reviews Genetics* (Vol. 19, Issue 8, pp. 491–504). Nature Publishing Group. https://doi.org/10.1038/s41576-018-0016-z

Schroeder, B. O. (2019). Fight them or feed them: How the intestinal mucus layer manages the gut microbiota. In *Gastroenterology Report* (Vol. 7, Issue 1, pp. 3–12). Oxford University Press. https://doi.org/10.1093/gastro/goy052

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., … Hayes, V. M. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature*, *463*(7283), 943–947. https://doi.org/10.1038/nature08795

Sehgal, R., Sheahan, K., O'Connell, P. R., Hanly, A. M., Martin, S. T., & Winter, D. C. (2014). Lynch Syndrome: An updated review. *Genes*, *5*(3), 497–507. https://doi.org/10.3390/genes5030497

Shah, N., Hou, Y. C. C., Yu, H. C., Sainger, R., Caskey, C. T., Venter, J. C., & Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American Journal of Human Genetics*, *102*(4), 609–619. https://doi.org/10.1016/j.ajhg.2018.02.019

Shamseldin, H. E., Maddirevula, S., Faqeih, E., Ibrahim, N., Hashem, M., Shaheen, R., & Alkuraya, F. S. (2017). Increasing the sensitivity of clinical exome sequencing through improved filtration strategy. *Genetics in Medicine*, *19*(5), 593–598. https://doi.org/10.1038/gim.2016.155

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shannon, Paul, Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C. C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., … Mali, P. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, *14*(6), 573–576. https://doi.org/10.1038/nmeth.4225

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shimaoka, I., Kamide, K., Ohishi, M., Katsuya, T., Akasaka, H., Saitoh, S., Sugimoto, K., Oguro, R., Congrains, A., Fujisawa, T., Shimamoto, K., Ogihara, T., & Rakugi, H. (2010). Association of gene polymorphism of the fat-mass and obesity-associated gene with insulin resistance in Japanese. *Hypertension Research*, *33*(3), 214–218. https://doi.org/10.1038/hr.2009.215

Sidiq, T., Yoshihama, S., Downs, I., & Kobayashi, K. S. (2016). Nod2: A critical

regulator of ileal microbiota and Crohn's disease. In *Frontiers in Immunology* (Vol. 7, Issue SEP). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2016.00367

Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., & Moreau, Y. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, *10*(11), 1083–1084. https://doi.org/10.1038/nmeth.2656

Simms, L. A., Doecke, J. D., Walsh, M. D., Huang, N., Fowler, E. V., & Radford-Smith, G. L. (2008). Reduced α-defensin expression is associated with inflammation and not NOD2 mutation status in ileal Crohn's disease. *Gut*, *57*(7), 903–910. https://doi.org/10.1136/gut.2007.142588

Sioutos, N., Coronado, S. de, Haber, M. W., Hartel, F. W., Shaiu, W. L., & Wright, L. W. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, *40*(1), 30–43. https://doi.org/10.1016/j.jbi.2006.02.013

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *The American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H.,

Puviindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H. K., Naranjo, S., Acemel, R. D., … Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–375. https://doi.org/10.1038/nature13138

Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., LePichon, J. B., Miller, N. A., Thiffault, I., Dinwiddie, D. L., Twist, G., Noll, A., Heese, B. A., Zellmer, L., Atherton, A. M., Abdelmoity, A. T., Safina, N., Nyp, S. S., Zuccarelli, B., … Kingsmore, S. F. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science Translational Medicine*, *6*(265). https://doi.org/10.1126/scitranslmed.3010076

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., Boone, C., & Andrews, B. (2006). Mapping pathways and phenotypes by systematic gene overexpression. *Molecular Cell*, *21*(3), 319–330. https://doi.org/10.1016/j.molcel.2005.12.011

Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., & Altman, R. B. (2019). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Biocomputing 2020*, 463–474. https://doi.org/10.1142/9789811215636_0041

Soskic, B., Cano-Gamez, E., Smyth, D. J., Rowan, W. C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D. F., Larminie, C. G. C., Bronson, P. G., Willé, D., & Trynka, G. (2019). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics*, *51*(10), 1486–

1493. https://doi.org/10.1038/s41588-019-0493-9

Statement, I. C. (2000). Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment. *American Journal of Respiratory and Critical Care Medicine*, *161*(2), 646–664. https://doi.org/10.1164/ajrccm.161.2.ats3-00

Stavropoulos, D. J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R. K. C., Szego, M. J., Hayeems, R. Z., Shaul, R. Z., Brudno, M., Girdea, M., Frey, B., Alipanahi, B., Ahmed, S., Babul-Hirji, R., Porras, R. B., … Pinto, D. (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *Npj Genomic Medicine*, *1*, 15012. https://doi.org/10.1038/npjgenmed.2015.12

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 Update. *Human Mutation*, *21*(6), 577–581. https://doi.org/10.1002/humu.10212

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., & Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. In *Human Genetics* (Vol. 136, Issue 6, pp. 665–677). Springer Verlag. https://doi.org/10.1007/s00439-017-1779-6

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical

or genomical? *PLoS Biology*, *13*(7). https://doi.org/10.1371/journal.pbio.1002195

Stringer, S., Wray, N. R., Kahn, R. S., & Derks, E. M. (2011). Underestimated effect sizes in GWAS: Fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027964

Strober, W., & Watanabe, T. (2011). NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. In *Mucosal Immunology* (Vol. 4, Issue 5, pp. 484–495). Mucosal Immunol. https://doi.org/10.1038/mi.2011.29

Strober, Warren, Asano, N., Fuss, I., Kitani, A., & Watanabe, T. (2014). Cellular and molecular mechanisms underlying NOD2 risk-associated polymorphisms in Crohn's disease. In *Immunological Reviews* (Vol. 260, Issue 1, pp. 249–260). Blackwell Publishing Ltd. https://doi.org/10.1111/imr.12193

Strom, S. P., Lee, H., Das, K., Vilain, E., Nelson, S. F., Grody, W. W., & Deignan, J. L. (2014). Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genetics in Medicine*, *16*(7), 510–515. https://doi.org/10.1038/gim.2013.183

Süel, G. (2011). Use of Fluorescence Microscopy to Analyze Genetic Circuit Dynamics. In *Methods in enzymology* (Vol. 497, pp. 275–293). Methods Enzymol. https://doi.org/10.1016/b978-0-12-385075-1.00013-5

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Jensen, L. J., & Von Mering, C. (2018). STRING v11: protein-protein association networks with increased

coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*, 607–613. https://doi.org/10.1093/nar/gky1131

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. In *Epigenetics and Chromatin* (Vol. 8, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13072-015-0050-4

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J. B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., … McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, *47*(7), 717–726. https://doi.org/10.1038/ng.3304

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(Database issue). https://doi.org/10.1093/nar/gky1055

Thomas, P. D., Hill, D. P., Mi, H., Osumi-Sutherland, D., Van Auken, K., Carbon, S.,

Balhoff, J. P., Albou, L. P., Good, B., Gaudet, P., Lewis, S. E., & Mungall, C. J. (2019). Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. In *Nature Genetics* (Vol. 51, Issue 10, pp. 1429–1433). Nature Publishing Group. https://doi.org/10.1038/s41588-019-0500-1

Thul, P. J., & Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science*, *27*(1), 233–244. https://doi.org/10.1002/pro.3307

Tiwary, B. K. (2020). Computational medicine: Quantitative modeling of complex diseases. In *Briefings in Bioinformatics* (Vol. 21, Issue 2, pp. 429–440). Oxford University Press. https://doi.org/10.1093/bib/bbz005

Travassos, L. H., Carneiro, L. A. M., Ramjeet, M., Hussey, S., Kim, Y. G., Magalhes, J. G., Yuan, L., Soares, F., Chea, E., Le Bourhis, L., Boneca, I. G., Allaoui, A., Jones, N. L., Nũez, G., Girardin, S. E., & Philpott, D. J. (2010). Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nature Immunology*, *11*(1), 55–62. https://doi.org/10.1038/ni.1823

Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., & Katagiri, F. (2009). Network properties of robust immunity in plants. *PLoS Genetics*, *5*(12). https://doi.org/10.1371/journal.pgen.1000772

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *237*(641), 37–72. https://doi.org/10.1098/rstb.1952.0012

Umar, A., Buermeyer, A. B., Simon, J. A., Thomas, D. C., Clark, A. B., Liskay, R. M., & Kunkel, T. A. (1996). Requirement for PCNA in DNA Mismatch Repair at a Step Preceding DNA Resynthesis. *Cell*, *87*(1), 65–73. https://doi.org/10.1016/S0092-8674(00)81323-9

Underhill, D. M. (2007). Collaboration between the innate immune receptors dectin-1, TLRs, and Nods. In *Immunological Reviews* (Vol. 219, Issue 1, pp. 75–87). Immunol Rev. https://doi.org/10.1111/j.1600-065X.2007.00548.x

van Rooij, J., Arp, P., Broer, L., Verlouw, J., van Rooij, F., Kraaij, R., Uitterlinden, A., & Verkerk, A. J. M. H. (2020). Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time. *Genetics in Medicine*. https://doi.org/10.1038/s41436-020-0900-8

Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew, I., Uhlén, M., Pedersen, B. K., & Nielsen, J. (2015). Proteome- and Transcriptome-Driven Reconstruction of the Human Myocyte Metabolic Network and Its Use for Identification of Markers for Diabetes. *Cell Reports*, *11*(6), 921–933. https://doi.org/10.1016/j.celrep.2015.04.010

Vihinen, M. (2014). Variation Ontology for annotation of variation effects and mechanisms. *Genome Research*, *24*(2), 356–364. https://doi.org/10.1101/gr.157495.113

Vilar, E., Mork, M. E., Cuddy, A., Borras, E., Bannon, S. A., Taggart, M. W., Ying, J., Broaddus, R. R., Luthra, R., Rodriguez-Bigas, M. A., Lynch, P. M., & You, Y. Q. N. (2014). Role of microsatellite instability-low as a diagnostic biomarker

of Lynch syndrome in colorectal cancer. *Cancer Genetics*, *207*(10–12), 495–502. https://doi.org/10.1016/j.cancergen.2014.10.002

Vilela, E. G., da Gama Torres, H. O., Martin, F. P., de Lourdes de Abreu Ferrari, M., Andrade, M. M., & da Cunha, A. S. (2012). Evaluation of inflammatory activity in Crohn's disease and ulcerative colitis. *World Journal of Gastroenterology*, *18*(9), 872–881. https://doi.org/10.3748/wjg.v18.i9.872

Vincent, A. L., Jordan, C. A., Cadzow, M. J., Merriman, T. R., McGhee, C. N., YS., R., WJ, C., HY, P., JM, L., CB, C., KR, D., CA, J., H, O., Y, W., H, T., J, F., J, Z., R, S., ND, G., … A, A. (2014). Mutations in the Zinc Finger Protein Gene, *ZNF469* , Contribute to the Pathogenesis of Keratoconus. *Investigative Opthalmology & Visual Science*, *55*(9), 5629. https://doi.org/10.1167/iovs.14-14532

Visser, U., Abeyruwan, S., Vempati, U., Smith, R. P., Lemmon, V., & Schürer, S. C. (2011). BioAssay Ontology (BAO): A semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-257

Vivian-Griffiths, T., Baker, E., Schmidt, K. M., Bracher-Smith, M., Walters, J., Artemiou, A., Holmans, P., O'Donovan, M. C., Owen, M. J., Pocklington, A., & Escott-Price, V. (2019). Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *180*(1), 80–85. https://doi.org/10.1002/ajmg.b.32705

Voss, E., Wehkamp, J., Wehkamp, K., Stange, E. F., Schröder, J. M., & Harder, J.

(2006). NOD2/CARD15 mediates induction of the antimicrobial peptide human beta-defensin-2. *Journal of Biological Chemistry*, *281*(4), 2005–2011. https://doi.org/10.1074/jbc.M511044200

Vu, V., Verster, A. J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G. T., Moffat, J., & Fraser, A. G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell*, *162*(2), 391–402. https://doi.org/10.1016/j.cell.2015.06.037

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, M. H., Zhou, Y. Q., & Chen, Y. Q. (2002). Macrophage-stimulating protein and RON receptor tyrosine kinase: Potential regulators of macrophage inflammatory activities. In *Scandinavian Journal of Immunology* (Vol. 56, Issue 6, pp. 545–553). Scand J Immunol. https://doi.org/10.1046/j.1365-3083.2002.01177.x

Wang, X. R. ober., & Li, C. (2014). Decoding F508del misfolding in cystic fibrosis. In *Biomolecules* (Vol. 4, Issue 2, pp. 498–509). Biomolecules. https://doi.org/10.3390/biom4020498

Wang, Yan, & Wang, J.-G. (2018). Genome-Wide Association Studies of Hypertension and Several Other Cardiovascular Diseases. *Pulse*, *6*(3–4), 169–186. https://doi.org/10.1159/000496150

Wang, Yinhua, Díaz Arenas, C., Stoebel, D. M., & Cooper, T. F. (2013). Genetic background affects epistatic interactions between two beneficial mutations.

Biology Letters, *9*(1). https://doi.org/10.1098/rsbl.2012.0328

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–270. https://doi.org/10.1002/humu.22

Wehkamp, J., Koslowski, M., Wang, G., & Stange, E. F. (2008). Barrier dysfunction due to distinct defensin deficiencies in small intestinal and colonic crohn' s disease. In *Mucosal Immunology* (Vol. 1, pp. 67–74). Mucosal Immunol. https://doi.org/10.1038/mi.2008.48

Wehkamp, Jan, Salzman, N. H., Porter, E., Nuding, S., Weichenthal, M., Petras, R. E., Shen, B., Schaeffeler, E., Schwab, M., Linzmeier, R., Feathers, R. W., Chu, H., Lima, H., Fellermann, K., Ganz, T., Stange, E. F., & Bevins, C. L. (2005). Reduced Paneth cell α-defensins in ileal Crohn's disease. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(50), 18129–18134. https://doi.org/10.1073/pnas.0505256102

Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. In *Nature Reviews Genetics* (Vol. 15, Issue 11, pp. 722–733). Nature Publishing Group. https://doi.org/10.1038/nrg3747

Wek, R. C., Jiang, H.-Y., & Anthony, T. G. (2006). Coping with stress: eIF2 kinases and translational control. *Biochemical Society Transactions*, *34*(1), 7. https://doi.org/10.1042/bst20060007

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., Sanna, S., … Hirschhorn, J. N. (2009). Six new loci associated with body

mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, *41*(1), 25–34. https://doi.org/10.1038/ng.287

Wilson, S. S., Tocchi, A., Holly, M. K., Parks, W. C., & Smith, J. G. (2015). A small intestinal organoid model of non-invasive enteric pathogen-epithelial cell interactions. *Mucosal Immunology*, *8*(2), 352–361. https://doi.org/10.1038/mi.2014.72

Wolfert, M. A., Murray, T. F., Boons, G. J., & Moore, J. N. (2002). The origin of the synergistic effect of muramyl dipeptide with endotoxin and peptidoglycan. *Journal of Biological Chemistry*, *277*(42), 39179–39186. https://doi.org/10.1074/jbc.M204885200

Yamamoto-Furusho, J. K., Barnich, N., Hisamatsu, T., & Podolsky, D. K. (2010). MDP-NOD2 stimulation induces HNP-1 secretion, which contributes to NOD2 antibacterial function. *Inflammatory Bowel Diseases*, *16*(5), 736–742. https://doi.org/10.1002/ibd.21144

Yamamoto, S., & Ma, X. (2009). Role of Nod2 in the development of Crohn's disease. In *Microbes and Infection* (Vol. 11, Issue 12, pp. 912–918). Microbes Infect. https://doi.org/10.1016/j.micinf.2009.06.005

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017a). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 *NAGLU* (Human N-acetyl-glucosaminidase) and *UBE2I* (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017b). Ensemble variant interpretation

methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Younesi, E., Malhotra, A., Gündel, M., Scordis, P., Kodamullil, A. T., Page, M., Müller, B., Springstubbe, S., Wüllner, U., Scheller, D., & Hofmann-Apitius, M. (2015). PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. *Theoretical Biology & Medical Modelling*, *12*, 20. https://doi.org/10.1186/s12976-015-0017-y

Yuan, J., & Berg, H. C. (2013). Ultrasensitivity of an adaptive bacterial motor. *Journal of Molecular Biology*, *425*(10), 1760–1764. https://doi.org/10.1016/j.jmb.2013.02.016

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, *353*(2), 459–473. https://doi.org/10.1016/j.jmb.2005.08.020

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics, 7*(1), 166. https://doi.org/10.1186/1471-2105-7-166

Zawistowski, M., Reppell, M., Wegmann, D., St Jean, P. L., Ehm, M. G., Nelson, M. R., Novembre, J., & Zöllner, S. (2014). Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *European Journal of Human Genetics*, *22*(9), 1137–1144. https://doi.org/10.1038/ejhg.2013.297

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., … Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, *40*(5), 638–645. https://doi.org/10.1038/ng.120

Zeigler, V. L., Gillette, P. C., Crawford, F. A., Wiles, H. B., & Fyfe, D. A. (1990). New approaches to treatment of incessant ventricular tachycardia in the very young. *Journal of the American College of Cardiology*, *16*(3), 681–685. https://doi.org/10.1016/0735-1097(90)90360-2

Zhang, Y., Fan, H., Xu, J., Xiao, Y., Xu, Y., Li, Y., & Li, X. (2013). Network analysis reveals functional cross-links between disease and inflammation genes. *Scientific Reports*, *3*. https://doi.org/10.1038/srep03426

Acharya, S., Wilson, T., Gradia, S., Kane, M. F., Guerrette, S., Marsischky, G. T., Kolodner, R., & Fishel, R. (1996). hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13629–13634. https://doi.org/10.1073/pnas.93.24.13629

Adhikari, A. N., Gallagher, R. C., Wang, Y., Currier, R. J., Amatuni, G., Bassaganyas, L., Chen, F., Kundu, K., Kvale, M., Mooney, S. D., Nussbaum, R. L., Randi, S. S., Sanford, J., Shieh, J. T., Srinivasan, R., Sunderam, U., Tang, H., Vaka, D., Zou, Y., … Brenner, S. E. (2020). The role of exome sequencing in

newborn screening for inborn errors of metabolism. *Nature Medicine*, *26*(9), 1392–1397. https://doi.org/10.1038/s41591-020-0966-5

Adolph, T. E., Niederreiter, L., Blumberg, R. S., & Kaser, A. (2012). Endoplasmic reticulum stress and inflammation. *Digestive Diseases*, *30*(4), 341–346. https://doi.org/10.1159/000338121

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In *Current Protocols in Human Genetics* (pp. 7.20.1-7.20.41). John Wiley & Sons, Inc. https://doi.org/10.1002/0471142905.hg0720s76

Ahluwalia, B., Moraes, L., Magnusson, M. K., & Öhman, L. (2018). Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies. In *Scandinavian Journal of Gastroenterology* (Vol. 53, Issue 4, pp. 379–389). Taylor and Francis Ltd. https://doi.org/10.1080/00365521.2018.1447597

Al Nabhani, Z., Dietrich, G., Hugot, J. P., & Barreau, F. (2017). Nod2: The intestinal gate keeper. In *PLoS Pathogens* (Vol. 13, Issue 3). Public Library of Science. https://doi.org/10.1371/journal.ppat.1006177

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J. H., Yang, J., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. https://doi.org/10.1038/nature09410

Allot, A., Peng, Y., Wei, C. H., Lee, K., Phan, L., & Lu, Z. (2018). LitVar: A semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Research*, *46*(W1), W530–W536. https://doi.org/10.1093/nar/gky355

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151

Andersen, M. E., Yang, R. S. H., French, C. T., Chubb, L. S., & Dennison, J. E. (2002). Molecular circuits, biological switches, and nonlinear dose-response relationships. In *Environmental Health Perspectives* (Vol. 110, Issue SUPPL. 6, pp. 971–978). Public Health Services, US Dept of Health and Human Services. https://doi.org/10.1289/ehp.02110s6971

Asgari, Y., Khosravi, P., Zabihinpour, Z., & Habibi, M. (2018). Exploring candidate biomarkers for lung and prostate cancers using gene expression and flux variability analysis. *Integrative Biology (United Kingdom)*, *10*(2), 113–120. https://doi.org/10.1039/c7ib00135e

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics* (Vol. 25, Issue 1, pp. 25–29). Nat Genet. https://doi.org/10.1038/75556

Atreya, R., & Siegmund, B. (2018). Development of therapy for and prediction of

IBD - Getting personal. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 15, Issue 2). Nature Publishing Group. https://doi.org/10.1038/nrgastro.2017.166

Auer, P. L., Reiner, A. P., Wang, G., Kang, H. M., Abecasis, G. R., Altshuler, D., Bamshad, M. J., Nickerson, D. A., Tracy, R. P., Rich, S. S., & Leal, S. M. (2016). Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *American Journal of Human Genetics*, *99*(4), 791–801. https://doi.org/10.1016/j.ajhg.2016.08.012

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. B. (1988). Genetic disorders in children and young adults: A population study. *American Journal of Human Genetics*, *42*(5), 677–693. /pmc/articles/PMC1715177/?report=abstract

Barnich, N., Aguirre, J. E., Reinecker, H. C., Xavier, R., & Podolsky, D. K. (2005). Membrane recruitment of NOD2 in intestinal epithelial cells is essential for nuclear factor-KB activation in muramyl dipeptide recognition. *Journal of Cell Biology*, *170*(1), 21–26. https://doi.org/10.1083/jcb.200502153

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D.,

Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., … Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, *40*(8), 955–962. https://doi.org/10.1038/ng.175

Bartosova, Z., Fridrichova, I., Bujalkova, M., Wolf, B., Ilencikova, D., Krizan, P., Hlavcak, P., Palaj, J., Lukac, L., Lukacova, M., Böör, A., Haider, R., Jiricny, J., Nyström-Lahti, M., & Marra, G. (2003). Novel *MLH1* and *MSH2* germline mutations in the first HNPCC families identified in Slovakia. *Human Mutation*, *21*(4), 449–449. https://doi.org/10.1002/humu.9127

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., & Boone, C. (2013). Genetic Interaction Networks: Toward an Understanding of Heritability. *Annual Review of Genomics and Human Genetics*, *14*(1), 111–133. https://doi.org/10.1146/annurev-genom-082509-141730

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Bear, C. E. (2020). A Therapy for Most with Cystic Fibrosis. *Cell*, *180*(2), 211. https://doi.org/10.1016/j.cell.2019.12.032

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B.

A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., Adam, S., Bulman, D. E., Jones, S. J. M., Avard, D., Nguyen, M. T., Rousseau, F., Marshall, C., Wintle, R. F., Shen, Y., … Boycott, K. M. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, *94*(6), 809–817. https://doi.org/10.1016/j.ajhg.2014.05.003

Beltrame, L., Calura, E., Popovici, R. R., Rizzetto, L., Rivero Guedez, D., Donato, M., Romualdi, C., Draghici, S., & Cavalieri, D. (2011). *The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways*. *27*(15), 2127–2133. https://doi.org/10.1093/bioinformatics/btr339

Blüthgen, N., & Herzel, H. (2003). How robust are switches in intracellular signaling cascades? *Journal of Theoretical Biology*, *225*(3), 293–300. https://doi.org/10.1016/S0022-5193(03)00247-9

Bordbar, A., McCloskey, D., Zielinski, D. C., Sonnenschein, N., Jamshidi, N., & Palsson, B. O. (2015). Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *Cell Systems*, *1*(4), 283–292. https://doi.org/10.1016/j.cels.2015.10.003

Borm, M. E. A., van Bodegraven, A. A., Mulder, C. J. J., Kraal, G., & Bouma, G. (2008). The effect of NOD2 activation on TLR2-mediated cytokine responses is dependent on activation dose and NOD2 genotype. *Genes and Immunity*, *9*(3), 274–278. https://doi.org/10.1038/gene.2008.9

Borodinov, N., Neumayer, S., Kalinin, S. V., Ovchinnikova, O. S., Vasudevan, R. K.,

& Jesse, S. (2019). Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy. *Npj Computational Materials*, *5*(1), 1–8. https://doi.org/10.1038/s41524-019-0148-5

Boué, S., Talikka, M., Westra, J. W., Hayes, W., Di Fabio, A., Park, J., Schlage, W. K., Sewer, A., Fields, B., Ansari, S., Martin, F., Veljkovic, E., Kenney, R., Peitsch, M. C., & Hoeng, J. (2015). Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database : The Journal of Biological Databases and Curation*, *2015*, bav030. https://doi.org/10.1093/database/bav030

Bourke, S. J. (2006). Interstitial lung disease: progress and problems. *Postgraduate Medical Journal*, *82*(970), 494–499. https://doi.org/10.1136/pgmj.2006.046417

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., Prlic, A., Sastry, A., Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M. T., Nielsen, J., … Palsson, B. O. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272–281. https://doi.org/10.1038/nbt.4072

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012.

https://doi.org/10.1093/nar/gky1120

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The single global macromolecular structure archive. In *Methods in Molecular Biology* (Vol. 1607, pp. 627–641). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7000-1_26

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., … Worthington, J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. https://doi.org/10.1038/nature05911

Byrne, A. B., Weirauch, M. T., Wong, V., Koeva, M., Dixon, S. J., Stuart, J. M., & Roy, P. J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. *Journal of Biology*, *6*(3). https://doi.org/10.1186/jbiol58

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. In *Frontiers in Genetics* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fgene.2020.00424

Cardinale, S., & Cambray, G. (2017). Genome-wide analysis of E. coli cell-gene interactions. *BMC Systems Biology*, *11*(1). https://doi.org/10.1186/s12918-017-0494-1

Carter, C. O. (1977). Monogenic disorders. *Journal of Medical Genetics*, *14*(5), 316–320. https://doi.org/10.1136/jmg.14.5.316

Caruso, R., Warner, N., Inohara, N., & Núñez, G. (2014). NOD1 and NOD2: Signaling, host defense, and inflammatory disease. In *Immunity* (Vol. 41, Issue 6, pp. 898–908). Cell Press. https://doi.org/10.1016/j.immuni.2014.12.010

Cassa, C. A., Tong, M. Y., & Jordan, D. M. (2013). Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Human Mutation*, *34*(9), 1216–1220. https://doi.org/10.1002/humu.22375

Celebi, R., Uyar, H., Yasar, E., Gumus, O., Dikenelli, O., & Dumontier, M. (2019). Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3284-5

Chan, A. Y., Punwani, D., Kadlecek, T. A., Cowan, M. J., Olson, J. L., Mathes, E. F., Sunderam, U., Fu, S. M., Srinivasan, R., Kuriyan, J., Brenner, S. E., Weiss, A., & Puck, J. M. (2016). A novel human autoimmune syndrome caused by combined hypomorphic and activating mutations in ZAP-70. *Journal of Experimental Medicine*, *213*(2), 155–165. https://doi.org/10.1084/jem.20150888

Chan, A. Y., Punwani, D., Kadlecek, T. A., Cowan, M. J., Olson, J. L., Mathes, E. F., Sunderam, U., Man Fu, S., Srinivasan, R., Kuriyan, J., Brenner, S. E., Weiss, A., & Puck, J. M. (2016). A novel human autoimmune syndrome caused by combined hypomorphic and activating mutations in ZAP-70. *Journal of Experimental Medicine*, *213*(2), 155–165.

Chandonia, J., Adhikari, A., Carraro, M., Chhibber, A., Cutting, G. R., Fu, Y., Gasparini, A., Jones, D. T., Kramer, A., Kundu, K., Lam, H. Y. K., Leonardi, E., Moult, J., Pal, L. R., Searls, D. B., Shah, S., Sunyaev, S., Tosatto, S. C. E., Yin,

Y., & Buckley, B. A. (2017). Lessons from the CAGI-4 Hopkins clinical panel challenge. *Human Mutation*, *38*(9), 1155–1168. https://doi.org/10.1002/humu.23225

Chao, K. L., Gorlatova, N. V., Eisenstein, E., & Herzberg, O. (2014). Structural Basis for the Binding Specificity of Human Recepteur d'Origine Nantais (RON) Receptor Tyrosine Kinase to Macrophage-stimulating Protein. *Journal of Biological Chemistry*, *289*(43), 29948–29960. https://doi.org/10.1074/jbc.M114.594341

Chauhan, S., Mandell, M. A., & Deretic, V. (2016). Mechanism of action of the tuberculosis and Crohn disease risk factor IRGM in autophagy. *Autophagy*, *12*(2), 429–431. https://doi.org/10.1080/15548627.2015.1084457

Cheadle, J. P., Meredith, A. L., & al-Jader, L. N. (1992). A new missense mutation (R1283M) in exon 20 of the cystic fibrosis transmembrane conductance regulator gene. *Human Molecular Genetics*, *1*(2), 123–125. http://www.ncbi.nlm.nih.gov/pubmed/1284468

Chen, I. Y., Agrawal, M., Horng, S., & Sontag, D. (2019). Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. *Biocomputing 2020*, 19–30. https://doi.org/10.1142/9789811215636_0003

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1653-z

Chibucos, M. C., Siegele, D. A., Hu, J. C., & Giglio, M. (2017). The evidence and conclusion ontology (ECO): Supporting GO annotations. In *Methods in Molecular Biology* (Vol. 1446, pp. 245–259). Humana Press Inc. https://doi.org/10.1007/978-1-4939-3743-1_18

Chou, I. C., & Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. In *Mathematical Biosciences* (Vol. 219, Issue 2, pp. 57–83). Math Biosci. https://doi.org/10.1016/j.mbs.2009.03.002

Chow, C. Y., Kelsey, K. J. P., Wolfner, M. F., & Clark, A. G. (2016). Candidate genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in Drosophila. *Human Molecular Genetics*, *25*(4), 651–659. https://doi.org/10.1093/HMG/DDV502

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., & Kingsmore, S. F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *Npj Genomic Medicine*, *3*(1). https://doi.org/10.1038/s41525-018-0053-8

Cleynen, I., & Halfvarsson, J. (2019). How to approach understanding complex trait

genetics - inflammatory bowel disease as a model complex trait. In *United European gastroenterology journal* (Vol. 7, Issue 10, pp. 1426–1430). NLM (Medline). https://doi.org/10.1177/2050640619891120

Cohen, P. R. (2015). DARPA's Big Mechanism program. *Physical Biology*, *12*(4), 045008. https://doi.org/10.1088/1478-3975/12/4/045008

Come, J. H., Fraser, P. E., & Lansbury, P. T. (1993). A kinetic model for amyloid formation in the prion diseases: Importance of seeding. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(13), 5959–5963. https://doi.org/10.1073/pnas.90.13.5959

Condren, M. E., & Bradshaw, M. D. (2013). Ivacaftor: A Novel Gene-Based Therapeutic Approach for Cystic Fibrosis. *The Journal of Pediatric Pharmacology and Therapeutics*, *18*(1), 8–13. https://doi.org/10.5863/1551-6776-18.1.8

Cornish, A., & Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, *2015*, 456479. https://doi.org/10.1155/2015/456479

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Van Leeuwen, J., Van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353*(6306). https://doi.org/10.1126/science.aaf1420

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the*

Life Sciences. University of Chicago Press.

Cutting, G. R. (2015). Cystic fibrosis genetics: From molecular understanding to clinical application. In *Nature Reviews Genetics* (Vol. 16, Issue 1, pp. 45–56). Nature Publishing Group. https://doi.org/10.1038/nrg3849

Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P. L., Babbi, G., Lena, P. Di, Casadio, R., Edwards, M., Gifford, D., Jones, D. T., Sundaram, L., Bhat, R. R., Li, X., Pal, L. R., Kundu, K., Yin, Y., Moult, J., Jiang, Y., … Morgan, A. A. (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*, *38*(9), 1182–1192. https://doi.org/10.1002/humu.23280

Danilkovitch, A., Donley, S., Skeel, A., & Leonard, E. J. (2000). Two Independent Signaling Pathways Mediate the Antiapoptotic Action of Macrophage-Stimulating Protein on Epithelial Cells. *Molecular and Cellular Biology*, *20*(6), 2218–2227. https://doi.org/10.1128/mcb.20.6.2218-2227.2000

Darden, L., Kundu, K., Pal, L. R., & Moult, J. (2018). Harnessing formal concepts of biological mechanism to analyze human disease. *PLoS Computational Biology*, *14*(12). https://doi.org/10.1371/journal.pcbi.1006540

De Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S. G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/ng.3760

Dehghan, A. (2018). Genome-wide association studies. In *Methods in Molecular Biology* (Vol. 1793, pp. 37–49). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7868-7_4

Devuyst, O. (2015). The 1000 genomes project: Welcome to a new world. In *Peritoneal Dialysis International* (Vol. 35, Issue 7, pp. 676–677). Multimed Inc. https://doi.org/10.3747/pdi.2015.00261

Drummond, J. T., Li, G. M., Longley, M. J., & Modrich, P. (1995). Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science (New York, N.Y.), 268*(5219), 1909–1912. https://doi.org/10.1126/science.7604264

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Economou, M., Trikalinos, T. A., Loizou, K. T., Tsianos, E. V., & Ioannidis, J. P. A. (2004). Differential effects of NOD2 variants on Crohn's disease risk and

phenotype in diverse populations: A metaanalysis. *American Journal of Gastroenterology*, *99*(12), 2393–2404. https://doi.org/10.1111/j.1572-0241.2004.40304.x

Edwards, S. L., Beesley, J., French, J. D., & Dunning, M. (2013). Beyond GWASs: Illuminating the dark road from association to function. In *American Journal of Human Genetics* (Vol. 93, Issue 5, pp. 779–797). Cell Press. https://doi.org/10.1016/j.ajhg.2013.10.012

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, *6*(5). https://doi.org/10.1186/gb-2005-6-5-r44

Ellner, S. P., & Guckenheimer, J. (2006). *Dynamic Models in Biology | Princeton University Press*. https://press.princeton.edu/books/paperback/9780691125893/dynamic-models-in-biology

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Duenas Roca, C., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., … D'eustachio, P. (2017). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*, 649–655. https://doi.org/10.1093/nar/gkx1132

Fajac, I., Viel, M., Sublemontier, S., Hubert, D., Bienvenu, T., Pasteur, M., Helliwell, S., Houghton, S., Webb, S., Foreraker, J., Coulden, R., Flower, C., Bilton, D., Keogan, M., Eskandari, S., Snyder, P., Kreman, M., Zampighi, G., Welsh, M.,

… Warnock, D. (2008). Could a defective epithelial sodium channel lead to bronchiectasis. *Respiratory Research*, *9*(1), 46. https://doi.org/10.1186/1465-9921-9-46

Fang, H., Wu, Y., Narzisi, G., ORawe, J. A., Barrón, L. T. J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., Lyon, G. J., Gudmundsson, J., Sulem, P., Gudbjartsson, D., Masson, G., Agnarsson, B., Benediktsdottir, K., Sigurdsson, A., Magnusson, O., Gudjonsson, S., … Rothberg, J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, *6*(10), 89. https://doi.org/10.1186/s13073-014-0089-z

Farh, K. K. H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Femminella, G. D., Thayanandan, T., Calsolaro, V., Komici, K., Rengo, G., Corbi, G., & Ferrara, N. (2018). Imaging and molecular mechanisms of Alzheimer's disease: A review. In *International Journal of Molecular Sciences* (Vol. 19, Issue 12). MDPI AG. https://doi.org/10.3390/ijms19123702

Ferrell, J. E., & Machleder, E. M. (1998). The biochemical basis of an all-or-none cell fate switch in xenopus oocytes. *Science*, *280*(5365), 895–898. https://doi.org/10.1126/science.280.5365.895

Fischer, S., & Neurath, M. F. (2017). Precision Medicine in Inflammatory Bowel

Diseases. *Clinical Pharmacology and Therapeutics*, *102*(4), 623–632. https://doi.org/10.1002/cpt.793

Frank, S. A. (2013). Input-output relations in biological systems: Measurement, information and the Hill equation. In *Biology Direct* (Vol. 8, Issue 1). Biol Direct. https://doi.org/10.1186/1745-6150-8-31

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125. https://doi.org/10.1038/ng.717

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., & Morris, Q. (2018). GeneMANIA update 2018. *Web Server Issue Published Online*, *46*. https://doi.org/10.1093/nar/gky311

Galardini, M., Busby, B. P., Vieitez, C., Dunham, A. S., Typas, A., & Beltrao, P. (2019). The impact of the genetic background on gene deletion phenotypes in Saccharomyces cerevisiae . *Molecular Systems Biology*, *15*(12). https://doi.org/10.15252/msb.20198831

Gámez-Pozo, A., Trilla-Fuertes, L., Berges-Soria, J., Selevsek, N., López-Vacas, R., Díaz-Almirón, M., Nanni, P., Arevalillo, J. M., Navarro, H., Grossmann, J., Gayá Moreno, F., Gómez Rioja, R., Prado-Vázquez, G., Zapater-Moros, A., Main, P., Feliú, J., Martínez Del Prado, P., Zamora, P., Ciruelos, E., … Fresno Vara, J. Á. (2017). Functional proteomics outlines the complexity of breast

cancer molecular subtypes. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-10493-w

Gersemann, M., Becker, S., Kübler, I., Koslowski, M., Wang, G., Herrlinger, K. R., Griger, J., Fritz, P., Fellermann, K., Schwab, M., Wehkamp, J., & Stange, E. F. (2009). Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis. *Differentiation*, *77*(1), 84–94. https://doi.org/10.1016/j.diff.2008.09.008

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, a, Schenck, a, Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B., Kleefstra, T., Brunner, H. G., … Veltman, J. a. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. https://doi.org/10.1038/nature13394

Giral, H., Landmesser, U., & Kratzer, A. (2018). Into the Wild: GWAS Exploration of Non-coding RNAs. In *Frontiers in Cardiovascular Medicine* (Vol. 5). Frontiers Media S.A. https://doi.org/10.3389/fcvm.2018.00181

Girardin, S. E., Boneca, I. G., Viala, J., Chamaillard, M., Labigne, A., Thomas, G., Philpott, D. J., & Sansonetti, P. J. (2003). Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *Journal of Biological Chemistry*, *278*(11), 8869–8872. https://doi.org/10.1074/jbc.C200651200

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., & König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting

coronary artery disease status. *Genetic Epidemiology*, *44*(2), 125–138. https://doi.org/10.1002/gepi.22279

Gorlatova, N., Chao, K., Pal, L. R., Araj, R. H., Galkin, A., Turko, I., Moult, J., & Herzberg, O. (2011). Protein characterization of a candidate mechanism SNP for Crohn's disease: The macrophage stimulating protein R689C substitution. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027269

Greenberg, S. A., & Amato, A. A. (2004). Uncertainties in the pathogenesis of adult dermatomyositis. In *Current Opinion in Neurology* (Vol. 17, Issue 3, pp. 359–364). Curr Opin Neurol. https://doi.org/10.1097/00019052-200406000-00018

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I., Fitzgerald, G. A., Dolinski, K., Grosser, T., & Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, *47*(6), 569–576. https://doi.org/10.1038/ng.3259

Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., Ainscough, B. J., Ramirez, C. A., Rieke, D. T., Kujan, L., Barnell, E. K., Wagner, A. H., Skidmore, Z. L., Wollam, A., Liu, C. J., Jones, M. R., Bilski, R. L., Lesurf, R., Feng, Y. Y., … Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. In *Nature Genetics* (Vol. 49, Issue 2, pp. 170–174). Nature Publishing Group. https://doi.org/10.1038/ng.3774

Grimes, C. L., Ariyananda, L. D. Z., Melnyk, J. E., & O'Shea, E. K. (2012). The innate immune protein Nod2 binds directly to MDP, a bacterial cell wall

fragment. *Journal of the American Chemical Society*, *134*(33), 13535–13537. https://doi.org/10.1021/ja303883c

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. In *Genome Biology* (Vol. 20, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1730-3

Gupta, S., Gellert, M., & Yang, W. (2012). Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. *Nature Structural and Molecular Biology*, *19*(1), 72–79. https://doi.org/10.1038/nsmb.2175

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, *3*(10), 1871–1878. https://doi.org/10.1371/journal.pcbi.0030189

Ha, N. T., Freytag, S., & Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. *European Journal of Human Genetics*, *22*(9), 1124–1130. https://doi.org/10.1038/ejhg.2013.304

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017a). Human genetic variation and the gut microbiome in disease. In *Nature Reviews Genetics* (Vol. 18, Issue 11, pp. 690–699). Nature Publishing Group. https://doi.org/10.1038/nrg.2017.63

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017b). Human genetic variation and the gut microbiome in disease. In *Nature Reviews Genetics* (Vol. 18, Issue 11, pp. 690–699). Nature Publishing Group. https://doi.org/10.1038/nrg.2017.63

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., & Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for

pairwise genetic interactions. *Nature Biotechnology*, *35*(5), 463–474. https://doi.org/10.1038/nbt.3834

Hanrahan, J. W., Matthes, E., Carlile, G., & Thomas, D. Y. (2017). Corrector combination therapies for F508del-CFTR. In *Current Opinion in Pharmacology* (Vol. 34, pp. 105–111). Elsevier Ltd. https://doi.org/10.1016/j.coph.2017.09.016

Harding, H. P., Novoa, I., Zhang, Y., Zeng, H., Wek, R., Schapira, M., & Ron, D. (2000). Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Molecular Cell*, *6*(5), 1099–1108. https://doi.org/10.1016/S1097-2765(00)00108-8

Harding, H. P., Zhang, Y., & Ron, D. (1999). Protein translation and folding are coupled by an endoplasmic- reticulum-resident kinase. *Nature*, *397*(6716), 271–274. https://doi.org/10.1038/16729

Harley, J. B., Alarcón-Riquelme, M. E., Criswell, L. A., Jacob, C. O., Kimberly, R. P., Moser, K. L., Tsao, B. P., Vyse, T. J., Langefeld, C. D., Nath, S. K., Guthridge, J. M., Cobb, B. L., Mirel, D. B., Marion, M. C., Williams, A. H., Divers, J., Wang, W., Frank, S. G., Namjou, B., … Kelly, J. A. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nature Genetics*, *40*(2), 204–210. https://doi.org/10.1038/ng.81

Hasnain, S. Z., Tauro, S., Das, I., Tong, H., Chen, A. H., Jeffery, P. L., McDonald, V., Florin, T. H., & McGuckin, M. A. (2013). IL-10 promotes production of intestinal mucus by suppressing protein misfolding and endoplasmic reticulum stress in goblet cells. *Gastroenterology*, *144*(2).

https://doi.org/10.1053/j.gastro.2012.10.043

Häuser, F., Deyle, C., Berard, D., Neukirch, C., Glowacki, C., Bickmann, J. K., Wenzel, J. J., Lackner, K. J., & Rossmann, H. (2012). Macrophage-stimulating protein polymorphism rs3197999 is associated with a gain of function: Implications for inflammatory bowel disease. *Genes and Immunity*, *13*(4), 321–327. https://doi.org/10.1038/gene.2011.88

Hayashi, R., Tsuchiya, K., Fukushima, K., Horita, N., Hibiya, S., Kitagaki, K., Negi, M., Itoh, E., Akashi, T., Eishi, Y., Okada, E., Araki, A., Ohtsuka, K., Fukuda, S., Ohno, H., Okamoto, R., Nakamura, T., Tanaka, S., Chayama, K., & Watanabe, M. (2016). Reduced human α-defensin 6 in noninflamed jejunal tissue of patients with Crohn's disease. *Inflammatory Bowel Diseases*, *22*(5), 1119–1128. https://doi.org/10.1097/MIB.0000000000000707

Hayes, B. (2013). Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1019, pp. 149–169). Methods Mol Biol. https://doi.org/10.1007/978-1-62703-447-0_6

He, C., Kraft, P., Chen, C., Buring, J. E., Paré, G., Hankinson, S. E., Chanock, S. J., Ridker, P. M., Hunter, D. J., & Chasman, D. I. (2009). Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nature Genetics*, *41*(6), 724–728. https://doi.org/10.1038/ng.385

Heazlewood, C. K., Cook, M. C., Eri, R., Price, G. R., Tauro, S. B., Taupin, D., Thornton, D. J., Chin, W. P., Crockford, T. L., Cornall, R. J., Adams, R., Kato, M., Nelms, K. A., Hong, N. A., Florin, T. H. J., Goodnow, C. C., & McGuckin, M. A. (2008). Aberrant mucin assembly in mice causes endoplasmic reticulum

stress and spontaneous inflammation resembling ulcerative colitis. *PLoS Medicine*, *5*(3), 0440–0460. https://doi.org/10.1371/journal.pmed.0050054

Henry, V. J., Goelzer, A., Ferré, A., Fischer, S., Dinh, M., Loux, V., Froidevaux, C., & Fromion, V. (2017). The bacterial interlocked process ONtology (BiPON): A systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of Biomedical Semantics*, *8*(1). https://doi.org/10.1186/s13326-017-0165-6

Hillmer, R. A. (2015). Systems Biology for Biologists. In *PLoS Pathogens* (Vol. 11, Issue 5). Public Library of Science. https://doi.org/10.1371/journal.ppat.1004786

Himmelstein, Daniel S., & Baranzini, S. E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Computational Biology*, *11*(7). https://doi.org/10.1371/journal.pcbi.1004259

Himmelstein, Daniel Scott, Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*. https://doi.org/10.7554/eLife.26726

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Hol, W. G., Halie, L. M., & Sander, C. (1981). Dipoles of the alpha-helix and beta-

sheet: their role in protein folding. *Nature*, *294*(5841), 532–536. http://www.ncbi.nlm.nih.gov/pubmed/7312043

Hu, Z., Yu, C., Furutsuki, M., Andreoletti, G., Ly, M., Hoskins, R., Adhikari, A. N., & Brenner, S. E. (2019). VIPdb, a genetic Variant Impact Predictor Database. *Human Mutation*, *40*(9), 1202–1214. https://doi.org/10.1002/humu.23858

Hucka, M., Bergmann, F. T., Dräger, A., Hoops, S., Keating, S. M., Le Novère, N., Myers, C. J., Olivier, B. G., Sahle, S., Schaff, J. C., Smith, L. P., Waltemath, D., & Wilkinson, D. J. (2018). The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. *Journal of Integrative Bioinformatics*, *15*(1). https://doi.org/10.1515/jib-2017-0081

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., & Thomas, G. (2001a). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, *411*(6837), 599–603. https://doi.org/10.1038/35079107

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., & Thomas, G. (2001b). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, *411*(6837), 599–603. https://doi.org/10.1038/35079107

Hugot, J. P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., Lee, J. C.,

Beaugerie, L., Naom, I., Dupas, J. L., Van Gossum, A., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C. G., Lennard-Jones, J. E., Cortot, A., Colombel, J. F., & Thomas, G. (1996). Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature*, *379*(6568), 821–823. https://doi.org/10.1038/379821a0

Hui, K. Y., Fernandez-Hernandez, H., Hu, J., Schaffner, A., Pankratz, N., Hsu, N. Y., Chuang, L. S., Carmi, S., Villaverde, N., Li, X., Rivas, M., Levine, A. P., Bao, X., Labrias, P. R., Haritunians, T., Ruane, D., Gettler, K., Chen, E., Li, D., … Peter, I. (2018). Functional variants in the LRRK2 gene confer shared effects on risk for Crohn's disease and Parkinson's disease. *Science Translational Medicine*, *10*(423). https://doi.org/10.1126/scitranslmed.aai7795

Hwang, S., Kim, E., Lee, I., Marcotte, E. M., Church, G. M., Lunshof, J. E., Bamshad, M. J., Do, R., Kathiresan, S., Abecasis, G. R., Pereira, P. C. B., Yang, Y., Cirulli, E. T., Goldstein, D. B., Tennessen, J. A., Renkema, K. Y., Stokman, M. F., Giles, R. H., Knoers, N. V., … Karolchik, D. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*, 17875. https://doi.org/10.1038/srep17875

Ideker, T., & Krogan, N. J. (2012). Differential network biology. In *Molecular Systems Biology* (Vol. 8). Mol Syst Biol. https://doi.org/10.1038/msb.2011.99

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456. https://arxiv.org/abs/1502.03167v3

Ison, J., ˇ Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J.,

Lopez, R., Pettifer, S., Rice, P., & Kelso, J. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, *42*(22), 13534–13544. https://doi.org/10.1093/nar/gku1206

Johansson, M. E. V., Ambort, D., Pelaseyed, T., Schütte, A., Gustafsson, J. K., Ermund, A., Subramani, D. B., Holmén-Larsson, J. M., Thomsson, K. A., Bergström, J. H., Van Der Post, S., Rodriguez-Piñeiro, A. M., Sjövall, H., Bäckström, M., & Hansson, G. C. (2011). Composition and functional role of the mucus layers in the intestine. In *Cellular and Molecular Life Sciences* (Vol. 68, Issue 22, pp. 3635–3641). Cell Mol Life Sci. https://doi.org/10.1007/s00018-011-0822-3

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Philip Schumm, L., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., … Whittaker, P. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, *491*(7422), 119–124. https://doi.org/10.1038/nature11582

Kadayakkara, D. K., Beatty, P. L., Turner, M. S., Janjic, J. M., Ahrens, E. T., & Finn, O. J. (2010). Inflammation driven by overexpression of the hypoglycosylated abnormal Mucin 1 (MUC1) links inflammatory bowel disease and pancreatitis.

Pancreas, *39*(4), 510–515. https://doi.org/10.1097/MPA.0b013e3181bd6501

Kambouris, M., Maroun, R. C., Ben-Omran, T., Al-Sarraj, Y., Errafii, K., Ali, R., Boulos, H., Curmi, P. A., El-Shanti, H., Seelow, D., Schuelke, M., Hildebrandt, F., Nürnberg, P., Fiser, A., Modeller, S., Shen, M., Sali, A., Laskowski, R., MacArthur, M., … Fay, J. (2014). Mutations in zinc finger 407 [ZNF407] cause a unique autosomal recessive cognitive impairment syndrome. *Orphanet Journal of Rare Diseases*, *9*(1), 80. https://doi.org/10.1186/1750-1172-9-80

Kametani, F., & Hasegawa, M. (2018). Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease. In *Frontiers in Neuroscience* (Vol. 12, Issue JAN). Frontiers Media S.A. https://doi.org/10.3389/fnins.2018.00025

Kammermeier, J., Drury, S., James, C. T., Dziubak, R., Ocaka, L., Elawad, M., Beales, P., Lench, N., Uhlig, H. H., Bacchelli, C., & Shah, N. (2014). Targeted gene panel sequencing in children with very early onset inflammatory bowel disease--evaluation and prospective analysis. *Journal of Medical Genetics*, *51*(11), 748–755. https://doi.org/10.1136/jmedgenet-2014-102624

Kanehisa, M, & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. http://www.ncbi.nlm.nih.gov/pubmed/10592173

Kanehisa, Minoru, Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*, 353–361. https://doi.org/10.1093/nar/gkw1092

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D.,

Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kaser, A., & Blumberg, R. S. (2014). ATG16L1 Crohn's disease risk stresses the endoplasmic reticulum of Paneth cells. In *Gut* (Vol. 63, Issue 7, pp. 1038–1039). BMJ Publishing Group. https://doi.org/10.1136/gutjnl-2013-306103

Kaser, A., Lee, A. H., Franke, A., Glickman, J. N., Zeissig, S., Tilg, H., Nieuwenhuis, E. E. S., Higgins, D. E., Schreiber, S., Glimcher, L. H., & Blumberg, R. S. (2008). XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease. *Cell*, *134*(5), 743–756. https://doi.org/10.1016/j.cell.2008.07.021

Keeling, M. J. (2005). Models of foot-and-mouth disease. In *Proceedings of the Royal Society B: Biological Sciences* (Vol. 272, Issue 1569, pp. 1195–1202). Royal Society. https://doi.org/10.1098/rspb.2004.3046

Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G., & Lowe, J. B. (1995). Sequence and expression of a candidate for the human Secretor blood group α(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *Journal of Biological Chemistry*, *270*(9), 4640–4649. https://doi.org/10.1074/jbc.270.9.4640

Kendler, K. S., & Neale, M. C. (2009). "Familiality" or Heritability. *Archives of General Psychiatry*, *66*(4), 452.

https://doi.org/10.1001/archgenpsychiatry.2009.14

Kikuchi, M., Ogishima, S., Mizuno, S., Miyashita, A., Kuwano, R., Nakaya, J., & Tanaka, H. (2015). Network-based analysis for uncovering mechanisms underlying Alzheimer's disease. In *Systems Biology of Alzheimer's Disease* (Vol. 1303, pp. 479–491). Springer New York. https://doi.org/10.1007/978-1-4939-2627-5_29

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, *28*(23), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591

Kinoshita, J., & Clark, T. (2007). Alzforum. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 401, pp. 365–381). Methods Mol Biol. https://doi.org/10.1007/978-1-59745-520-6_19

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Knodler, L. A., & Celli, J. (2011). Eating the strangers within: Host control of intracellular bacteria via xenophagy. In *Cellular Microbiology* (Vol. 13, Issue 9, pp. 1319–1327). Cell Microbiol. https://doi.org/10.1111/j.1462-5822.2011.01632.x

Kohl, P., & Noble, D. (2009). Systems biology and the virtual physiological human. In *Molecular Systems Biology* (Vol. 5). Mol Syst Biol.

https://doi.org/10.1038/msb.2009.51

Konopka, T., & Smedley, D. (2020). Incremental data integration for tracking genotype-disease associations. *PLoS Computational Biology*, *16*(1). https://doi.org/10.1371/journal.pcbi.1007586

Korennykh, A. V., Egea, P. F., Korostelev, A. A., Finer-Moore, J., Zhang, C., Shokat, K. M., Stroud, R. M., & Walter, P. (2009). The unfolded protein response signals through high-order assembly of Ire1. *Nature*, *457*(7230), 687–693. https://doi.org/10.1038/nature07661

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(8), 1073–1081. https://doi.org/10.1038/nprot.2009.86

L. Kretschmann, K., Eyob, H., S. Buys, S., & L. Welm, A. (2010). The Macrophage Stimulating Protein/Ron Pathway as a Potential Therapeutic Target to Impede Multiple Mechanisms Involved in Breast Cancer Progression. *Current Drug Targets*, *11*(9), 1157–1168. https://doi.org/10.2174/138945010792006825

Laksshman, S., Bhat, R. R., Viswanath, V., & Li, X. (2017). DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. *Human Mutation*, *38*(9), 1217–1224. https://doi.org/10.1002/humu.23272

Lancaster, M. A., & Huch, M. (2019). Disease modelling in human organoids. *DMM Disease Models and Mechanisms*, *12*(7). https://doi.org/10.1242/dmm.039347

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R.

(2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., Mcdaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*. https://doi.org/10.1093/nar/gkx1153

Lang, W. H., Coats, J. E., Majka, J., Hura, G. L., Lin, Y., Rasnik, I., & McMurray, C. T. (2011). Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(42). https://doi.org/10.1073/pnas.1105461108

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-559

Lécine, P., Esmiol, S., Métais, J. Y., Nicoletti, C., Nourry, C., McDonald, C., Nunez, G., Hugot, J. P., Borg, J. P., & Ollendorff, V. (2007). The NOD2-RICK complex signals from the plasma membrane. *Journal of Biological Chemistry*, *282*(20), 15197–15207. https://doi.org/10.1074/jbc.M606242200

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T.,

Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. In *Genome Medicine* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00742-5

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. In *Nature Reviews Microbiology* (Vol. 10, Issue 4, pp. 291–305). Nature Publishing Group. https://doi.org/10.1038/nrmicro2737

Li, Han, Korennykh, A. V., Behrman, S. L., & Walter, P. (2010). Mammalian endoplasmic reticulum stress sensor IRE1 signals by dynamic clustering. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16113–16118. https://doi.org/10.1073/pnas.1010580107

Li, Heng. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)*, *30*(20), 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, J., Horstman, B., & Chen, Y. (2011). Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics*, *27*(13). https://doi.org/10.1093/bioinformatics/btr227

Li, M., Zhang, S., Qiu, Y., He, Y., Chen, B., Mao, R., Cui, Y., Zeng, Z., & Chen, M. (2017). Upregulation of miR-665 promotes apoptosis and colitis in inflammatory

bowel disease by repressing the endoplasmic reticulum stress components XBP1 and ORMDL3. *Cell Death and Disease*, *8*(3). https://doi.org/10.1038/cddis.2017.76

Li, Y., Cho, H., Wang, F., Canela-Xandri, O., Luo, C., Rawlik, K., Archacki, S., Xu, C., Tenesa, A., Chen, Q., & Wang, Q. K. (2020). Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *Journal of the American Heart Association*, *9*(7), e014146. https://doi.org/10.1161/JAHA.119.014146

Lilyquist, J., Ruddy, K. J., Vachon, C. M., & Couch, F. J. (2018). Common genetic variation and breast cancer Risk—Past, present, and future. In *Cancer Epidemiology Biomarkers and Prevention* (Vol. 27, Issue 4, pp. 380–394). American Association for Cancer Research Inc. https://doi.org/10.1158/1055-9965.EPI-17-1144

Lin, Z., Wang, Z., Hegarty, J. P., Lin, T. R., Wang, Y., Deiling, S., Wu, R., Thomas, N. J., & Floros, J. (2017). Genetic association and epistatic interaction of the interleukin-10 signaling pathway in pediatric inflammatory bowel disease. *World Journal of Gastroenterology*, *23*(27), 4897–4909. https://doi.org/10.3748/wjg.v23.i27.4897

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W. W. L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M. T., Cordeiro, D., Cytrynbaum, C., Dell, S. D., … Marshall, C. R. (2018). Improved diagnostic yield compared with targeted gene sequencing panels

suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine*, *20*(4), 435–443. https://doi.org/10.1038/gim.2017.119

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). Nat Genet. https://doi.org/10.1038/ng.2653

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018a). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298. https://doi.org/10.1038/nmeth.4627

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018b). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298. https://doi.org/10.1038/nmeth.4627

Ma, X., Dai, Z., Sun, K., Zhang, Y., Chen, J., Yang, Y., Tso, P., Wu, G., & Wu, Z. (2017). Intestinal epithelial cell endoplasmic reticulum stress and inflammatory bowel disease pathogenesis: An update review. *Frontiers in Immunology*, *8*(OCT), 1271. https://doi.org/10.3389/fimmu.2017.01271

Madian, A. G., Wheeler, H. E., Jones, R. B., & Dolan, M. E. (2012). Relating human genetic variation to variation in drug responses. In *Trends in Genetics* (Vol. 28, Issue 10, pp. 487–495). Trends Genet. https://doi.org/10.1016/j.tig.2012.06.008

Maher, B. (2008). Personal genomes: The case of the missing heritability. In *Nature* (Vol. 456, Issue 7218, pp. 18–21). Nature Publishing Group. https://doi.org/10.1038/456018a

Malhotra, A., Younesi, E., Gündel, M., Müller, B., Heneka, M. T., & Hofmann-Apitius, M. (2014). ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & Dementia*, *10*(2), 238–246. https://doi.org/10.1016/j.jalz.2013.02.009

Mallah, N., Zapata-Cachafeiro, M., Aguirre, C., Ibarra-García, E., Palacios–Zabalza, I., Macías-García, F., Domínguez-Muñoz, J. E., Piñeiro-Lamas, M., Ibáñez, L., Vidal, X., Vendrell, L., Martin-Arias, L., Sáinz-Gil, M., Velasco-González, V., & Figueiras, A. (2020). Polymorphisms Involved in Platelet Activation and Inflammatory Response on Aspirin-Related Upper Gastrointestinal Bleeding: A Case-Control Study. *Frontiers in Pharmacology*, *11*. https://doi.org/10.3389/fphar.2020.00860

Mallott, J., Kwan, A., Church, J., Gonzalez-Espinosa, D., Lorey, F., Tang, L. F., Sunderam, U., Rana, S., Srinivasan, R., Brenner, S. E., & Puck, J. (2013). Newborn screening for SCID identifies patients with ataxia telangiectasia. *Journal of Clinical Immunology*, *33*(3), 540–549. https://doi.org/10.1007/s10875-012-9846-1

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. In *Nature* (Vol. 461, Issue 7265, pp. 747–753). Nature. https://doi.org/10.1038/nature08494

Marín de Mas, I., Aguilar, E., Zodda, E., Balcells, C., Marin, S., Dallmann, G.,

Thomson, T. M., Papp, B., & Cascante, M. (2018). Model-driven discovery of long-chain fatty acid metabolic reprogramming in heterogeneous prostate cancer cells. *PLoS Computational Biology*, *14*(1). https://doi.org/10.1371/journal.pcbi.1005914

Martín-López, J. V., & Fishel, R. (2013). The mechanism of mismatch repair and the functional analysis of mismatch repair defects in Lynch syndrome. *Familial Cancer*, *12*(2), 159–168. https://doi.org/10.1007/s10689-013-9635-x

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U. S., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., Thomas, E., Scott, R. H., Baple, E., Tucci, A., Brittain, H., de Burca, A., Ibañez, K., Kasperaviciute, D., Smedley, D., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. In *Nature Genetics* (Vol. 51, Issue 11, pp. 1560–1565). Nature Publishing Group. https://doi.org/10.1038/s41588-019-0528-2

Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., Nogueira, M. M., He, F., Parton, A., Lemonnier, N., Gawron, P., Gebel, S., Hainaut, P., Ollert, M., Dogrusoz, U., … Auffray, C. (2018). Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *Npj Systems Biology and Applications*, *4*(1). https://doi.org/10.1038/s41540-018-0059-y

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., Donnelly, P., Green, E., Guyer, M., Schrijver, I., Aziz, N., Farkas, D., Furtado,

M., Gonzalez, A., Greiner, T., Grody, W., Hambuch, T., Kalman, L., Kant, J., … Trajanoski, Z. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, *6*(3), 26. https://doi.org/10.1186/gm543

Mcgovern, D. P. B., Jones, M. R., Taylor, K. D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., Dutridge, D., Fleshner, P., Shih, D. Q., Melmed, G., Mengesha, E., King, L., Pressman, S., Haritunians, T., Guo, X., … Rotter, J. I. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Human Molecular Genetics*, *19*(17), 3468–3476. https://doi.org/10.1093/hmg/ddq248

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., … Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527–1541. https://doi.org/10.1101/gr.091868.109

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., Cunningham, F., Eisenstein, M., Weil, M., Chen, A., Visscher, P., Brown, M., McCarthy, M., Yang, J., Pierre, A. Saint, Génin, E., Zuk, O., Schaffner, S., … Liu, X. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Mendez, A. S., Alfaro, J., Morales-Soto, M. A., Dar, A. C., McCullagh, E., Gotthardt, K., Li, H., Acosta-Alvear, D., Sidrauski, C., Korennykh, A. V., Bernales, S.,

Shokat, K. M., & Walter, P. (2015). Endoplasmic reticulum stress-independent activation of unfolded protein response kinases by a small molecule ATP-mimic. *ELife*, *4*(MAY). https://doi.org/10.7554/eLife.05434

Meryhew, N. L., Kimberly, R. P., Messner, R. P., & Runquist, O. A. (1986). Mononuclear phagocyte system in SLE. II. A kinetic model of immune complex handling in systemic lupus erythematosus. *The Journal of Immunology*, *137*(1).

Mina, E., Thompson, M., Kaliyaperumal, R., Zhao, J., van der Horst, E., Tatum, Z., Hettne, K. M., Schultes, E. A., Mons, B., & Roos, M. (2015). Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study. *Journal of Biomedical Semantics*, *6*(1), 5. https://doi.org/10.1186/2041-1480-6-5

Miskovic, L., Tokic, M., Fengos, G., & Hatzimanikatis, V. (2015). Rites of passage: Requirements and standards for building kinetic models of metabolic phenotypes. In *Current Opinion in Biotechnology* (Vol. 36, pp. 146–153). Elsevier Ltd. https://doi.org/10.1016/j.copbio.2015.08.019

Moehle, C., Ackermann, N., Langmann, T., Aslanidis, C., Kel, A., Kel-Margoulis, O., Schmitz-Madry, A., Zahn, A., Stremmel, W., & Schmitz, G. (2006). Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *Journal of Molecular Medicine*, *84*(12), 1055–1066. https://doi.org/10.1007/s00109-006-0100-2

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., Von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S. A. G., Wong, K. C. C., Illig, T., Vogelberg, C., Weiland, S.

K., Von Mutius, E., … Cookson, W. O. C. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, *448*(7152), 470–473. https://doi.org/10.1038/nature06014

Mohanan, V., Nakata, T., Desch, A. N., Lévesque, C., Boroughs, A., Guzman, G., Cao, Z., Creasey, E., Yao, J., Boucher, G., Charron, G., Bhan, A. K., Schenone, M., Carr, S. A., Reinecker, H. C., Daly, M. J., Rioux, J. D., Lassen, K. G., & Xavier, R. J. (2018). C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science*, *359*(6380), 1161–1166. https://doi.org/10.1126/science.aan0814

Monteleone, G., Caruso, R., & Pallone, F. (2012). Role of Smad7 in inflammatory bowel diseases. *World Journal of Gastroenterology*, *18*(40), 5664–5668. https://doi.org/10.3748/wjg.v18.i40.5664

Monteleone, G., Neurath, M. F., Ardizzone, S., Di Sabatino, A., Fantini, M. C., Castiglione, F., Scribano, M. L., Armuzzi, A., Caprioli, F., Sturniolo, G. C., Rogai, F., Vecchi, M., Atreya, R., Bossa, F., Onali, S., Fichera, M., Corazza, G. R., Biancone, L., Savarino, V., … Pallone, F. (2015). Mongersen, an oral SMAD7 antisense oligonucleotide, and crohn's disease. *New England Journal of Medicine*, *372*(12), 1104–1113. https://doi.org/10.1056/NEJMoa1407250

Morosky, S. A., Zhu, J., Mukherjee, A., Sarkar, S. N., & Coyne, C. B. (2011). Retinoic acid-induced gene-I (RIG-I) associates with nucleotide-binding oligomerization domain-2 (NOD2) to negatively regulate inflammatory signaling. *Journal of Biological Chemistry*, *286*(32), 28574–28583. https://doi.org/10.1074/jbc.M111.227942

Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N., Listgarten, J., Younger, S. T., Bernstein, B. E., Root, D. E., & Doench, J. G. (2018). Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. *Nature Biotechnology*, *36*(2), 179–189. https://doi.org/10.1038/nbt.4048

Negroni, A., Pierdomenico, M., Cucchiara, S., & Stronati, L. (2018). NOD2 and inflammation: Current insights. In *Journal of Inflammation Research* (Vol. 11, pp. 49–60). Dove Medical Press Ltd. https://doi.org/10.2147/JIR.S137606

Neurath, M. F. (2014). New targets for mucosal healing and therapy in inflammatory bowel diseases. In *Mucosal Immunology* (Vol. 7, Issue 1, pp. 6–19). Mucosal Immunol. https://doi.org/10.1038/mi.2013.73

Novère, N. Le, Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., … Kitano, H. (2009a). The Systems Biology Graphical Notation. In *Nature Biotechnology* (Vol. 27, Issue 8, pp. 735–741). Nat Biotechnol. https://doi.org/10.1038/nbt.1558

Novère, N. Le, Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., … Kitano, H. (2009b). The Systems Biology Graphical Notation. *Nature Biotechnology*, *27*(8), 735–741. https://doi.org/10.1038/nbt.1558

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y. Y.,

Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., & Topper, S. (2017). Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine*, *19*(10), 1105–1117. https://doi.org/10.1038/gim.2017.37

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nũez, G., & Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, *411*(6837), 603–606. https://doi.org/10.1038/35079114

Okazaki, T., Murata, M., Kai, M., Adachi, K., Nakagawa, N., Kasagi, ‖noriko, Matsumura, W., Maegaki, Y., & Nanba, E. (2016). Clinical Diagnosis of Mendelian Disorders Using a Comprehensive Gene-Targeted Panel Test for Next-Generation Sequencing. *Yonago Acta Medica*, *59*, 118–125.

Okumura, R., & Takeda, K. (2018). Maintenance of intestinal homeostasis by mucosal barriers. In *Inflammation and Regeneration* (Vol. 38, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s41232-018-0063-z

Onge, R. P. S., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R. W., Nislow, C., Roth, F. P., & Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics*, *39*(2), 199–206. https://doi.org/10.1038/ng1948

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data.

Briefings in Bioinformatics, *15*(2), 256–278. https://doi.org/10.1093/bib/bbs086

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., & Baralle, F. E. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Human Molecular Genetics*, *12*(10), 1111–1120. https://doi.org/10.1093/hmg/ddg131

Pal, L. R., Chao, K. L., Moult, J., & Herzberg, O. (2017). On the interpretation of gasdermin-B expression quantitative trait loci data. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 114, Issue 38, pp. E7863–E7864). National Academy of Sciences. https://doi.org/10.1073/pnas.1712734114

Pal, L. R., Kundu, K., Yin, Y., & Moult, J. (2017). CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. *Human Mutation*, *38*(9), 1225–1234. https://doi.org/10.1002/humu.23256

Pal, L. R., Yu, C. H., Mount, S. M., & Moult, J. (2015). Insights from GWAS: Emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics*, *16*(8). https://doi.org/10.1186/1471-2164-16-S8-S4

Pansarasa, O., Bordoni, M., Drufuca, L., Diamanti, L., Sproviero, D., Trotti, R., Bernuzzi, S., Salvia, S. La, Gagliardi, S., Ceroni, M., & Cereda, C. (2018). Lymphoblastoid cell lines as a model to understand amyotrophic lateral sclerosis disease mechanisms. In *DMM Disease Models and Mechanisms* (Vol. 11, Issue 3). Company of Biologists Ltd. https://doi.org/10.1242/dmm.031625

Paone, P., & Cani, P. D. (2020). Mucus barrier, mucins and gut microbiota: the

expected slimy partners? *Gut*, gutjnl-2020-322260. https://doi.org/10.1136/gutjnl-2020-322260

Park, J.-H., Kim, Y.-G., McDonald, C., Kanneganti, T.-D., Hasegawa, M., Body-Malapel, M., Inohara, N., & Núñez, G. (2007). RICK/RIP2 Mediates Innate Immune Responses Induced through Nod1 and Nod2 but Not TLRs. *The Journal of Immunology*, *178*(4), 2380–2386. https://doi.org/10.4049/jimmunol.178.4.2380

Patel, J. P., Puck, J. M., Srinivasan, R., Brown, C., Sunderam, U., Kundu, K., Brenner, S. E., Gatti, R. A., & Church, J. A. (2015a). Nijmegen Breakage Syndrome Detected by Newborn Screening for T Cell Receptor Excision Circles (TRECs). *Journal of Clinical Immunology*, *35*(2), 227. https://doi.org/10.1007/s10875-015-0136-6

Patel, J. P., Puck, J. M., Srinivasan, R., Brown, C., Sunderam, U., Kundu, K., Brenner, S. E., Gatti, R. A., & Church, J. A. (2015b). Nijmegen breakage syndrome detected by newborn screening for T cell receptor excision circles (TRECs). *Journal of Clinical Immunology*, *35*(2), 227–233. https://doi.org/10.1007/s10875-015-0136-6

PDBe-KB consortium. (2020). PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Research*, *48*(Database issue). https://doi.org/10.1093/nar/gkz853

Pegoraro, G., & Misteli, T. (2017). High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. In *Trends in Genetics* (Vol. 33, Issue 9, pp. 604–615). Elsevier Ltd. https://doi.org/10.1016/j.tig.2017.06.005

Peter, I., Mitchell, A. A., Ozelius, L., Erazo, M., Hu, J., Doheny, D., Abreu, M. T., Present, D. H., Ullman, T., Benkov, K., Korelitz, B. I., Mayer, L., & Desnick, R. J. (2011). Evaluation of 22 genetic variants with Crohn's Disease risk in the Ashkenazi Jewish population: a case-control study. *BMC Medical Genetics*, *12*. https://doi.org/10.1186/1471-2350-12-63

Phillips, P. C. (2008). Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. In *Nature Reviews Genetics* (Vol. 9, Issue 11, pp. 855–867). Nat Rev Genet. https://doi.org/10.1038/nrg2452

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., Zandi, P. P., Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M., Smith, S., Middle, C., Rodesch, M., Albert, T., Hannon, G., McCombie, W., Henson, J., … Sham, P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14. https://doi.org/10.1186/1479-7364-8-14

Pon, A., Jewison, T., Su, Y., Liang, Y., Knox, C., Maciejewski, A., Wilson, M., & Wishart, D. S. (2015). Pathways with PathWhiz. *Nucleic Acids Research*, *43*. https://doi.org/10.1093/nar/gkv399

Posey, J. E. (2019). Genome sequencing and implications for rare disorders. In *Orphanet Journal of Rare Diseases* (Vol. 14, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13023-019-1127-0

Posey, J. E., O'Donnell-Luria, A. H., Chong, J. X., Harel, T., Jhangiani, S. N., Coban Akdemir, Z. H., Buyske, S., Pehlivan, D., Carvalho, C. M. B., Baxter, S., Sobreira, N., Liu, P., Wu, N., Rosenfeld, J. A., Kumar, S., Avramopoulos, D.,

White, J. J., Doheny, K. F., Witmer, P. D., … Lupski, J. R. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. In *Genetics in Medicine* (Vol. 21, Issue 4, pp. 798–812). Nature Publishing Group. https://doi.org/10.1038/s41436-018-0408-7

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756-63. https://doi.org/10.1093/nar/gkt1114

Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal*, *38*(10), 1499–1511. https://doi.org/10.1002/aic.690381003

Punwani, D., Zhang, Y., Yu, J., Cowan, M. J., Rana, S., Kwan, A., Adhikari, A. N., Lizama, C. O., Mendelsohn, B. A., Fahl, S. P., Chellappan, A., Srinivasan, R., Brenner, S. E., Wiest, D. L., & Puck, J. M. (2016). Multisystem Anomalies in Severe Combined Immunodeficiency with Mutant BCL11B. *New England Journal of Medicine*, *375*(22), 2165–2176. https://doi.org/10.1056/NEJMoa1509164

Ramos, M. P. M., Ribeiro, C., & Soares, A. J. (2019). A kinetic model of T cell autoreactivity in autoimmune diseases. *Journal of Mathematical Biology*, *79*(6–7), 2005–2031. https://doi.org/10.1007/s00285-019-01418-4

Resat, H., Petzold, L., & Pettigrew, M. F. (2009). Kinetic modeling of biological

systems. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 541, pp. 311–335). Methods Mol Biol. https://doi.org/10.1007/978-1-59745-243-4_14

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015a). *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. https://doi.org/10.1038/gim.2015.30

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015b). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine, 17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Ridley, K., & Condren, M. (2020). Elexacaftor-tezacaftor-ivacaftor: The first triple-combination cystic fibrosis transmembrane conductance regulator modulating therapy. *Journal of Pediatric Pharmacology and Therapeutics, 25*(3), 192–197. https://doi.org/10.5863/1551-6776-25.3.192

Robinson, P. N., Kohler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., & Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research, 24*(2), 340–348. https://doi.org/10.1101/gr.160325.113

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., & Tress, M. L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, *41*(D1), D110–D117. https://doi.org/10.1093/nar/gks1058

Sadedin, S. P., Dashnow, H., James, P. A., Bahlo, M., Bauer, D. C., Lonie, A., Lunke, S., Macciocca, I., Ross, J. P., Siemering, K. R., Stark, Z., White, S. M., Taylor, G., Gaff, C., Oshlack, A., & Thorne, N. P. (2015). Cpipe: A shared variant detection pipeline designed for diagnostic settings. *Genome Medicine*, *7*(1). https://doi.org/10.1186/s13073-015-0191-x

Salem, M., Ammitzboell, M., Nys, K., Seidelin, J. B., & Nielsen, O. H. (2015). ATG16L1: A multifunctional susceptibility factor in crohn disease. In *Autophagy* (Vol. 11, Issue 4, pp. 585–594). Taylor and Francis Inc. https://doi.org/10.1080/15548627.2015.1017187

Salomon, M. P., Li, W. L. S., Edlund, C. K., Morrison, J., Fortini, B. K., Win, A. K., Conti, D. V., Thomas, D. C., Duggan, D., Buchanan, D. D., Jenkins, M. A., Hopper, J. L., Gallinger, S., Le Marchand, L., Newcomb, P. A., Casey, G., & Marjoram, P. (2016). GWASeq: Targeted re-sequencing follow up to GWAS. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-2459-y

Sands, B. E., Feagan, B. G., Sandborn, W. J., Schreiber, S., Peyrin-Biroulet, L., Colombel, J. F., Rossiter, G., Usiskin, K., Ather, S., Zhan, X., & D'Haens, G. (2020). Mongersen (GED-0301) for active Crohn's disease: Results of a phase 3 study. *American Journal of Gastroenterology*, *115*(5), 738–745. https://doi.org/10.14309/ajg.0000000000000493

Santoso, J. W., & McCain, M. L. (2020). Neuromuscular disease modeling on a chip. In *DMM Disease Models and Mechanisms* (Vol. 13, Issue 7). Company of Biologists Ltd. https://doi.org/10.1242/dmm.044867

Sarkar, A., Duncan, M., Hart, J., Hertlein, E., Guttridge, D. C., & Wewers, M. D. (2006). ASC Directs NF-κB Activation by Regulating Receptor Interacting Protein-2 (RIP2) Caspase-1 Interactions. *The Journal of Immunology*, *176*(8), 4979–4986. https://doi.org/10.4049/jimmunol.176.8.4979

Sarkar, J., Dwivedi, G., Chen, Q., Sheu, I. E., Paich, M., Chelini, C. M., D'Alessandro, P. M., & Burns, S. P. (2019). A long-Term mechanistic computational model of physiological factors driving the onset of type 2 diabetes in an individual. *PLoS ONE*, *13*(2). https://doi.org/10.1371/journal.pone.0192472

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. In *Nature Reviews Genetics* (Vol. 19, Issue 8, pp. 491–504). Nature Publishing Group. https://doi.org/10.1038/s41576-018-0016-z

Schroeder, B. O. (2019). Fight them or feed them: How the intestinal mucus layer manages the gut microbiota. In *Gastroenterology Report* (Vol. 7, Issue 1, pp. 3–12). Oxford University Press. https://doi.org/10.1093/gastro/goy052

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., … Hayes, V. M. (2010). Complete Khoisan and Bantu genomes from

southern Africa. *Nature*, *463*(7283), 943–947. https://doi.org/10.1038/nature08795

Sehgal, R., Sheahan, K., O'Connell, P. R., Hanly, A. M., Martin, S. T., & Winter, D. C. (2014). Lynch Syndrome: An updated review. *Genes*, *5*(3), 497–507. https://doi.org/10.3390/genes5030497

Shah, N., Hou, Y. C. C., Yu, H. C., Sainger, R., Caskey, C. T., Venter, J. C., & Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American Journal of Human Genetics*, *102*(4), 609–619. https://doi.org/10.1016/j.ajhg.2018.02.019

Shamseldin, H. E., Maddirevula, S., Faqeih, E., Ibrahim, N., Hashem, M., Shaheen, R., & Alkuraya, F. S. (2017). Increasing the sensitivity of clinical exome sequencing through improved filtration strategy. *Genetics in Medicine*, *19*(5), 593–598. https://doi.org/10.1038/gim.2016.155

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shannon, Paul, Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A.,

Kuo, C. C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., … Mali, P. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, *14*(6), 573–576. https://doi.org/10.1038/nmeth.4225

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shimaoka, I., Kamide, K., Ohishi, M., Katsuya, T., Akasaka, H., Saitoh, S., Sugimoto, K., Oguro, R., Congrains, A., Fujisawa, T., Shimamoto, K., Ogihara, T., & Rakugi, H. (2010). Association of gene polymorphism of the fat-mass and obesity-associated gene with insulin resistance in Japanese. *Hypertension Research*, *33*(3), 214–218. https://doi.org/10.1038/hr.2009.215

Sidiq, T., Yoshihama, S., Downs, I., & Kobayashi, K. S. (2016). Nod2: A critical regulator of ileal microbiota and Crohn's disease. In *Frontiers in Immunology* (Vol. 7, Issue SEP). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2016.00367

Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., & Moreau, Y. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, *10*(11), 1083–1084. https://doi.org/10.1038/nmeth.2656

Simms, L. A., Doecke, J. D., Walsh, M. D., Huang, N., Fowler, E. V., & Radford-Smith, G. L. (2008). Reduced α-defensin expression is associated with inflammation and not NOD2 mutation status in ileal Crohn's disease. *Gut*, *57*(7),

903–910. https://doi.org/10.1136/gut.2007.142588

Sioutos, N., Coronado, S. de, Haber, M. W., Hartel, F. W., Shaiu, W. L., & Wright, L. W. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, *40*(1), 30–43. https://doi.org/10.1016/j.jbi.2006.02.013

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *The American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puviindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H. K., Naranjo, S., Acemel, R. D., … Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–375. https://doi.org/10.1038/nature13138

Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., LePichon, J. B., Miller, N. A., Thiffault, I., Dinwiddie, D. L., Twist, G., Noll, A., Heese, B. A., Zellmer, L., Atherton, A. M., Abdelmoity, A. T., Safina, N., Nyp, S. S., Zuccarelli, B., … Kingsmore, S. F. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science Translational Medicine*, *6*(265).

https://doi.org/10.1126/scitranslmed.3010076

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., Boone, C., & Andrews, B. (2006). Mapping pathways and phenotypes by systematic gene overexpression. *Molecular Cell*, *21*(3), 319–330. https://doi.org/10.1016/j.molcel.2005.12.011

Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., & Altman, R. B. (2019). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Biocomputing 2020*, 463–474. https://doi.org/10.1142/9789811215636_0041

Soskic, B., Cano-Gamez, E., Smyth, D. J., Rowan, W. C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D. F., Larminie, C. G. C., Bronson, P. G., Willé, D., & Trynka, G. (2019). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics*, *51*(10), 1486–1493. https://doi.org/10.1038/s41588-019-0493-9

Statement, I. C. (2000). Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment. *American Journal of Respiratory and Critical Care Medicine*, *161*(2), 646–664. https://doi.org/10.1164/ajrccm.161.2.ats3-00

Stavropoulos, D. J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R. K. C., Szego, M. J., Hayeems, R. Z., Shaul, R. Z., Brudno, M., Girdea, M., Frey, B., Alipanahi, B., Ahmed, S., Babul-Hirji, R., Porras, R. B., … Pinto, D. (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *Npj Genomic Medicine*, *1*, 15012.

https://doi.org/10.1038/npjgenmed.2015.12

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 Update. *Human Mutation*, *21*(6), 577–581. https://doi.org/10.1002/humu.10212

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., & Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. In *Human Genetics* (Vol. 136, Issue 6, pp. 665–677). Springer Verlag. https://doi.org/10.1007/s00439-017-1779-6

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7). https://doi.org/10.1371/journal.pbio.1002195

Stringer, S., Wray, N. R., Kahn, R. S., & Derks, E. M. (2011). Underestimated effect sizes in GWAS: Fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027964

Strober, W., & Watanabe, T. (2011). NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. In *Mucosal Immunology* (Vol. 4, Issue 5, pp. 484–495). Mucosal Immunol. https://doi.org/10.1038/mi.2011.29

Strober, Warren, Asano, N., Fuss, I., Kitani, A., & Watanabe, T. (2014). Cellular and

molecular mechanisms underlying NOD2 risk-associated polymorphisms in Crohn's disease. In *Immunological Reviews* (Vol. 260, Issue 1, pp. 249–260). Blackwell Publishing Ltd. https://doi.org/10.1111/imr.12193

Strom, S. P., Lee, H., Das, K., Vilain, E., Nelson, S. F., Grody, W. W., & Deignan, J. L. (2014). Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genetics in Medicine, 16*(7), 510–515. https://doi.org/10.1038/gim.2013.183

Süel, G. (2011). Use of Fluorescence Microscopy to Analyze Genetic Circuit Dynamics. In *Methods in enzymology* (Vol. 497, pp. 275–293). Methods Enzymol. https://doi.org/10.1016/b978-0-12-385075-1.00013-5

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Jensen, L. J., & Von Mering, C. (2018). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research, 47*, 607–613. https://doi.org/10.1093/nar/gky1131

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. In *Epigenetics and Chromatin* (Vol. 8, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13072-015-0050-4

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C.,

Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J. B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., … McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, *47*(7), 717–726. https://doi.org/10.1038/ng.3304

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(Database issue). https://doi.org/10.1093/nar/gky1055

Thomas, P. D., Hill, D. P., Mi, H., Osumi-Sutherland, D., Van Auken, K., Carbon, S., Balhoff, J. P., Albou, L. P., Good, B., Gaudet, P., Lewis, S. E., & Mungall, C. J. (2019). Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. In *Nature Genetics* (Vol. 51, Issue 10, pp. 1429–1433). Nature Publishing Group. https://doi.org/10.1038/s41588-019-0500-1

Thul, P. J., & Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science*, *27*(1), 233–244. https://doi.org/10.1002/pro.3307

Tiwary, B. K. (2020). Computational medicine: Quantitative modeling of complex diseases. In *Briefings in Bioinformatics* (Vol. 21, Issue 2, pp. 429–440). Oxford

University Press. https://doi.org/10.1093/bib/bbz005

Travassos, L. H., Carneiro, L. A. M., Ramjeet, M., Hussey, S., Kim, Y. G., Magalhes, J. G., Yuan, L., Soares, F., Chea, E., Le Bourhis, L., Boneca, I. G., Allaoui, A., Jones, N. L., Nũez, G., Girardin, S. E., & Philpott, D. J. (2010). Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nature Immunology*, *11*(1), 55–62. https://doi.org/10.1038/ni.1823

Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., & Katagiri, F. (2009). Network properties of robust immunity in plants. *PLoS Genetics*, *5*(12). https://doi.org/10.1371/journal.pgen.1000772

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *237*(641), 37–72. https://doi.org/10.1098/rstb.1952.0012

Umar, A., Buermeyer, A. B., Simon, J. A., Thomas, D. C., Clark, A. B., Liskay, R. M., & Kunkel, T. A. (1996). Requirement for PCNA in DNA Mismatch Repair at a Step Preceding DNA Resynthesis. *Cell*, *87*(1), 65–73. https://doi.org/10.1016/S0092-8674(00)81323-9

Underhill, D. M. (2007). Collaboration between the innate immune receptors dectin-1, TLRs, and Nods. In *Immunological Reviews* (Vol. 219, Issue 1, pp. 75–87). Immunol Rev. https://doi.org/10.1111/j.1600-065X.2007.00548.x

van Rooij, J., Arp, P., Broer, L., Verlouw, J., van Rooij, F., Kraaij, R., Uitterlinden, A., & Verkerk, A. J. M. H. (2020). Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar

classification over time. *Genetics in Medicine*. https://doi.org/10.1038/s41436-020-0900-8

Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew, I., Uhlén, M., Pedersen, B. K., & Nielsen, J. (2015). Proteome- and Transcriptome-Driven Reconstruction of the Human Myocyte Metabolic Network and Its Use for Identification of Markers for Diabetes. *Cell Reports*, *11*(6), 921–933. https://doi.org/10.1016/j.celrep.2015.04.010

Vihinen, M. (2014). Variation Ontology for annotation of variation effects and mechanisms. *Genome Research*, *24*(2), 356–364. https://doi.org/10.1101/gr.157495.113

Vilar, E., Mork, M. E., Cuddy, A., Borras, E., Bannon, S. A., Taggart, M. W., Ying, J., Broaddus, R. R., Luthra, R., Rodriguez-Bigas, M. A., Lynch, P. M., & You, Y. Q. N. (2014). Role of microsatellite instability-low as a diagnostic biomarker of Lynch syndrome in colorectal cancer. *Cancer Genetics*, *207*(10–12), 495–502. https://doi.org/10.1016/j.cancergen.2014.10.002

Vilela, E. G., da Gama Torres, H. O., Martin, F. P., de Lourdes de Abreu Ferrari, M., Andrade, M. M., & da Cunha, A. S. (2012). Evaluation of inflammatory activity in Crohn's disease and ulcerative colitis. *World Journal of Gastroenterology*, *18*(9), 872–881. https://doi.org/10.3748/wjg.v18.i9.872

Vincent, A. L., Jordan, C. A., Cadzow, M. J., Merriman, T. R., McGhee, C. N., YS., R., WJ, C., HY, P., JM, L., CB, C., KR, D., CA, J., H, O., Y, W., H, T., J, F., J, Z., R, S., ND, G., … A, A. (2014). Mutations in the Zinc Finger Protein Gene, *ZNF469* , Contribute to the Pathogenesis of Keratoconus. *Investigative*

Opthalmology & Visual Science, *55*(9), 5629. https://doi.org/10.1167/iovs.14-14532

Visser, U., Abeyruwan, S., Vempati, U., Smith, R. P., Lemmon, V., & Schürer, S. C. (2011). BioAssay Ontology (BAO): A semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-257

Vivian-Griffiths, T., Baker, E., Schmidt, K. M., Bracher-Smith, M., Walters, J., Artemiou, A., Holmans, P., O'Donovan, M. C., Owen, M. J., Pocklington, A., & Escott-Price, V. (2019). Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *180*(1), 80–85. https://doi.org/10.1002/ajmg.b.32705

Voss, E., Wehkamp, J., Wehkamp, K., Stange, E. F., Schröder, J. M., & Harder, J. (2006). NOD2/CARD15 mediates induction of the antimicrobial peptide human beta-defensin-2. *Journal of Biological Chemistry*, *281*(4), 2005–2011. https://doi.org/10.1074/jbc.M511044200

Vu, V., Verster, A. J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G. T., Moffat, J., & Fraser, A. G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell*, *162*(2), 391–402. https://doi.org/10.1016/j.cell.2015.06.037

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, M. H., Zhou, Y. Q., & Chen, Y. Q. (2002). Macrophage-stimulating protein and RON receptor tyrosine kinase: Potential regulators of macrophage inflammatory activities. In *Scandinavian Journal of Immunology* (Vol. 56, Issue 6, pp. 545–553). Scand J Immunol. https://doi.org/10.1046/j.1365-3083.2002.01177.x

Wang, X. R., & Li, C. (2014). Decoding F508del misfolding in cystic fibrosis. In *Biomolecules* (Vol. 4, Issue 2, pp. 498–509). Biomolecules. https://doi.org/10.3390/biom4020498

Wang, Yan, & Wang, J.-G. (2018). Genome-Wide Association Studies of Hypertension and Several Other Cardiovascular Diseases. *Pulse*, *6*(3–4), 169–186. https://doi.org/10.1159/000496150

Wang, Yinhua, Díaz Arenas, C., Stoebel, D. M., & Cooper, T. F. (2013). Genetic background affects epistatic interactions between two beneficial mutations. *Biology Letters*, *9*(1). https://doi.org/10.1098/rsbl.2012.0328

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–270. https://doi.org/10.1002/humu.22

Wehkamp, J., Koslowski, M., Wang, G., & Stange, E. F. (2008). Barrier dysfunction due to distinct defensin deficiencies in small intestinal and colonic crohn' s disease. In *Mucosal Immunology* (Vol. 1, pp. 67–74). Mucosal Immunol. https://doi.org/10.1038/mi.2008.48

Wehkamp, Jan, Salzman, N. H., Porter, E., Nuding, S., Weichenthal, M., Petras, R. E., Shen, B., Schaeffeler, E., Schwab, M., Linzmeier, R., Feathers, R. W., Chu, H., Lima, H., Fellermann, K., Ganz, T., Stange, E. F., & Bevins, C. L. (2005).

Reduced Paneth cell α-defensins in ileal Crohn's disease. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(50), 18129–18134. https://doi.org/10.1073/pnas.0505256102

Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. In *Nature Reviews Genetics* (Vol. 15, Issue 11, pp. 722–733). Nature Publishing Group. https://doi.org/10.1038/nrg3747

Wek, R. C., Jiang, H.-Y., & Anthony, T. G. (2006). Coping with stress: eIF2 kinases and translational control. *Biochemical Society Transactions*, *34*(1), 7. https://doi.org/10.1042/bst20060007

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., Sanna, S., … Hirschhorn, J. N. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, *41*(1), 25–34. https://doi.org/10.1038/ng.287

Wilson, S. S., Tocchi, A., Holly, M. K., Parks, W. C., & Smith, J. G. (2015). A small intestinal organoid model of non-invasive enteric pathogen-epithelial cell interactions. *Mucosal Immunology*, *8*(2), 352–361. https://doi.org/10.1038/mi.2014.72

Wolfert, M. A., Murray, T. F., Boons, G. J., & Moore, J. N. (2002). The origin of the synergistic effect of muramyl dipeptide with endotoxin and peptidoglycan. *Journal of Biological Chemistry*, *277*(42), 39179–39186. https://doi.org/10.1074/jbc.M204885200

Yamamoto-Furusho, J. K., Barnich, N., Hisamatsu, T., & Podolsky, D. K. (2010). MDP-NOD2 stimulation induces HNP-1 secretion, which contributes to NOD2 antibacterial function. *Inflammatory Bowel Diseases*, *16*(5), 736–742. https://doi.org/10.1002/ibd.21144

Yamamoto, S., & Ma, X. (2009). Role of Nod2 in the development of Crohn's disease. In *Microbes and Infection* (Vol. 11, Issue 12, pp. 912–918). Microbes Infect. https://doi.org/10.1016/j.micinf.2009.06.005

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017a). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 *NAGLU* (Human N-acetyl-glucosaminidase) and *UBE2I* (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017b). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Younesi, E., Malhotra, A., Gündel, M., Scordis, P., Kodamullil, A. T., Page, M., Müller, B., Springstubbe, S., Wüllner, U., Scheller, D., & Hofmann-Apitius, M. (2015). PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. *Theoretical Biology & Medical Modelling*, *12*, 20. https://doi.org/10.1186/s12976-015-0017-y

Yuan, J., & Berg, H. C. (2013). Ultrasensitivity of an adaptive bacterial motor.

Journal of Molecular Biology, *425*(10), 1760–1764. https://doi.org/10.1016/j.jmb.2013.02.016

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, *353*(2), 459–473. https://doi.org/10.1016/j.jmb.2005.08.020

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, *7*(1), 166. https://doi.org/10.1186/1471-2105-7-166

Zawistowski, M., Reppell, M., Wegmann, D., St Jean, P. L., Ehm, M. G., Nelson, M. R., Novembre, J., & Zöllner, S. (2014). Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *European Journal of Human Genetics*, *22*(9), 1137–1144. https://doi.org/10.1038/ejhg.2013.297

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., … Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, *40*(5), 638–645. https://doi.org/10.1038/ng.120

Zeigler, V. L., Gillette, P. C., Crawford, F. A., Wiles, H. B., & Fyfe, D. A. (1990). New approaches to treatment of incessant ventricular tachycardia in the very young. *Journal of the American College of Cardiology*, *16*(3), 681–685.

https://doi.org/10.1016/0735-1097(90)90360-2

Zhang, Y., Fan, H., Xu, J., Xiao, Y., Xu, Y., Li, Y., & Li, X. (2013). Network analysis reveals functional cross-links between disease and inflammation genes. *Scientific Reports*, *3*. https://doi.org/10.1038/srep03426